



Comprehensive performance analysis of objective quality metrics for digital holography



Ayyoub Ahar^{a,b,*}, Tobias Birnbaum^{a,b}, Maksymilian Chlipala^c, Weronika Zaperty^c, Saeed Mahmoudpour^{a,b}, Tomasz Kozacki^c, Malgorzata Kujawinska^c, Peter Schelkens^{a,b}

^a Dept. of Electronics and Informatics (ETRO), Vrije Universiteit Brussel (VUB), Pleinlaan 2, B-1050 Brussels, Belgium

^b imec, Kapeldreef 75, B-3001 Leuven, Belgium

^c Warsaw University of Technology, Institute of Micromechanics and Photonics, 8 Sw. A. Boboli St., 02-525 Warsaw, Poland

ARTICLE INFO

MSC:
00-01
99-00

Keywords:

Digital holography
Fourier holography
Holographic display
Hologram quality assessment
Visual quality assessment
Visual quality metrics

ABSTRACT

Objective quality assessment of digital holograms has proven to be a challenging task. While prediction of perceptual quality of the recorded 3D content from the holographic wavefield is an open problem; perceptual quality assessment from content after rendering, requires a time-consuming rendering step and a multitude of possible viewports. In this research, we use 96 Fourier holograms of the recently released HoloDB database to evaluate the performance of well-known and state-of-the-art image quality metrics on digital holograms. We compare the reference holograms with their distorted versions: (i) before rendering on the real and imaginary parts of the quantized complex-wavefield, (ii) after converting Fourier to Fresnel holograms, (iii) after rendering, on the quantized amplitude of the reconstructed data, and (iv) after subsequently removing speckle noise using a Wiener filter. For every experimental track, the quality metric predictions are compared to the Mean Opinion Scores (MOS) gathered on a 2D screen, light field display and a holographic display. Additionally, a statistical analysis of the results and a discussion on the performance of the metrics are presented. The tests demonstrate that while for each test track a few quality metrics present a highly correlated performance compared to the multiple sets of available MOS, none of them demonstrates a consistently high-performance across all four test-tracks.

1. Introduction

Predicting perceived visual quality for 3D media in general is highly desired. In this regard, objective quality evaluation of 3D content has been particularly pursued based on the type of the utilized technology to capture the depth information plus visual parallax. For example in [1–4] methods are introduced to evaluate the stereoscopic scenes. Alternatively, for Depth-Image-Based-Rendering techniques, which are essentially based on providing a depth value per regular 2D image pixel, methods like [5–9] are proposed for evaluation. For light field (LF) imaging, some recent efforts have proved to be efficient in their perceptual analysis, among them [10–14].

One of the emerging plenoptic modalities, which essentially can provide the most complete set of visual depth cues compared to the other 3D data modalities [15], is digital holography. Holography can present faithful reconstruction of the real object, with continuous parallax (within the recorded Field of View (FoV) and the freedom to refocus on unlimited number of focal distances — just like watching the real 3D scene.

However, perceptual quality prediction of the holographic content requires taking into account extra layers of complexity, on top of the general difficulties associated to quality prediction on plenoptic content [15,16], such as stereoscopic or light field content. A major complexity added to the Visual Quality Assessment (VQA) of digital holograms, consists of predicting the visual quality of the captured 3D scene from the complex-valued holographic fringes, instead of analysing several reconstructed view-ports of a scene involving an expensive rendering step. Even without considering the perceptual quality and visual appearance aspects, direct calculation of a mathematical error between a pair of complex-valued data is a challenge of its own (i.e. given the wrapping nature of phase and unboundedness of the magnitude, quantifying the maximum possible error or in general defining a bounded measure, which gives the relative error compared to the magnitude of the reference data, is not straightforward). Nonetheless, in [17,18], the authors have recently proposed a framework which can potentially be utilized with this extent.

The next major challenge for objective holographic VQA, is introduced by the lack of comprehensive perceptually-annotated holographic data sets. This comes from the fact that creating such data sets

* Corresponding author at: Dept. of Electronics and Informatics (ETRO), Vrije Universiteit Brussel (VUB), Pleinlaan 2, B-1050 Brussels, Belgium.
E-mail address: ayyoub.ahar@vub.be (A. Ahar).

is on its own a challenging task. First of all, collecting appropriately diverse sets of holograms that take into account the different characteristics of the many types of holograms is critical. This diversity can be addressed via an increased complexity of the recorded scene, *i.e.* via the number of objects available in the scene, the objects' depth, the distance from the recording sensor, the presence of spatial occlusions, and the object materials and surface structures, *e.g.* diffused, transparent, specular etc. Various types of hologram production are possible, *e.g.* computer generated hologram (CGH) or optically recorded hologram (ORH). But each type is limited by (virtual) sensor resolution, pixel pitch, reference wavelength, numerical aperture, sources of noise etc. All of the mentioned factors contribute to the attributes of the produced hologram. Some humble efforts to create publicly available holographic data sets can be found in [19,20]. Apart from data selection, conducting subjective experiments requires inventive but reproducible methodologies and scoring protocols for such plenoptic content.

Yet another issue for holographic subjective experiments arises from the fact that holographic displays with acceptable visual attributes are still rare and mostly operate under laboratory conditions, which require advanced technical skills to be properly configured. This makes it difficult to directly evaluate the holograms. To address this issue, most of the related researches, have utilized other types of displays to conduct holographic subjective experiments. In an early effort, Lehtimäki et al. [21,22] utilized a stereoscopic display to investigate applicable visual depth cues and appearance of a handful of optically recorded holograms. In [23], the perceptual quality of a limited set of holograms was even studied after printing on glass plates. Later in [24], a large 2D screen with high resolution was utilized to conduct subjective experiment on the central view of CGHs. In [25,26], a light field display was proposed to render a set of CGHs. Nonetheless, the most comprehensive subjective experiment for holographic content to this date is provided in [27] where a subjective experiment was conducted on a holographic display [28], a light field display [29], and a regular 2D display [30] using the same data set. A diverse set of both CGH and ORH was generated from single and multi-objects scenes carefully picked to represent various depths, recording distances and surface materials. The 96 holograms of this data set, called HoloDB, and the Mean Opinion Scores (MOS) gathered from each display setup have been made publicly available [31].

An additional complexity is the presence of speckle noise which is due to coherent light being used for the recording and displaying of holograms [32]. The speckle stems from the interference of multiple wavefronts, each of which is generated from a different illuminated point on the scene surfaces and is super-positioned with the reference light beam. Although, those interfering wavefronts are generated from *e.g.* the reflection of the highly coherent reference beam, the slightly different distances of the points to the recording sensor introduces spatial incoherence. Constructive, as well as destructive interference show-up in the reconstruction as very dark and bright spots, randomly positioned which alter the appearance of the main recorded data. The visibility of the speckle noise may vary significantly from object to object depending on the smoothness level of the object surface or, in case of an in-line recording setup, based on the thickness of the recorded sample. In [33], Bianco et al. provide a good review of the many proposed solutions for the suppression of speckle noise. More recently, in [34] and [35], state-of-the-art denoising methods have been evaluated both objectively and subjectively, using different sets of digital holograms and various objective metrics.

To the best of our knowledge, a full-fledged holographic perceptual quality metric is yet to be introduced due to the obstacles we reviewed above. Nevertheless, optimized compression techniques for digital holograms [15,16,36–38] and sophisticated numerical methods for CGH generation [39–44] continue to advance. Most recently, in [45] an extensive objective evaluation of compression artefacts on a set of CGHs is provided. However, only Signal-to-Noise Ratio (SNR) for hologram plane evaluations and Peak Signal-to-Noise Ratio (PSNR) for reconstructed holograms are utilized as the evaluation criteria. Moreover,

ORHs have not been included in the evaluation of [45]. Consequently, lack of reliable perceptual quality predictors is being felt more than ever. As a first step towards designing such algorithms, we report in this paper a comprehensive analysis on the prediction performance of state-of-the-art Image Quality Metrics (IQMs), based on the subjective scores provided by the HoloDB data set. We also explore their strengths and weaknesses with regard to the studied holographic data and summarize their behaviour when used to compare the holograms before and after rendering. Additionally, we study a conventional speckle denoising filter and compare the performance of the IQMs before and after denoising. Furthermore, to ensure that the results of our study is not biased to the characteristics of Fourier holograms, we study the performance of IQMs after a lossless, numerical conversion to the more conventional Fresnel hologram type.

The main objectives and novelties of this manuscript include:

- analysis and comparison of the accuracy of the predicted visual quality by the IQMs operating on:
 - the hologram plane;
 - numerically reconstructed holograms for different viewing angles and focal depth planes;
 - numerically reconstructed holograms after speckle denoising;
- assessing the impact of the hologram type, Fresnel or Fourier, on the prediction performance of the IQMs;
- definition of general guidelines for deployment of currently available IQMs on holographic content.

In Section 2, we discuss the characteristics of the tested data, we explain the experimental pipeline and the conversion process of Fourier holograms to Fresnel holograms. The tested IQMs as well as their input requirements are also discussed here, along with the utilized tools for the statistical analysis of the results. The test results along with statistical analysis and discussion about the performance of quality metrics are provided in Section 3. Finally, Section 4 provides the concluding remarks.

2. Experimental pipeline and test methodology

In this section, we describe the technical details required for the test data and experimental setup, the IQMs under test and the deployed methods for statistical analysis.

2.1. Description of test data

The Fourier Holograms of the HoloDB database [27] are used for our experiments. The database consists of 8 reference holograms, 4 of which were captured from the macroscopic real objects and the other 4 were numerically generated from the point clouds. In the optical recording off-axis lensless Fourier holography was employed to obtain holograms of comparatively large objects with wide angular FoV while maximizing the used Space Bandwidth Product (SBP) [46,47]. Compared to regular Fresnel holography, this method circumvent the common limitations on the sensor pixel pitch and resolution [48,49]. From the ORHs, on-axis holograms were obtained by bandpass filtering of positive frequencies components. The computer generated holograms were generated immediately as matching on-axis Fourier holograms. Fig. 1 shows the centre view of the rendered reference holograms in their frontal focal distances.

2.1.1. Compression specifications

Each reference hologram later was separated into its algebraic components and quantized to 8 bits per pixel (bpp). Those quantized real and imaginary parts were encoded separately using the JPEG 2000 [50, 51], intra H.265/HEVC [52] and wave atom coding (WAC) [53]. All three encoders compressed the reference holograms at bitrates of 0.25

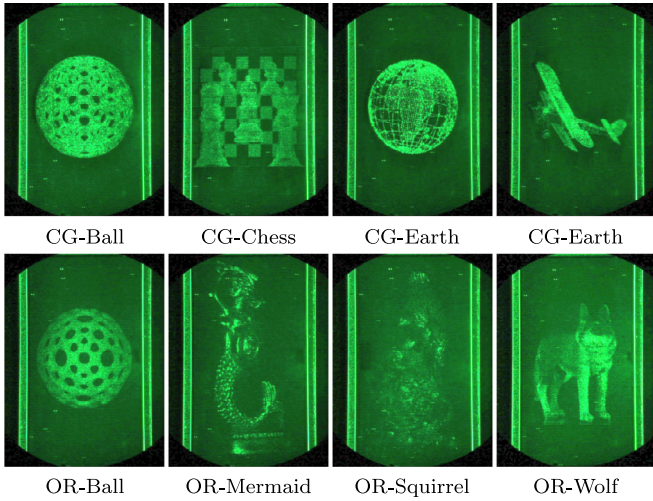


Fig. 1. Centre views of the reference holograms optically reconstructed in the holographic display. The top row contains the 4 CGHs numerically generated from point-clouds and the bottom row shows the ORHs recorded from real objects.

bpp, 0.5 bpp, 0.75 bpp, and 1.5 bpp (bpp=bits per complex-valued pixel) creating a data set of 96 compressed holograms. We remark, that upon compression with JPEG 2000 aliasing artefacts were introduced in the reconstructions at the lower two bitrates. In these cases, fine scales of the 4 level Mallat CDF 9/7 wavelet decomposition were suppressed, leading to a downsampling of the hologram by a factor of 2 per suppressed scale. This shrank the aliasing free cone [15] and caused the aliasing after reconstruction [54]. The test subjects were instructed to ignore these artefacts during the subjective experiments where the MOS values were produced.

2.1.2. Rating modalities

Along with each hologram a total of 12 MOS scores are provided in HoloDB, accounting for renders from 2 focal distances and 2 perspectives (centre view and right-corner view) which was displayed in 3 display setups (Holographic display, Light Field (LF) Display and regular 2D Monitor (2D)). The exceptions are the CG-Chess which was reconstructed in 3 focal distances (because of its deep volume of 310 mm compared to recording distance of 491 mm which includes multiple chess-pieces positioned in 3 rows) and the OR-Mermaid which was rendered in a single focal distance (due to its narrow thickness of only 5 mm compared to the recording distance of 450 mm). These selected perspectives and focal distances represent the most critical rendering points provided by the 3D box of the recorded scenes. They were chosen to limit the number of test sequences and avoid prolongation of the subjective experiment. For the current objective experiments particularly after the reconstruction, we could eventually test several other view points as it is outlined in [55]. However, since benchmarking the performance of tested quality prediction methods strictly relies on the MOS, we have limited our evaluations to those viewpoints which already have been annotated with the MOS.

2.2. Quality assessment pipeline

The typical procedure of visual quality assessment for a set of compressed digital images consists of simply comparing their perceptual appearance right after and before the compression step. However, in digital holography the quality assessment can be performed in more different points through the processing pipeline due to additional steps being involved for visualizing the 3D content in each hologram. Below, the implemented operations in our assessment pipeline and the assessment points are sequentially explained:

- Quality assessment in hologram domain: The real and imaginary parts of holograms are originally restored in floating point format. Though, the current implementation of the codecs like HEVC does not support such inputs and thus the algebraic components of holograms are quantized to 8bpp prior to the compression. Hereafter, the 8bpp hologram (8bpp for each of the real and imaginary channels) acts as the reference and undergoes the same operations as its compressed version. Choosing the 8bpp format is also related to the strict input requirements of some of the tested quality metrics as depicted in Table 1. The reference is being encoded–decoded. The first measurement point(QA_1) is here where the input of the encoder and the output of the decoder are given to the quality metrics to predict the visual quality of the holograms after reconstruction. These predictions are then compared w.r.t. the MOS. To assess the impact of the hologram type on the performance of quality metrics, the compressed and reference holograms are converted to Fresnel type. The second measurement point(QA_2) is here where similar quality evaluations as the QA_1 are performed.
- Quality assessment after reconstructing the objects: Visualization of the holographic 3D contents, mainly requires two additional steps. First, the data values of the quantized hologram needs to be scaled back to the original data range of the source hologram. Second, each hologram should be back-propagated to reveal the captured 3D content. This process is repeated for both the reference and the compressed(decoded) holograms. After the centre and corner views are rendered from both the hologram pair, those views are quantized to 8bpp again (to meet the input requirements of IQMs). Before quantization, data range clipping is performed to ensure few extreme values resulted from the speckle noise does not create an unnecessary expansion of dynamic range. Then again they are given to the quality metrics to predict their visual quality which is compared to their corresponding MOS(QA_3). Finally, to evaluate the impact of the speckle noise on the performance of the quality measures, a speckle denoising process is performed on the extracted views. Then in QA_4 track, the same quality predictions and performance evaluations as the QA_3 are conducted.

Fig. 2 shows an abstract scheme of the experimental pipeline and the four steps for which we conducted the quality evaluations in this paper. For the sake of clarity, we have omitted separate processing lines for the real and imaginary parts in the represented hologram domain.

2.3. Conversion of fourier to Fresnel holograms

Aside of the Fourier form of the considered holograms, which makes the most efficient use of the available space-bandwidth product, we also considered synthetically obtained Fresnel forms. Fourier holograms are marked by their wavefronts not being convergent and are thus under a planar reference wave reconstructed only in the conjugated plane of a lens. The wavefronts of Fresnel holograms converge under plane wave illumination without any additional lenses. Both forms are common and possess a substantially different space-frequency behaviour. Representing the same content as in a Fourier hologram H_{Four} in a Fresnel hologram H_{Fres} without aliasing, requires upsampling $US(\cdot)$ of H_{Four} , such that the effective pixel pitch is halved (sampled bandwidth doubled), and demodulated with a parabolic wavefront K of mono-chromatic light of wavelength λ .

$$H_{Fres} = e^{\frac{\pi i}{\lambda z}(x^2+y^2)} \cdot \underbrace{US(H_{Four})}_{=: K(z)} \quad (1)$$

The distance z in used in the kernel $K(z)$ is not the actual scene distance, as this would require upsampling by a large factor m . Instead, we compute z such that objects are brought into focus upon direct

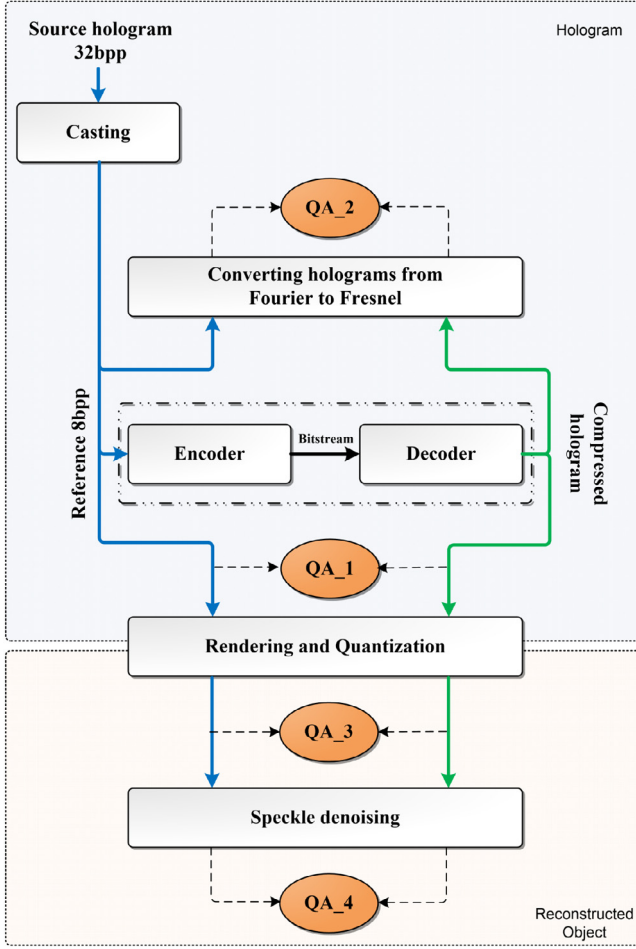


Fig. 2. A high-level scheme of the experimental pipeline showing where the predictions of the IQMs were compared with the MOS scores. The four Quality Assessment tracks in this paper are shown as QA_1 to QA_4 in the diagram.

observation as close to the hologram plane as possible without causing aliasing after upsampling. We will provide a general formula, but consider only $m = 2$ as the smallest integer m . Only upsampling by an integer m preserves the original samples and ensures invertibility.

Given a Fourier hologram rectangular hologram recorded with a square pixel pitch \hat{p} it is necessarily bandwidth-limited at $\pm(2\hat{p})^{-1}$ and it is sufficient to consider only the larger of the two hologram dimensions, i.e. \hat{N} px. If the scene centre is in focus after a Fourier transform of the hologram, the main ridge of the phase space footprint will be aligned with the spatial axis, cf. Fig. 3 and no additional offset of z needs to be considered. We start with the ansatz that the largest frequencies, $|(2\hat{p})^{-1}|$, should be located at the edges of the (of the major dimension) of the Fresnel hologram, i.e. at $|\xi| = N\hat{p}/2$, after upsampling ($p = \hat{p}/m$, $N = m\hat{N}$) and reference wave demodulation. In combination with the space-frequency law of Fresnel diffraction [56]

$$f_{\xi} = \frac{\xi}{\lambda z}, \quad (2)$$

it is straightforward to show that z is given as

$$z = \frac{m}{m-1} \frac{N p^2}{\lambda}. \quad (3)$$

Eq. (1) is reversible without any loss and thus testing the reconstruction only from either representation is sufficient. The conversion may be understood as obtaining a Fresnel hologram from its compact space-bandwidth representation [57]. Alternatively, the Fourier hologram can

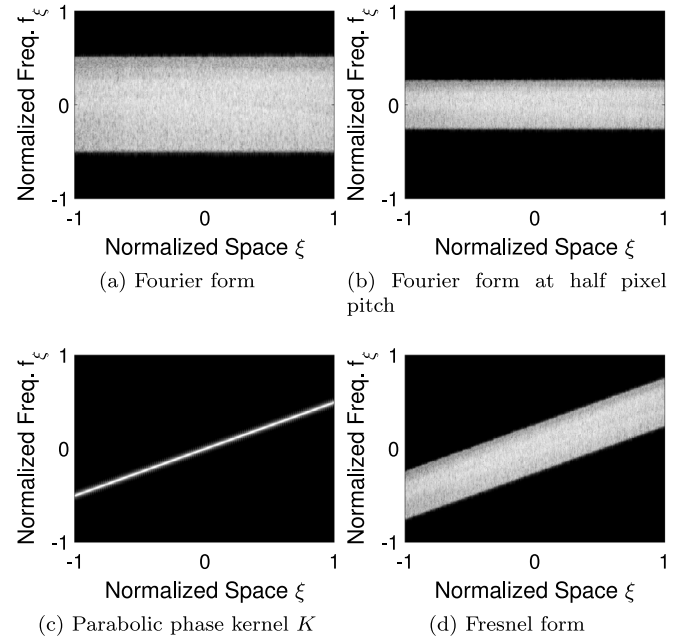


Fig. 3. Exemplary space-frequency footprint of 1D cross sections of the Fourier (a) and the Fresnel form (d) of the OR-Squirrel hologram. The parabolic wavefront (c) focuses the subsampled Fourier hologram (b) at a distance of ~ 52.4 cm in this case.

be interpreted as the interim hologram formed by a Fresnel hologram and the parabolic phase kernel K , which is used to facilitate numerical back-propagation of a hologram within the Fresnel approximation. The space-frequency footprints of either form and the parabolic kernel are given in Fig. 3. As we will see they will significantly influence the performance of the quality metrics.

2.4. Quality metrics

For the current experiment, we use the mathematical measures which can directly evaluate the complex data, including Peak Signal-to-Noise Ratio (PSNR), Mean Squared Error (MSE) and its normalized form (NMSE), where the MSE is normalized by the Frobenius norm of the reference hologram. We also test our recently proposed IQM called Sparseness Significance Ranking Measure (SSRM) [58] which is native on the complex domain and showed a good compatibility for the quality evaluation of a limited set of CGHs [59]. In its original form, SSRM operates solely in Fourier domain and predicts the similarity by comparing Fourier coefficients of reference and the impaired data. It calculates a separate quality score for the DC term which later is combined with the score of other coefficients. However, in some preliminary tests, we found out that in case of Fourier holograms, this may not necessarily result in a more robust performance. Consequently, we considered a version of this method which treats the DC term just like other Fourier coefficients. The results for testing this version are presented under the name of SSRMt. Since the compression of holograms was performed separately on the real and imaginary parts of the holograms, we can also separately evaluate those real and imaginary parts with the state-of-the-art IQMs –which by default can operate on the real-valued data – and then calculate the arithmetic mean to achieve one quality prediction score for each compressed hologram. The goal is to check whether they can relate their signal fidelity measurements on holograms, to the overall perceptual quality of the rendered scenes (represented by the provided MOS). It is also possible to decompose the complex values to their polar components (i.e. amplitude and phase). Although, since the wrapped phase is stored for holograms, quality predictions on such data can have a significant

Table 1
Information of the compared quality prediction algorithms. If necessary, the required dynamic range (DR) of input is identified.

IQM	Input requirements	Parameter settings	Underlying principle	
1	MSE	Real/Complex	N\A	Pixel-based squared difference measurement with average pooling
2	NMSE	Real/Complex	Normalized by Frobenius norm of reference data	Normalized pixel-based squared difference measurement
3	PSNR	Real/Complex	Real Peak = 255, Complex Peak = 1	Ratio between peak signal energy and pixel-based squared difference measurement on logarithmic scale
4	SSRM	Real/Complex	N\A	Expresses correlation between ranked amplitudes in frequency domain based on their sparseness significance
5	SS-RMT	Real/Complex	N\A	Adaptation of the SSRM where the DC term is not separately evaluated but is scored along with other coefficients
6	SSIM	Real	Dynamic Range(L)= 255 Exponents = [1,1,1] Regularization Constants: C1 = (0.01*L) ² C2 = (0.03*L) ² C3 = C2/2	Regularized weighted comparison of local luminance mean, variance (contrast) and covariance (structure)
7	IWSSIM	Real	Dynamic Range(L)= 255 Exponents = [1,1,1] Regularization Constants: C1 = (0.01*L) ² C2 = (0.03*L) ² C3 = C2/2	Calculates a scale variant weighted SSIM-based local similarity over each scale of a Laplacian pyramid decomposition. The scale weights are calculated based on Gaussian scale mixture model of natural images
8	MS-SSIM	Real	Dynamic Range(L)= 255 Regularization Constants: C1 = (0.01*L) ² C2 = (0.03*L) ² C3 = C2/2 Number of scales = 5	Decomposes the data to several resolutions scales by iterative low-pass filtering and downsampling and compares the contrast and similarity at each level based on original SSIM. The last stage involves luminance comparison. A weighted multiplication yield the final score.
9	UQI	Real	Block size = 8	Special case of SSIM without weight and regularization
10	GMSD	Real (DR 8:8 bit)	Regularization Constant: T = 170	Calculates the standard deviation of a similarity map calculated based on pixel-wise gradient magnitude comparison
11	FSIM	Real (DR 8:8 bit)	Regularization Constants: T1 = 0.85, T2 = 160, T3 = 200	Weighted similarity, calculated on image gradient magnitudes and phase congruency
12	NLPD	Real	Laplacian Pyramid levels = 5	RMSE of weighted Laplacian pyramid decomposition
13	VIFp	Real (DR 8:8 bit)	N\A	Mutual information is calculated after modelling the image source using Gaussian scale mixture model on wavelet coefficients extracted from steerable pyramid decomposition and modelling the distortion and HVS channels

negative impact on the performance of the quality prediction methods and further complicate interpretation of our test results. The tested IQMs include: FSIM [60], IWSSIM [61], MS-SSIM [62], VIF [63], NLP [64], GMSD [65]. The UQI [66] and SSIM [67] have been utilized in the context of holographic data e.g in [68,69] and [70], consequently

we added them also for the experiments. It should be noted that the machine learning based IQMs along with the ones which require the information from the chrominance channels of the colour images were omitted from this evaluation. The former group obviously require large training sets to be adapted for the case of holography while our holographic data set is too small for such purposes. The later group also require information from the colour channels while our holograms are all monochromatic. Table 1, summarizes the input requirements of the tested quality metrics and provides their parameter settings as it was used in this experiment. It also summarizes their main features and utilized underlying principle. For the experiments in QA_1 and QA_2 tracks where the input data are the complex-valued holograms, the real and imaginary parts of holograms were separately tested by all of the studied 13 methods. Although, for the methods 1 to 5 in the Table 1 which can directly be deployed for the complex valued data, we tested them on the complex values. Their results are hereafter distinguished by the suffix “_C”. Note that for the MSE, NMSE and the PSNR, calculation results averaged over real and imaginary parts in QA_1 and QA_2, are not reported due to being almost identical to those directly calculated on complex values.

2.5. Statistical analysis

In this experiment, we follow the guidelines of the Video Quality Experts Group(VQEG) [71] for evaluating the predictive performance of the tested IQMs. In this regard, three evaluation criteria are considered namely: prediction monotonicity, accuracy, and consistency.

The Spearman Rank-order Correlation Coefficient (SROCC) and Kendall’s tau Rank-Order Correlation (KROC) are utilized to measure the strength of the IQMs in predicting the rank-ordering of the MOS. Next, we utilize the Pearson Correlation Coefficient (PCC) to measure the linear correlation between the MOS and the predicted scores.

The score scale of several IQMs are not the same as the MOS and their score functions have a non-linear behaviour. To compensate, it is recommended in [71] to fit a logistic function on the predicted scores with the constraint of being monotonic in the fitted interval. Afterwards, measuring the PCC and the Root Mean Squared Error(RMSE) between the MOS and the fitted scores will provide an estimation of prediction accuracy for the tested IQMs. In this experiment, we utilized a generalized 4-parameter logistic function:

$$MOS_p = f(X; \beta) = \beta_1 + \left(\frac{\beta_2 - \beta_1}{1 + e^{-\beta_3(X - \beta_4)}} \right) \quad (4)$$

where X is the set of quality scores predicted by the objective measures and MOS_p is the set of fitted scores. It is recently utilized in [70], which automatically guaranties the monotone behaviour of the fit function.

Hereafter, the PCC measured before and after the logistic regression is referred as PCC_NoFit and PCC_Fitted respectively. The recommended measure of prediction consistency is the outlier ratio. It is simply calculated as the ratio of the number of fitted predictions out of the 95% confidence intervals of the MOS divided to the total number of MOS.

Additionally, we perform a statistical significance test to determine whether the difference between performance of two quality metrics(here represented by the absolute difference between the fitted scores and the MOS) is statistically significant. A two-sided t-test is performed [72] where the null hypothesis is that there is no difference between performance of quality metrics, against the alternative hypothesis that the difference in performance is significant. The null hypothesis is rejected at a 5% significance level.

3. Experimental results and analysis

3.1. Evaluation of quality metrics in hologram plane

In this section, we discuss the outcome of the first (QA_1) test track referring to the evaluation of the performance of the studied

IQAs in case of compressing Fourier holograms in the hologram plane, and a second (QA_2) test track performing this test after converting the reference and decoded Fourier holograms to Fresnel holograms. Following the procedure depicted in Fig. 2, for every complex-valued hologram, the same synthetic aperture used in [27] to extract the centre and right-corner views to display and obtain the MOS, was utilized to select exactly the same data from the original hologram and the corresponding compressed versions. Note, that the compression process was done on the real and imaginary parts of the complete holograms and then the synthetic apertures were applied, not vice versa. Thereafter, the algebraic components of those cropped compressed holograms were compared to their non-compressed unsigned 8 bit version, i.e. a cropped part of the input data to the encoder was compared to the same cropped part of the output of the decoder. The obtained results are compared to the MOS for the corresponding perspective averaged over the focal distances. In order to adhere to the page limits, only the overall statistics across all perspectives are presented, not the per view evaluations.

3.1.1. QA_1 - Evaluation on Fourier Holograms

The overall results of our statistical analysis for this experimental track is presented in Table 2. The results are shown based on each set of MOS obtained from conducting the subjective test for holographic display (MOS_OPT), the light field display (MOS_LF) and the regular 2D display (MOS_2D).

A quick look at the table makes it evident that SSIM ranks best based on almost all evaluation criteria, independent of which MOS it was compared to. Seen the fact that the majority of the tested IQMs previously shown to be superior to the SSIM at least for the case of classical digital images, such a good performance of SSIM here was not expected. In digital imaging, SSIM is designed to measure the visual similarities between the structural information of the compared pair of data. However, when measured in hologram domain in particular, there is no specific structure in the data and the holograms in general mostly appear as noise. So, how SSIM is able to make such an accurate prediction by both correct ranking of the distorted holograms (high SROCC) and linear correlation of the predicted scores compared to the MOS (high PCC)?

To answer this question, we took the CG-Ball hologram as an example and calculated the SSIM quality map for the real and imaginary parts of the compressed holograms separately. Fig. 4 shows these quality maps. Due to resemblance of the quality maps for real and imaginary parts we only show the results for the real parts. The top row shows the results for the CG-Ball compressed with HEVC-Intra mode. While not directly visible in the compressed data itself, a certain grid artefact (related to the boundaries of the HEVC code blocks) appears in the SSIM quality map where the data is normalized based on the luminance and the standard deviation of pixel values (and compared to the non-compressed data using SSIM formulae). In the second row, quality maps for the hologram compressed with JPEG 2000 are depicted. Here, also heavy blocking artefacts show up in SSIM quality maps which otherwise are not clearly distinguishable in the compressed data, when observed directly. These artefacts at least partially could be related to what after the reconstruction appears as the aliasing artefacts (see Section 2.1.1). In the third row, where quality maps for WAC are shown, no structured compression artefact are found. However, just like with the other two encoders, SSIM is able to correctly distinguish the compression levels by verifying those increases or decreases in the magnitude of similarity (i.e. changes in brightness of quality maps observed in passing from one quality map to another. See the brightness of Fig. 4.(a) to (d) or Fig. 4.(i) to (l) and also note Fig. 4.(i) compared to (e) and (a)). Overall, it appears that by amplification of the compression artefacts which otherwise are wrapped-up and masked by the heavily noisy environment of the hologram (i.e. the interference pattern), SSIM is able to rather easily predict the relative quality score and rank of the decoded holograms without having any explicit knowledge about the actual appearance of the objects after the reconstruction.

Another observation based on the results of Table 2 is that the multi-scale metrics are generally worse than single scale methods (such as SSIM, PSNR). This probably due to the fact that weights that assigned to the scores of each scale in multi-scale metrics. These weights are assigned based on characteristics of the human visual system (HVS) and psychophysical measurements of digital images and might not be relevant to complex holographic data before the reconstruction. However, it must be noted that even the predictions of the worst metrics in each category are not failing completely. For example, the UQI metric which seems to achieve the lowest correlation results, has an average SROCC of 0.8.

To better understand the significance of SSIM performance in relation to the other metrics, we have provided 3 significance tables, one per set of MOSs in Fig. 5. Here, it can be seen that the performance of the SSIM compared to the IQMs like MSE, NMSE, PSNR, GMSD, and NLDP is not significantly better, while compared to the other metrics it shows a clear superiority. Accounting for the MOS_LF scores shown in Fig. 5.(b), the results are even closer and most of the quality metrics are on par with respect to the criterion explained in Section 2.5. The reason for these results could be related to the nature of Fourier holography. The object-field wavefronts captured by a Fourier hologram are not convergent at any finite distance. Thus, we may interpret them as being convergent at a point at infinity. Propagation over infinite distances is described by the Fraunhofer propagation and essentially reduced to a Fourier transform. Henceforth, when analysing Fourier holograms we essentially study the Fourier domain of the in-focus light field of an object. Thus, the metrics which try to relate their local (e.g SSIM, GMSD) or global error measurements (e.g PSNR, MSE) to the perceptual quality often perform better than metrics which rely on internal transformations. (e.g SSRM has an internal Fourier transform which, when applied to a Fourier hologram, brings the data back to the object plane).

3.1.2. QA_2 - Evaluation on Fresnel holograms

The next test relates to the synthesized Fresnel holograms. The same analysis process as for the Fourier holograms in the previous section is deployed. The Fourier to Fresnel hologram conversion was detailed in Section 2.3. While the space-frequency behaviour of the holograms changed drastically (see Fig. 3, the data still represents wavefields with statistics very different from natural images).

Interestingly, VIFp shows the best performance across all 3 MOS values in Table 3. Although its rather lower PCC_NoFit values indicate that its behaviour with respect to the MOS is less linear compared to its competitors like SSIM — closely following behind the VIFp. Another point to be noticed is the rather larger gap between the top and worst performing metrics w.r.t each criterion. This gap becomes more evident in the significance tables of Fig. 6. Here, it is clear that the top 3 quality metrics namely VIFp, UQI and SSIM are rather similar while their predictions are significantly better than others.

3.2. Evaluation on reconstructed holograms

In this section we provide the evaluation results of the experimental tracks QA_3 and QA_4, i.e. the evaluations after reconstruction of the Fourier holograms and after reconstruction plus speckle denoising. Here, we used the reconstructions of the centre and right-corner views, and at different focal distances as provided by the HoloDB. These reconstructions are generated with the same synthetic aperture as was used in previous experimental tracks, as well as in [27] to obtain the MOS. Only the absolute amplitudes of the reconstructed wavefield are examined by the IQMs as it is the principal part of the light sensed by our eye and upon which the MOS scores are based. The predictions and the MOS scores for all views and focal depths per rate point and codec are averaged to facilitate a direct comparison of IQM performances before and after reconstruction. Also, similar to the previous test tracks QA_1 and QA_2, only the overall statistics are presented.

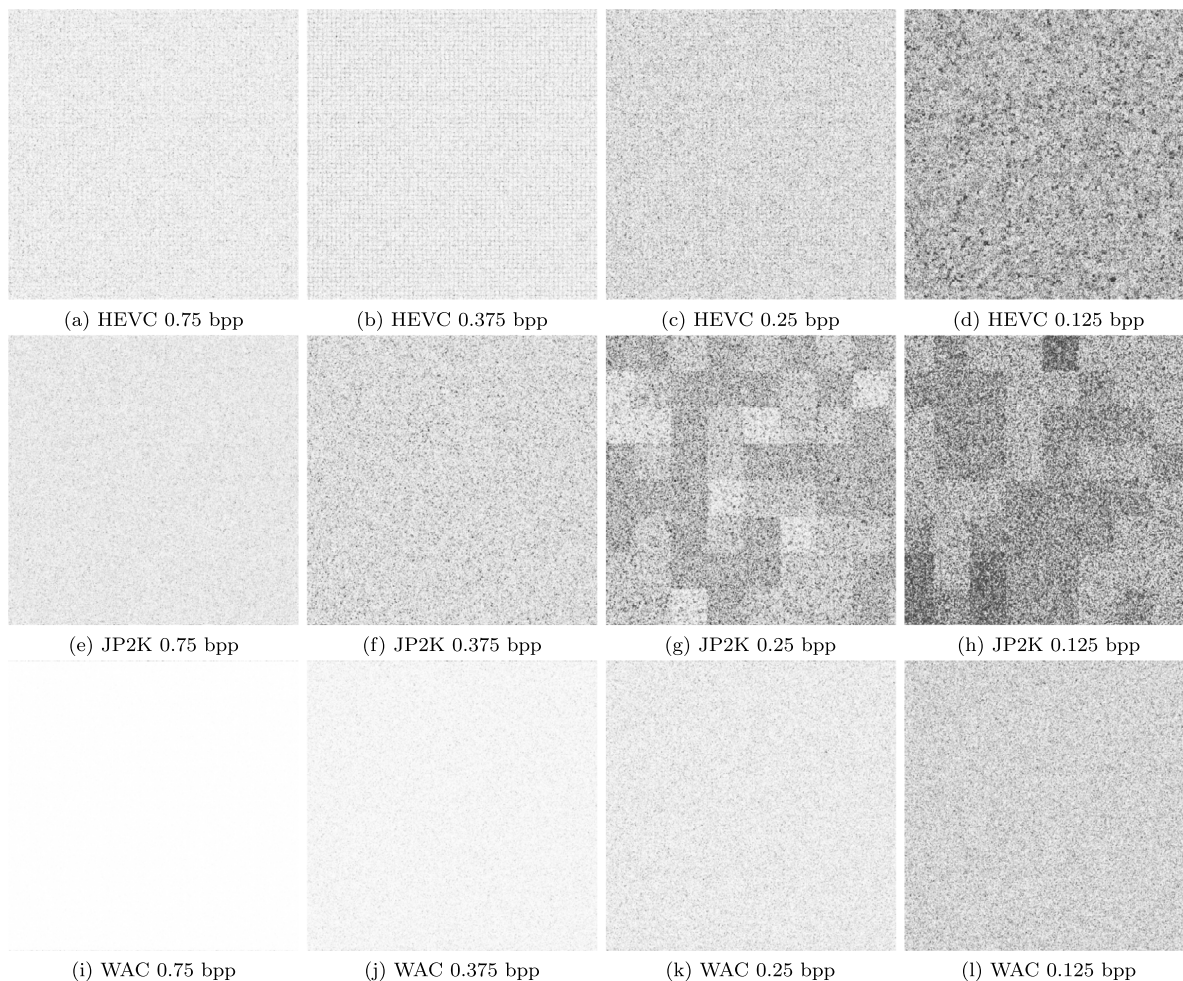


Fig. 4. SSIM quality maps for the real part of the compressed CG-Ball. Bright and dark areas account for high and low similarity accordingly. It appears that all encoders, especially HEVC-Intra mode and JPEG 2000, create particular compression artefacts, which clearly intensify when operating at lower bit rates. This creates a distinctive gradient over quality maps, enabling the SSIM to correctly rank and predict their relative quality score.

Table 2

Statistical evaluation of quality metrics for the Fourier holograms, rated after hologram decoding; thus before rendering. The statistics based on the MOS scores obtained from optical holographic display (OPT), the light field display (LF) and the regular 2D display (2D) and the IQM scores are separately shown. Quality metrics directly evaluated on the complex-valued data have “_C” as postfix.

		SSIM	SSRM	FSIM	GMSD	IWSSIM	MS-SSIM	UQI	VIFp	NLPD	SSRMt	SSRM_C	PSNR_C	SSRMt_C	MSE_C	NMSE_C
MOS-OPT	SROCC	0.9069	0.8382	0.8425	0.8855	0.8604	0.8833	0.8032	0.8365	0.8876	0.8379	0.8366	0.8821	0.8364	0.8821	0.8907
	KRCC	0.7314	0.6334	0.6424	0.698	0.6648	0.6913	0.6036	0.6331	0.6966	0.6332	0.6329	0.6851	0.6329	0.6851	0.6967
	PCC_NoFit	0.9088	0.8244	0.806	0.8708	0.8277	0.8593	0.7441	0.8138	0.8667	0.8245	0.8252	0.7389	0.8254	0.8731	0.8805
	PCC_Fitted	0.9129	0.8305	0.8429	0.8839	0.8565	0.8808	0.757	0.8329	0.8837	0.8305	0.8297	0.878	0.8297	0.8779	0.8873
	RMSE	0.4589	0.6261	0.6049	0.5258	0.5803	0.5323	0.7346	0.622	0.5262	0.6262	0.6275	0.538	0.6275	0.5382	0.5184
	OutlierRatio	0.6042	0.6771	0.6875	0.6458	0.6823	0.6354	0.7031	0.6719	0.651	0.6771	0.6979	0.5781	0.6979	0.5938	0.599
MOS-LF	SROCC	0.9012	0.8716	0.8762	0.8965	0.8847	0.8966	0.8158	0.8588	0.8898	0.8714	0.8725	0.8812	0.8726	0.8812	0.8894
	KRCC	0.7254	0.6927	0.6991	0.7225	0.7031	0.7176	0.6289	0.673	0.7078	0.6923	0.6939	0.6916	0.6942	0.6916	0.7022
	PCC_NoFit	0.8776	0.857	0.8151	0.8871	0.8187	0.8368	0.7286	0.8569	0.884	0.8572	0.8651	0.8028	0.8655	0.8633	0.8705
	PCC_Fitted	0.9006	0.8718	0.8774	0.8929	0.8817	0.8961	0.8285	0.8588	0.8897	0.8718	0.8729	0.8814	0.8731	0.8819	0.8888
	RMSE	0.4531	0.5107	0.5002	0.4694	0.4919	0.4626	0.5838	0.5341	0.476	0.5107	0.5086	0.4925	0.5084	0.4915	0.4778
	OutlierRatio	0.5729	0.599	0.6198	0.5833	0.5781	0.5781	0.6406	0.6146	0.5938	0.599	0.6042	0.6354	0.6042	0.6198	0.6302
MOS-2D	SROCC	0.9001	0.8365	0.8386	0.8795	0.8608	0.8685	0.7901	0.8249	0.8738	0.8364	0.8365	0.8775	0.8365	0.8775	0.8846
	KRCC	0.7215	0.6424	0.6438	0.6892	0.6713	0.6703	0.5859	0.6218	0.6804	0.6419	0.6427	0.6864	0.6427	0.6864	0.6967
	PCC_NoFit	0.8651	0.8229	0.7708	0.8767	0.7904	0.8026	0.7105	0.8346	0.878	0.8232	0.8316	0.8168	0.832	0.8561	0.8625
	PCC_Fitted	0.9052	0.8454	0.8482	0.8859	0.8682	0.8771	0.7286	0.8366	0.8874	0.8454	0.8455	0.8894	0.8456	0.89	0.8938
	RMSE	0.4697	0.5903	0.5854	0.5126	0.5484	0.5309	0.757	0.6054	0.5094	0.5902	0.59	0.5052	0.5899	0.504	0.4957
	OutlierRatio	0.6354	0.7552	0.7396	0.6667	0.6719	0.6875	0.7604	0.7396	0.6458	0.7552	0.7604	0.6198	0.7604	0.6302	0.6198

3.2.1. QA₃ - Evaluation on reconstructed Fourier holograms

Table 4 provides the statistical results for the IQMs, measured immediately after the reconstruction process, with respect to the measured MOSs. It is interesting to note that at this point, the reconstructions are much more similar to natural images. However, their statistical properties in general do not exactly follow that of natural imagery. The reconstructions are contaminated for example by speckle noise and do contain out of focus objects as it was explained in Section 1.

Nonetheless, the IQMs are expected to predict perceptual quality much better than prior to reconstruction since most of them are optimized for natural images.

For the majority of the tested IQMs we find a very competitive performance with the top ranks shared among the MSE family (MSE, NMSE, and PSNR) as well as the SSRM and SSRMt measures. However, the performances of the SSIM, UQI, and VIFp are this time however ranked worst — being in strong contrast to their performance before

MOS-OPT	LOI	VIFp	SSRM	SSRM_C	SSRMt	SSRMt_C	FSIM	IWSSIM	PSNR_C	MSE_C	GMSD	MSSIM	NLPD	SSIM	NMSE_C
UQI	0	0	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1
SSRM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VIFp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SSRMt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SSRM_C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SSRMt_C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
FSIM	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
IWSSIM	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MSSIM	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
NLPD	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SSIM	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
NMSE_C	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

(a) MOS-OPT

MOS-LF	LOI	VIFp	SSRM	SSRM_C	SSRMt	SSRMt_C	FSIM	IWSSIM	PSNR_C	MSE_C	GMSD	MSSIM	NLPD	SSIM	NMSE_C
UQI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VIFp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SSRM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SSRM_C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SSRMt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SSRMt_C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
FSIM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
IWSSIM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PSNR_C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MSE_C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GMSD	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MSSIM	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NLPD	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SSIM	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NMSE_C	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0

(b) MOS-LF

MOS-2D	LOI	VIFp	SSRM	SSRM_C	SSRMt	SSRMt_C	FSIM	IWSSIM	PSNR_C	MSE_C	GMSD	MSSIM	NLPD	SSIM	NMSE_C
UQI	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
SSRM	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SSRM_C	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SSRMt	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SSRMt_C	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
FSIM	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
IWSSIM	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PSNR_C	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MSE_C	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GMSD	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
MSSIM	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
NLPD	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SSIM	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
NMSE_C	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

(c) MOS-2D

Fig. 5. Statistical significance tables for the evaluation of quality metrics on Fourier holograms, rated after decoding. The statistics are separately shown based on the MOS scores obtained from optical holographic display (OPT) (a), the light field display (LF) (b) and the regular 2D display (2D) (c). The significance of the IQM performance in each row compared to the others is quantified by 1, 0 and -1 representing significantly better performance, no significant difference and significantly worse performance respectively.

Table 3

Statistical evaluation of quality metrics for the synthesized Fresnel holograms, rated before hologram reconstruction. The statistics based on the MOS scores obtained from optical holographic display (OPT), the light field display (LF) and the regular 2D display (2D) and the IQM scores are separately shown. Quality metrics directly evaluated on the complex-valued data are prefixed by “_C”.

	SSIM	SSRM	FSIM	GMSD	IWSSIM	MS-SSIM	UQI	VIFp	NLPD	SSRMt	SSRM_C	PSNR_C	SSRMt_C	MSE_C	NMSE_C
MOS-OPT															
SROCC	0.8912	0.7076	0.5179	0.8463	0.791	0.8769	0.8922	0.9088	0.7632	0.7078	0.6928	0.8015	0.693	0.8015	0.8112
KRCC	0.7038	0.5142	0.3614	0.6563	0.5912	0.6802	0.7067	0.7343	0.5642	0.5143	0.5021	0.6047	0.5022	0.6047	0.6117
PCC_NoFit	0.8912	0.6547	0.4607	0.8212	0.7596	0.854	0.8893	0.8045	0.7326	0.6548	0.6441	0.6954	0.6445	0.7699	0.783
PCC_Fitted	0.8944	0.6876	0.4833	0.8304	0.7721	0.8716	0.8953	0.916	0.7388	0.6878	0.6715	0.7878	0.672	0.7885	0.8036
RMSE	0.5027	0.8163	0.9841	0.6264	0.7144	0.5512	0.5007	0.4509	0.7576	0.816	0.833	0.6924	0.8325	0.6914	0.6691
OutlierRatio	0.6042	0.7396	0.8073	0.6302	0.724	0.6198	0.599	0.5625	0.6719	0.7396	0.7344	0.6406	0.7344	0.651	0.651
MOS-LF															
SROCC	0.887	0.6973	0.5102	0.8399	0.7561	0.8687	0.8858	0.9067	0.7597	0.6968	0.6842	0.8148	0.6835	0.8148	0.8253
KRCC	0.7022	0.5138	0.3674	0.6598	0.5644	0.68	0.7004	0.733	0.5714	0.5134	0.5055	0.6199	0.5045	0.6199	0.6297
PCC_NoFit	0.8724	0.6494	0.4672	0.8348	0.7105	0.8196	0.8621	0.855	0.7501	0.6494	0.642	0.7675	0.642	0.7889	0.803
PCC_Fitted	0.8859	0.7096	0.5034	0.8378	0.7594	0.8698	0.8849	0.9064	0.7636	0.7097	0.6945	0.8155	0.6946	0.8155	0.8278
RMSE	0.4837	0.7345	0.9008	0.5692	0.6782	0.5144	0.4855	0.4403	0.6732	0.7344	0.7501	0.6034	0.7499	0.6034	0.5849
OutlierRatio	0.625	0.724	0.8385	0.6198	0.6563	0.6354	0.6354	0.5625	0.7292	0.724	0.7292	0.6771	0.7292	0.6719	0.6875
MOS-2D															
SROCC	0.8901	0.6928	0.5213	0.8413	0.7769	0.8639	0.8874	0.901	0.7635	0.6923	0.6805	0.8227	0.6797	0.8227	0.8344
KRCC	0.7062	0.5017	0.3679	0.652	0.5757	0.6678	0.7021	0.7205	0.5711	0.5011	0.4923	0.6184	0.4913	0.6184	0.6315
PCC_NoFit	0.8674	0.6494	0.4745	0.851	0.7308	0.8074	0.8541	0.8667	0.7687	0.6493	0.6434	0.7922	0.6431	0.7976	0.812
PCC_Fitted	0.8953	0.7324	0.543	0.8599	0.8009	0.8814	0.8932	0.9063	0.7967	0.7323	0.7193	0.8408	0.7192	0.8408	0.8512
RMSE	0.4924	0.7525	0.928	0.564	0.6618	0.5221	0.497	0.4671	0.668	0.7525	0.7677	0.5982	0.7679	0.5982	0.58
OutlierRatio	0.6563	0.7604	0.8073	0.6667	0.7188	0.6615	0.6927	0.6146	0.7448	0.7604	0.7604	0.6771	0.7604	0.6771	0.6875

MOS-OPT	FSIM	SSRM	SSRM_C	SSRMt	SSRMt_C	FSIM	IWSSIM	PSNR_C	MSE_C	GMSD	MSSIM	NLPD	SSIM	NMSE_C
FSIM	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
SSRM	1	0	0	0	0	0	0	0	0	0	0	0	0	0
SSRM_C	1	0	0	0	0	0	0	0	0	0	0	0	0	0
SSRMt	1	0	0	0	0	0	0	0	0	0	0	0	0	0
SSRMt_C	1	0	0	0	0	0	0	0	0	0	0	0	0	0
FSIM	1	0	0	0	0	0	0	0	0	0	0	0	0	0
IWSSIM	1	0	0	0	0	0	0	0	0	0	0	0	0	0
PSNR_C	1	1	1	1	1	1	1	1	1	1	1	1	1	1
MSE_C	1	1	1	1	1	1	1	1	1	1	1	1	1	1
NMSE_C	1	1	1	1	1	1	1	1	1	1	1	1	1	1
GMSD	1	1	1	1	1	1	1	1	1	1	1	1	1	1
MSSIM	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SSIM	1	1	1	1	1	1	1	1	1	1	1	1	1	1
UQI	1	1	1	1	1	1	1	1	1	1	1	1	1	1
VIFp	1	1	1	1	1	1	1	1	1	1	1	1	1	1

(a) MOS-OPT

MOS-LF	FSIM	SSRM	SSRM_C	SSRMt	SSRMt_C	FSIM	IWSSIM	PSNR_C	MSE_C	GMSD	MSSIM	NLPD	SSIM	NMSE_C
FSIM	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
SSRM	1	0	0	0	0	0	0	0	0	0	0	0	0	0
SSRM_C	1	0	0	0	0	0	0	0	0	0	0	0	0	0
SSRMt	1	0	0	0	0									

Table 4

Statistical evaluation of quality metrics for the reconstructed Fourier holograms. The statistics based on the MOS scores obtained from optical holographic display (OPT), the light field display (LF) and the regular 2D display (2D) and the IQM scores are separately shown.

	PSNR	SSIM	SSRM	FSIM	GMSD	IWSSIM	MS-SSIM	UQI	VIFp	NLPD	SSRMt	MSE	NMSE	
MOS-OPT	SROCC	0.9319	0.5164	0.9218	0.7166	0.9185	0.9124	0.8394	0.4897	0.7682	0.9077	0.9319	0.9347	0.9306
	KRCC	0.7727	0.3648	0.7608	0.5322	0.7433	0.7343	0.6437	0.3468	0.5626	0.7267	0.7723	0.7771	0.7688
	PCC_NoFit	0.8045	0.4847	0.9106	0.7036	0.897	0.9006	0.8256	0.5047	0.7002	0.882	0.9284	0.9267	0.9179
	PCC_Fitted	0.9352	0.5044	0.9265	0.7161	0.9168	0.914	0.8389	0.5126	0.7628	0.9076	0.9294	0.9374	0.9341
	RMSE	0.3981	0.9707	0.4229	0.7847	0.4488	0.456	0.6118	0.9652	0.7268	0.4719	0.4149	0.3916	0.4012
	OutlierRatio	0.5156	0.875	0.474	0.6875	0.5521	0.6094	0.6458	0.8333	0.7083	0.5677	0.5208	0.5417	0.5365
MOS-LF	SROCC	0.9343	0.5201	0.9333	0.758	0.9293	0.9175	0.8452	0.4383	0.7729	0.9278	0.9223	0.9378	0.9379
	KRCC	0.7824	0.3782	0.7801	0.5672	0.7787	0.7475	0.6742	0.3004	0.5725	0.7676	0.7673	0.7878	0.7878
	PCC_NoFit	0.8609	0.5042	0.9293	0.7751	0.9249	0.915	0.847	0.4528	0.7476	0.9229	0.9131	0.9058	0.9031
	PCC_Fitted	0.9309	0.5266	0.9302	0.778	0.9291	0.9166	0.8511	0.4614	0.7753	0.9289	0.9168	0.9338	0.9361
	RMSE	0.3808	0.8862	0.3827	0.655	0.3856	0.4169	0.5473	0.9249	0.6584	0.3861	0.4162	0.3731	0.3668
	OutlierRatio	0.5469	0.9167	0.5104	0.7448	0.526	0.5781	0.7292	0.7969	0.6979	0.5573	0.5104	0.5417	0.5417
MOS-2D	SROCC	0.9312	0.5496	0.9284	0.8154	0.9276	0.9204	0.8655	0.4161	0.8034	0.922	0.9178	0.9344	0.9314
	KRCC	0.7737	0.3923	0.773	0.6315	0.776	0.7523	0.6996	0.2855	0.6015	0.7569	0.7601	0.7804	0.7771
	PCC_NoFit	0.875	0.5455	0.9304	0.7988	0.9299	0.9166	0.8611	0.443	0.7782	0.923	0.9104	0.8891	0.8827
	PCC_Fitted	0.9334	0.5617	0.9312	0.8162	0.9341	0.9226	0.8697	0.4435	0.8069	0.9265	0.9275	0.9357	0.9345
	RMSE	0.3966	0.9143	0.4028	0.6385	0.3945	0.4262	0.5454	0.9905	0.6528	0.4158	0.4132	0.3898	0.3933
	OutlierRatio	0.5417	0.875	0.4688	0.6667	0.474	0.5573	0.6406	0.8333	0.7344	0.5469	0.4688	0.5156	0.526

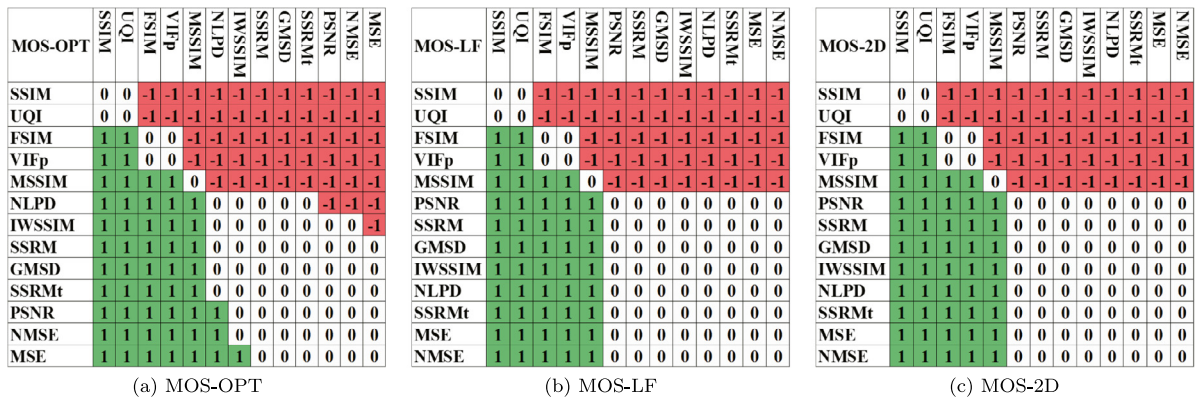


Fig. 7. Statistical significance tables for the evaluation of quality metrics on the reconstructed Fourier holograms. The statistics are separately shown based on the MOS scores obtained from optical holographic display (OPT) (a), the light field display (LF) (b) and the regular 2D display (2D) (c) and the IQM scores.

Solely the NLPD can be considered an outlier as it significantly drops in performance after the denoising step. One reason for this can be the sensitivity of the NLPD to the modifications on the image gradient which is exactly what a Wigner filter would flatten out and hence potentially directly impacting the prediction performance of NLPD. Although, the exact reason for such behaviour is not known to us. The statistical significance tables in Fig. 8 also emphasize our observations.

3.3. Global analysis

After reporting the results in each track, in this section a cross-track analysis of the IQM performances is provided. To do so, we present the overall scatter plots of quality predictions for each IQM in Fig. 10. For each IQM, their quality predictions for QA_1 to QA_4 vs the MOS of the holographic setup are shown together. The scattered points from each test track are accompanied with a logistic fit curve calculated as explained in Section 2.5. The results from the QA_1 to QA_4 are colour coded with blue, orange, green, and yellow, respectively. Here, in the cases of MSE 10(d), NMSE 10(g), PSNR 10(h), and VIFp 10(m), graphs reveal a distinctive shift when the measurements repeat in different steps of the holographic processing pipeline, i.e. these IQMs show a similar behaviour overall but in a shifted score range for each test track. For the case of PSNR, MSE, and VIFp, these data-range shifts occur exactly with the same order of the test tracks progressions. While these plots show that measures like the MSE family in general exhibit a similar behaviour independent of the point in the processing pipeline, where they have been calculated, these range shifts imply that their values are strictly comparable only with the measurements performed in the same place. As an example, a PSNR of 30 dB does not necessarily correspond to the same visual quality when the measurements are done once on the hologram and once on the denoised reconstruction. For the

IQMs which are designed to always provide a bounded measurement this is usually not an issue, though their overall behaviour compared to the MOS can drastically change depending on the QA step where they are used for quality prediction, e.g. in case of SSRM and SSIM. Overall, there seems to be a trade-off between generality and comparability. One should decide between having an unbounded measurement exclusively comparable with the measurements in the same test point but with stable behaviour across the processing pipeline, or having a bounded measurement but with varying behaviour dependent on where is used to predict the visual quality.

Next, we want to have a simple quantitative evaluation of the performance of each IQM w.r.t. all of the benchmarked criteria and compared to all three sets of MOS. To do so, within each row of the Tables 2–5, we ranked the IQMs (e.g. in Table 2, based on the SROCC criterion SSIM has the lowest rank so receives the highest score of 15 out of the 15 tested IQMs and on the other side UQI receives score of 1 for being the worst IQM w.r.t. the SROCC). Thus for each table, every IQM receives 18 ranking scores for the 6 evaluation criteria (i.e. SROCC, KRCC, PCC_NoFit, PCC_Fitted, RMSE and Outlier Ratio) and all 3 sets of MOS. Then, we calculated a column-wise sum of these ranking scores per IQM. That way, we obtained a compact indicator of the IQM performance per test track for which the results are depicted in the bar charts of Fig. 11. We could continue summing the ranking scores across all 4 test tracks, though due to significant changes in the rankings of the IQMs for each test track compared to others, we prefer to provide the ranking scores for each test track separately. The bar charts of Fig. 11 reveal the top performing IQMs in each track. Moreover, the ranking scores of the IQMs helps to see their relative performance compared to each other. For example, in Fig. 11(a), the plateau of the ranking scores from FSIM till VIFp shows that these IQMs w.r.t. all evaluation criteria are almost equally unreliable compared to the top three IQMs.

Table 5

Statistical evaluation of quality metrics for the reconstructed Fourier holograms after speckle noise removal. The statistics based on the MOS scores obtained from optical holographic display (OPT), the light field display (LF) and the regular 2D display (2D) and the IQM scores are separately shown.

	PSNR	SSIM	SSRM	FSIM	GMSD	IWSSIM	MS-SSIM	UQI	VIFp	NLPD	SSRMt	MSE	NMSE	
MOS-OPT	SROCC	0.919	0.675	0.9272	0.9011	0.9022	0.8867	0.8486	0.8926	0.7516	0.9261	0.919	0.9217	
	KRCC	0.7464	0.4974	0.7661	0.7144	0.7181	0.7256	0.6999	0.6472	0.7144	0.7618	0.7457	0.7473	
	PCC_NoFit	0.8308	0.6702	0.9162	0.8819	0.8797	0.8861	0.8589	0.7434	0.7443	0.7228	0.921	0.8852	0.8845
	PCC_Fitted	0.9208	0.6719	0.93	0.9035	0.9006	0.9032	0.8812	0.8369	0.8932	0.7343	0.9223	0.9222	0.9238
	RMSE	0.4383	0.8326	0.4133	0.4817	0.4887	0.4824	0.5314	0.6152	0.5054	0.7631	0.4343	0.4347	0.4303
	OutlierRatio	0.5469	0.7396	0.5156	0.6563	0.599	0.5521	0.5833	0.651	0.6198	0.75	0.5156	0.5417	0.5417
MOS-LF	SROCC	0.9271	0.7145	0.9282	0.9017	0.9211	0.9136	0.901	0.8647	0.885	0.7235	0.9069	0.9283	0.9313
	KRCC	0.7681	0.5413	0.7713	0.7217	0.7629	0.7452	0.7348	0.6781	0.6992	0.532	0.7458	0.7699	0.7718
	PCC_NoFit	0.8801	0.7315	0.9241	0.8993	0.9102	0.9094	0.8908	0.8107	0.8031	0.7109	0.8983	0.8748	0.8729
	PCC_Fitted	0.9261	0.7338	0.9243	0.907	0.9224	0.9145	0.9021	0.8613	0.8813	0.7427	0.9018	0.927	0.9296
	RMSE	0.3934	0.7083	0.3978	0.4391	0.4027	0.4218	0.4499	0.5297	0.4927	0.6981	0.4504	0.391	0.3842
	OutlierRatio	0.6198	0.849	0.5365	0.6771	0.4896	0.5521	0.6302	0.6406	0.6458	0.7135	0.5677	0.6094	0.5781
MOS-2D	SROCC	0.9306	0.7532	0.9262	0.9312	0.9014	0.9225	0.9107	0.8788	0.8924	0.722	0.9033	0.9312	0.9307
	KRCC	0.776	0.5743	0.7688	0.7733	0.7325	0.7597	0.749	0.6967	0.7044	0.5325	0.7365	0.7786	0.7771
	PCC_NoFit	0.8997	0.7703	0.9283	0.918	0.8896	0.9141	0.8918	0.8464	0.8268	0.7286	0.8989	0.8552	0.8506
	PCC_Fitted	0.9354	0.7704	0.9302	0.9395	0.9106	0.9277	0.9157	0.8911	0.8952	0.7831	0.9151	0.9353	0.9354
	RMSE	0.3909	0.7046	0.4057	0.3786	0.4567	0.4125	0.4442	0.5015	0.4925	0.6873	0.4456	0.3911	0.3908
	OutlierRatio	0.4896	0.7292	0.4427	0.5156	0.6042	0.5	0.5625	0.6302	0.6823	0.724	0.5052	0.4688	0.474

MOS-OPT	SSIM	NLPD	UQI	MSSIM	FSIM	GMSD	IWSSIM	VIFp	SSRMt	PSNR	SSRM	MSE	SSRM	NMSE
SSIM	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
NLPD	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
UQI	1	1	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
MSSIM	1	1	1	0	0	0	0	0	-1	-1	-1	-1	-1	-1
FSIM	1	1	1	0	0	0	0	0	0	-1	-1	-1	-1	-1
GMSD	1	1	1	0	0	0	0	0	0	0	0	0	-1	-1
IWSSIM	1	1	1	0	0	0	0	0	0	0	0	0	-1	-1
VIFp	1	1	1	0	0	0	0	0	0	0	0	0	0	-1
SSRMt	1	1	1	0	0	0	0	0	0	0	0	0	0	0
PSNR	1	1	1	0	0	0	0	0	0	0	0	0	0	0
SSRM	1	1	1	1	0	0	0	0	0	0	0	0	0	0
MSE	1	1	1	1	0	0	0	0	0	0	0	0	0	0
NMSE	1	1	1	1	0	0	0	0	0	0	0	0	0	0
SSRM	1	1	1	1	1	1	1	1	1	1	1	1	1	1

(a) MOS-OPT

MOS-LF	SSIM	NLPD	UQI	VIFp	MSSIM	FSIM	GMSD	IWSSIM	SSRMt	PSNR	SSRM	MSE	SSRM	NMSE
SSIM	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
NLPD	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
UQI	1	1	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
VIFp	1	1	0	0	0	0	0	0	-1	-1	-1	-1	-1	-1
FSIM	1	1	0	0	0	0	0	0	0	0	0	-1	-1	-1
GMSD	1	1	0	0	0	0	0	0	0	0	0	0	0	0
IWSSIM	1	1	0	0	0	0	0	0	0	0	0	0	0	0
SSRMt	1	1	1	0	0	0	0	0	0	0	0	0	0	0
PSNR	1	1	1	0	0	0	0	0	0	0	0	0	0	0
SSRM	1	1	1	1	0	0	0	0	0	0	0	0	0	0
MSE	1	1	1	1	0	0	0	0	0	0	0	0	0	0
NMSE	1	1	1	1	0	0	0	0	0	0	0	0	0	0

(b) MOS-LF

MOS-2D	SSIM	NLPD	UQI	VIFp	GMSD	IWSSIM	MSSIM	SSRMt	PSNR	SSRM	MSE	SSRM	NMSE
SSIM	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
NLPD	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
UQI	1	1	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1
VIFp	1	1	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1
GMSD	1	1	0	0	0	0	0	0	0	0	-1	-1	-1
IWSSIM	1	1	1	0	0	0	0	0	0	0	0	0	0
MSSIM	1	1	1	0	0	0	0	0	0	0	0	0	0
SSRMt	1	1	1	0	0	0	0	0	0	0	0	0	0
PSNR	1	1	1	0	0	0	0	0	0	0	0	0	0
SSRM	1	1	1	1	0	0	0	0	0	0	0	0	0
FSIM	1	1	1	1	0	0	0	0	0	0	0	0	0
MSE	1	1	1	1	0	0	0	0	0	0	0	0	0
NMSE	1	1	1	1	0	0	0	0	0	0	0	0	0

(c) MOS-2D

Fig. 8. Statistical significance tables for the evaluation of quality metrics on the reconstructed Fourier holograms after speckle denoising. The statistics based on the MOS scores obtained from optical holographic display (OPT) (a), the light field display (LF) (b) and the regular 2D display (2D) (c) and the IQM scores, are separately shown.

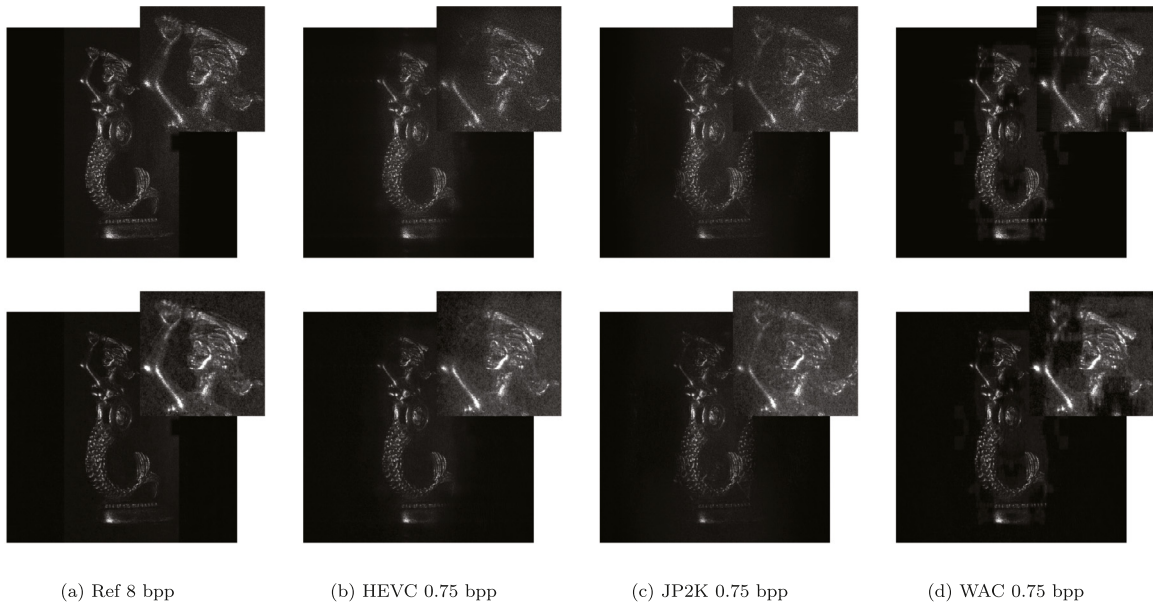


Fig. 9. Exemplary numerical reconstruction of OR-Mermaid from centre view. First row: Reference, encoded with HEVC, JP2K and WAC at 0.75 bpp, before speckle denoising(QA_3). Second row: same as first row, after speckle denoising(QA_4).

3.4. Final verdict

Our experiments for the first time provide a full view of the performance of the available quality measures for the case of digital holography w.r.t. MOS. We visualized the behaviour of each method

when tested across the four test tracks (Fig. 10) and quantified the performance of these quality measures based on several evaluation criteria (See Fig. 11). Considering all aspects tested in this research, we would like to finalize the analysis of the tested IQMs by providing a usage guideline such that helps interested user to choose the

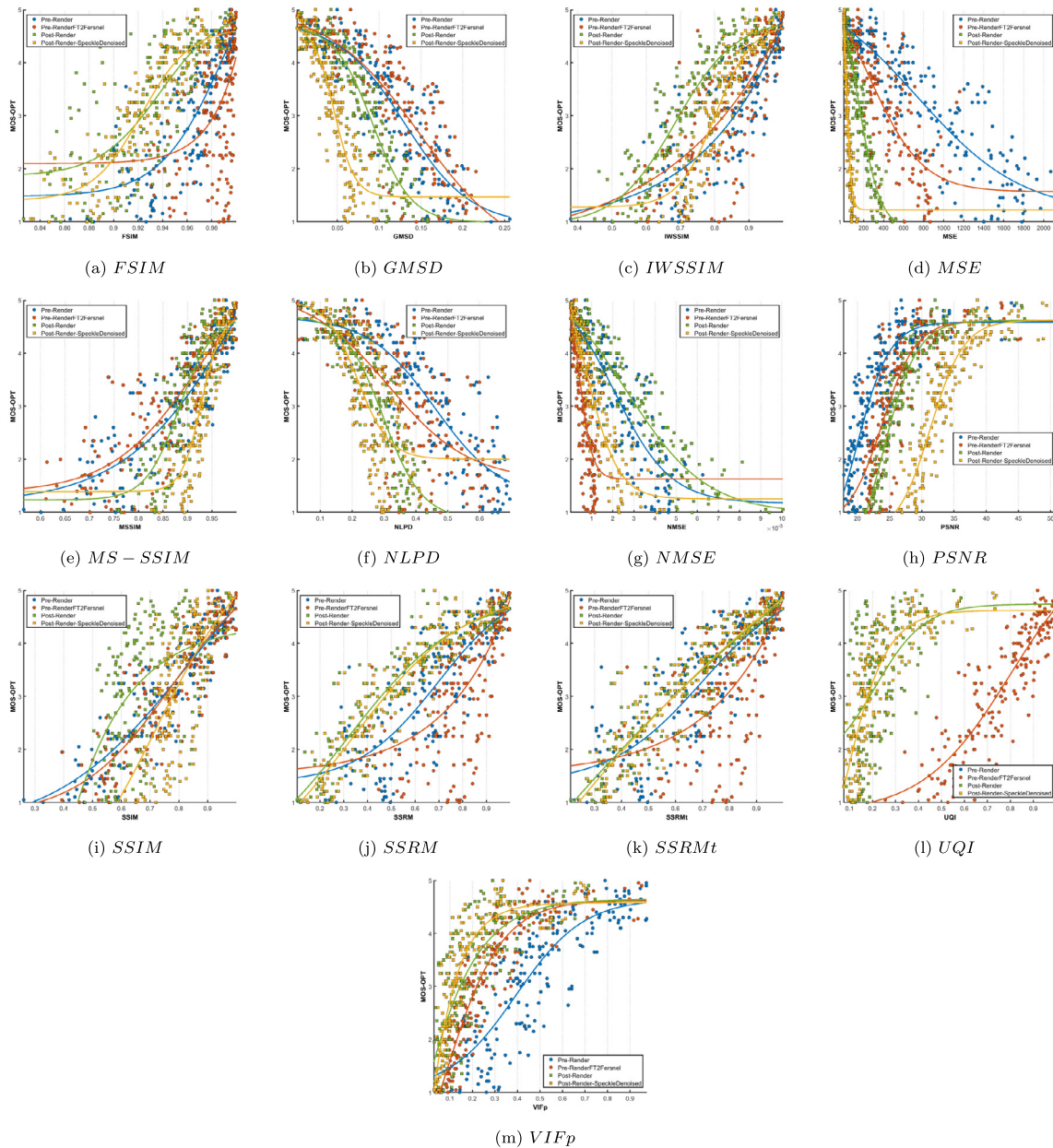


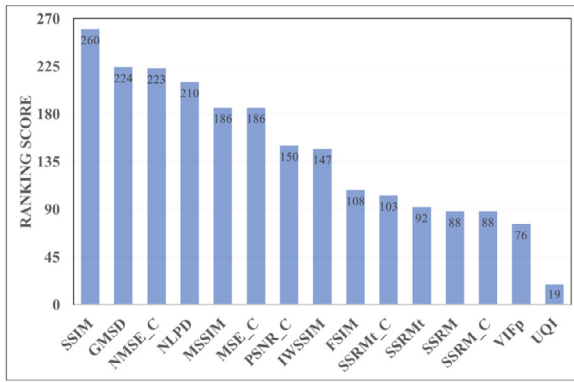
Fig. 10. Overall scatter plots of the quality metric performances for each of the four experimental tracks based on the MOS obtained from the holographic display setup (MOS-OPT). The scores from the centre view and the right corner view are combined and shown together per experimental track. Each data point for the two experimental tracks after reconstruction, represents the averaged MOS scores and predicted quality scores obtained from different focal distances per object. The results of the test tracks QA_1 to QA_4 are colour coded with blue, orange, green, and yellow, respectively.

most appropriate method for the holographic VQA tasks until more hologram-oriented quality metrics are designed. For each testing track, Table 6 provides our recommendations where the recommendations are categorized in three groups colour coded with green: usage is advised, light-blue: Neutral or results were not conclusive to provide a strong recommendation and red: the application is discouraged.

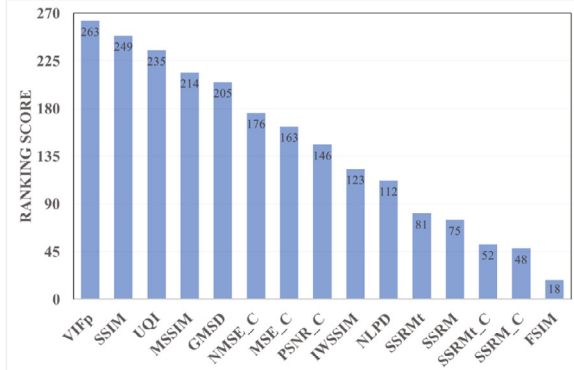
According to our experimental results, MSE and NMSE measurements, which are purely based on signal fidelity and do not take into account any perceptual aspect, did not show any major drawback and in fact are among the top three recommended methods, especially when measured after the reconstruction. PSNR follows almost the same performance across all testing tracks though it sometimes records a slightly poorer correlation with respect to the MOS scores. This might be particularly the case because the logarithmic scale of the PSNR fits well with the behaviour of the HVS when applied to the reconstruction plane, this connection is lost when applied to the hologram and the

logarithmic scale becomes without subject when compared to the HVS response measured on the reconstructed data. Moreover, please note that the MSE, NMSE and PSNR correlations with the MOS scores are higher in the reconstruction plane.

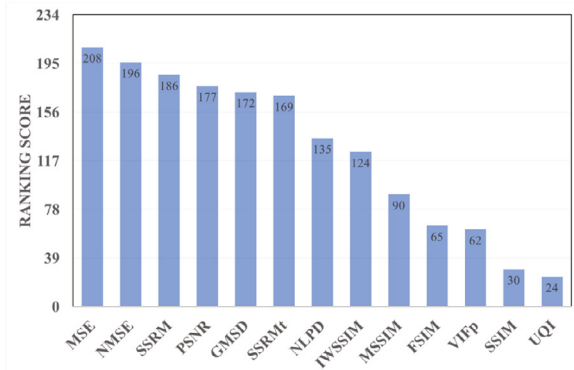
The SSIM and SSRM appear to demonstrate the most dramatic behaviour across all tracks. As mentioned earlier, while SSIM is found to be the top performing IQM prior to, and the worst after reconstruction, SSRM performs exactly the opposite. It performs poorly on holograms and jumps to the top of the list when tested on the reconstructed scenes. The SSRMt also demonstrates similar behaviour as the vanilla SSRM, which suggests a deeper connection between their identical operational core and their performance. Regarding the other members of the SSIM family (e.g. MS-SSIM, IWSSIM) though, it appears that application of more complex perceptual models dropped their performance in the hologram domain. After the reconstruction, their multi-scale analysis based measurements could only create an improvement on their



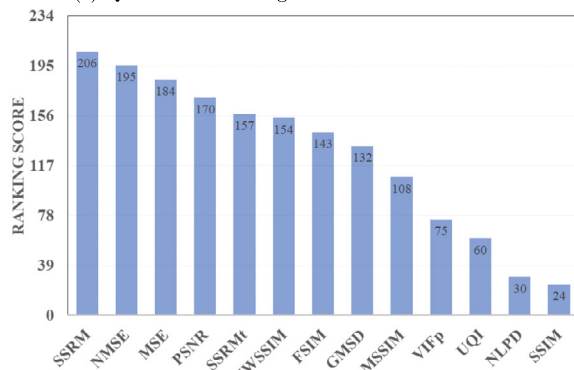
(a) QA.1: Fourier holograms prior to reconstruction



(b) QA.2: Fresnel holograms prior to reconstruction



(c) QA.3: Fourier holograms after reconstruction



(d) QA.4: Fourier holograms after reconstruction and denoising

Fig. 11. Total sum of the ranking scores of the IQMs ranked per row in the correlation tables Tables 2–5. The overall performance IQMs in each test track is summarized based on all the 6 evaluation criteria and the 3 sets of MOS used for the benchmarking.

performance over the default SSIM but not adequate to make them competitive. The case of IWSSIM can also be considered in conjunction

with the NLPD. Both of these methods benefit from the multi-resolution decomposition using Laplacian pyramid and both present a similar performance across our evaluation pipeline suggesting that such a decomposition may not necessarily be the most effective approach for VQA in digital holography. The other SSIM family member UQI, which does not benefit from the regularization in its operational core, is expected to perform worse than the more mathematically stable SSIM. Lack of regularization for this metric is known to cause instability especially in cases where the sum of squared means or sum of variances for the reference and distorted data gets close to zero. For this reason, we discourage its usage over the SSIM and in case of Fresnel holograms where it shows a competitive performance, we recommend utilizing it along with secondary quality prediction methods.

Also, the image-gradient based methods – which are normally able to make very accurate quality predictions in natural imagery – do not show an outstanding performance (e.g. FSIM, GMSD) suggesting that such approaches may not necessarily be effective for VQA in digital holography. However, the rather stronger performance of GMSD could be related to utilizing the standard deviation of the local gradient magnitude comparisons which enables a rather more stable prediction with more independence from the noisy nature of holograms and the speckle noise.

The VIFp which represents the information theoretic approach to the VQA and is been a long standing contender in the realm of IQA, also does not exhibit a reliable achievement across all of our experimental tracks. Only in one case when the Fresnel holograms are tested, outperforms others in most cases. This method exhibits a number of assumptions, which may or may not hold in case of holography causing large fluctuations in its performance. For example, it employs the natural image statistics by using the Gaussian Scale Mixture (GSM) to statistically model the wavelet coefficients after applying a steerable pyramid decomposition. This model is expected to substantially deviate from the statistics of the holographic signals. However, VIFp also models the distortion channel using an attenuating additive noise in wavelet domain which in this case potentially can relate well with the highly noisy nature of holograms. On top of them, in VIFp it is assumed that while passing through the HVS, the uncertainty level increases for perception of visual data. Thus the wavelet subbands for both the source and distortion channels are subject to an extra additive white Gaussian noise which models such increase in the uncertainty level. Now, wherever all these assumptions happen to fall in-line with the statistics of the tested holograms, a good prediction performance is foreseen. However, in most cases it is going to be the other way around. A re-adjustment on these assumptions based on the holographic statistical properties may potentially solve the unreliability issue for information theoretic based metrics like VIFp. Although, because of extremely noisy nature of holographic interference fringes and their strong dependence to the properties of the recorded scene, such statistical modelling may not be generalizable. We have recommended the usage of VIFp for the Fresnel holograms in Table 6. However, we strongly advise to use this metric along with other recommended IQMs due to its instability caused by the above explained reasons.

It should be noted that, these recommendations hold only within the bounds of current experiment (i.e. as long as the degradations are limited to compression artefacts, and if the same procedure as this research is utilized for processing, rendering and quality assessment of the holograms). Moreover, being the best among themselves does not necessarily translate these IQMs into being reliably accurate in predicting the visual quality of holograms. This gets more clear, when the achieved correlations w.r.t. MOS in this research is compared to the ones of the same IQMs tested on digital images. Especially, the achieved correlations before the reconstruction (reported in Tables 2 and 3), still have lots of room for improvement. Not to mention, our results showed that careful considerations should be taken into account for choosing the right quality prediction method among the available options based on the application and the characteristics of the holograms at the point

Table 6

Usage recommendation table of tested IQMs colour coded based on their overall performance in each test track. Green: advised to be utilized, Light blue: Neutral or results not conclusive within the bounds of this research and Red: usage is discouraged.

IQM	Fourier hologram (QA_1)	Fresnel hologram (QA_2)	Reconstructed hologram (QA_3)	Reconstructed hologram + speckle denoising (QA_4)
MSE				
NMSE				
PSNR				
SSRM				
SSRMt				
SSIM				
IWSSIM				
MS-SSIM				
UQI				
GMSD				
FSIM				
NLPD				
VIFp				

where the measurement occurs. Therefore, in spite of the provided recommendations, we emphasize that none of the available options are at the level that confidently alleviate the strong need for design and development of essential quality assessment algorithms for the holographic data.

4. Conclusions

In this research for the first time we utilized 96 holograms of HoloDB to make a systematic performance evaluation of the most viable options for the quality assessment of digital holograms. The results for 4 separate sets of evaluations are provided. We compared the reference holograms with the distorted versions: before rendering, on the real and imaginary parts of the complex-wavefield; after converting the Fourier holograms to regular Fresnel holograms; after rendering, on the quantized amplitude of the reconstructed data, and after speckle denoising. For every experimental track, the quality metric predictions are rigorously compared to three sets of MOS which were previously obtained via subjective experiments. Statistical analysis of IQM performances and a discussion on the behaviour of outstanding methods are presented. Finally their overall performance based on all of the utilized evaluation criteria and all three sets of MOS are summarized per test track to introduce the best performing quality metrics for each testing track. All aspects considered, turns out while for each test track, a couple of quality metrics present a significantly correlated performance compared to the multiple sets of available MOS, none of them show a consistently high-performance across all the four test-tracks. This emphasizes their sensitivity to the characteristics of the input and the position in the processing pipeline where the quality predictions are obtained. It also reveals once more the dire need to design efficient quality metrics for holographic content. We are hoping that this research provide a thorough understanding of the complexities involved in VQA of digital holograms and consequently act as the first systematic step towards designing advanced perceptual quality prediction methods for digital holograms.

CRedit authorship contribution statement

Ayyoub Ahar: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Tobias Birnbaum:** Conceptualization, Methodology, Software, Validation, Data curation, Writing - original draft, Writing - review & editing. **Maksymilian Chlipala:** Resources, Validation, Investigation, Writing - original draft. **Weronika Zaperty:** Resources, Validation. **Saeed Mahmoudpour:** Writing - review & editing. **Tomasz Kozacki:** Resources, Validation, Investigation, Writing - original draft, Supervision, Project administration, Funding acquisition. **Malgorzata Kujawska:** Resources, Validation, Investigation, Writing - original draft, Supervision, Project administration,

Funding acquisition. **Peter Schelkens:** Conceptualization, Methodology, Validation, Investigation, Writing - original draft, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC Grant Agreement n. 617779 (INTERFERE) and also the Cross-Ministry Giga KOREA Project (GigaKOREA GK20D0100) and the support from the WUT, China.

References

- [1] S. Li, Y. Ding, Y. Chang, No-reference stereoscopic image quality assessment based on cyclopean image and enhanced image, *Signal, Image Video Proc.* (2019) 1–9.
- [2] Y. Fan, M.-C. Larabi, F.A. Cheikh, C. Fernandez-Maloigne, No-reference quality assessment of stereoscopic images based on binocular combination of local features statistics, in: 2018 25th IEEE International Conference on Image Processing (ICIP), IEEE, 2018, pp. 3538–3542.
- [3] W. Zhou, Z. Chen, W. Li, Dual-stream interactive networks for no-reference stereoscopic image quality assessment, *IEEE Trans. Image Process.* (2019).
- [4] J. Xu, W. Zhou, Z. Chen, S. Ling, P.L. Callet, Predictive auto-encoding network for blind stereoscopic image quality assessment, 2019, *ArXiv Preprint ArXiv:1909.01738*.
- [5] F. Battisti, E. Bosc, M. Carli, P. Le Callet, S. Perugia, Objective image quality assessment of 3D synthesized views, *Signal Process., Image Commun.* 30 (2015) 78–88.
- [6] K. Gu, J. Qiao, S. Lee, H. Liu, W. Lin, P. Le Callet, Multiscale natural scene statistical analysis for no-reference quality evaluation of DIBR-synthesized views, *IEEE Trans. Broadcast.* (2019).
- [7] S. Ling, J. Li, Z. Che, X. Min, G. Zhai, P.L. Callet, Quality assessment of free-viewpoint videos by quantifying the elastic changes of multi-scale motion trajectories, 2019, *ArXiv Preprint ArXiv:1903.12107*.
- [8] Y. Zhou, L. Li, S. Wang, J. Wu, Y. Fang, X. Gao, No-reference quality assessment for view synthesis using dog-based edge statistics and texture naturalness, *IEEE Trans. Image Process.* (2019).
- [9] D.D. Sandić-Stanković, D.D. Kukolj, P. Le Callet, Fast blind quality assessment of DIBR-synthesized video based on high-high wavelet subband, *IEEE Trans. Image Process.* 28 (11) (2019) 5524–5536.
- [10] P. Paudyal, F. Battisti, M. Carli, Reduced reference quality assessment of light field images, *IEEE Trans. Broadcast.* 65 (1) (2019) 152–165.
- [11] Y. Fang, K. Wei, J. Hou, W. Wen, N. Imamoglu, Light field image quality assessment by local and global features of epipolar plane image, in: 2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM), IEEE, 2018, pp. 1–6.
- [12] W. Zhou, L. Shi, Z. Chen, Tensor oriented no-reference light field image quality assessment, 2019, *ArXiv Preprint ArXiv:1909.02358*.
- [13] P.A. Kara, R.R. Tamboli, O. Doronin, A. Cserkaszky, A. Barsi, Z. Nagy, M.G. Martini, A. Simon, The key performance indicators of projection-based light field visualization, *J. Inform. Display* 20 (2) (2019) 81–93.
- [14] L. Shi, W. Zhou, Z. Chen, J. Zhang, No-reference light field image quality assessment based on spatial-angular measurement, *IEEE Trans. Circuits Syst. Video Technol.* (2019).
- [15] D. Blinder, A. Ahar, S. Bettens, T. Birnbaum, A. Symeonidou, H. Ottevaere, C. Schretter, P. Schelkens, Signal processing challenges for digital holographic video display systems, *Signal Process., Image Commun.* 70 (2018) 114–130, <http://dx.doi.org/10.1016/j.image.2018.09.014>, URL: <http://www.sciencedirect.com/science/article/pii/S0923596518304855>.
- [16] P. Schelkens, T. Ebrahimi, A. Gilles, P. Gioia, K.-J. Oh, F. Pereira, C. Perra, A. Pinheiro, JPEG Pleno: Providing representation interoperability for holographic applications and devices, *ETRI J.* (2019) 93–108, <http://dx.doi.org/10.4218/etrij.2018-0509>.
- [17] A. Ahar, T. Birnbaum, C. Jaeh, P. Schelkens, A new similarity measure for complex valued data, in: *Digital Holography and Three-Dimensional Imaging*, Optical Society of America, 2017, pp. Tu1A–6.

- [18] A. Ahar, T. Birnbaum, C. Jaeh, P. Schelkens, A new similarity measure for complex amplitude holographic data, in: *Applications of Digital Image Processing XL*, 10396, International Society for Optics and Photonics, 2017, p. 1039611.
- [19] M.V. Bernardo, P. Fernandes, A. Arrifano, M. Antonini, E. Fonseca, P.T. Fiadeiro, A.M. Pinheiro, M. Pereira, Holographic representation: Hologram plane vs. object plane, *Signal Process., Image Commun.* 68 (2018) 193–206, <http://dx.doi.org/10.1016/j.image.2018.08.006>, URL: <http://www.sciencedirect.com/science/article/pii/S0923596518300407>.
- [20] D. Blinder, A. Ahar, A. Symeonidou, Y. Xing, T. Bruylants, C. Schretter, B. Pesquet-Popescu, F. Dufaux, A. Munteanu, P. Schelkens, Open access database for experimental validations of holographic compression engines, in: 2015 Seventh International Workshop on Quality of Multimedia Experience (Qomex), Ieee, New York, 2015, pp. 1–6, WOS:000375091800066.
- [21] T.M. Lehtimäki, K. Sääskilähti, R. Näsänen, T.J. Naughton, Visual perception of digital holograms on autostereoscopic displays, in: *Three-Dimensional Imaging, Visualization, and Display 2009*, 7329, International Society for Optics and Photonics, 2009, p. 73290C.
- [22] T.M. Lehtimäki, K. Sääskilähti, T. Pitkäaho, T.J. Naughton, Evaluation of perceived quality attributes of digital holograms viewed with a stereoscopic display, in: 2010 9th Euro-American Workshop on Information Optics, IEEE, 2010, pp. 1–3.
- [23] T.M. Lehtimäki, M. Niemelä, R. Näsänen, R.G. Reilly, T.J. Naughton, Using traditional glass plate holograms to study visual perception of future digital holographic displays, in: *Imaging and Applied Optics 2016*, Optical Society of America, 2016, p. JW4A.20, <http://dx.doi.org/10.1364/3D.2016.JW4A.20>.
- [24] A. Ahar, D. Blinder, T. Bruylants, C. Schretter, A. Munteanu, P. Schelkens, Subjective quality assessment of numerically reconstructed compressed holograms, in: *Applications of Digital Image Processing XXXVIII*, 9599, SPIE, United States, 2015, <http://dx.doi.org/10.1117/12.2189887>, 95990K–95990K–15.
- [25] A. Symeonidou, D. Blinder, B. Ceulemans, A. Munteanu, P. Schelkens, Three-dimensional rendering of computer-generated holograms acquired from point-clouds on light field displays, in: A.G. Tescher (Ed.), *SPIE Optics + Photonics, Applications of Digital Image Processing XXXIX*, 9971, Spie-Int Soc Optical Engineering, Bellingham, 2016, p. 99710S, WOS:000390023100024.
- [26] A. Symeonidou, D. Blinder, P. Schelkens, Colour computer-generated holography for point clouds utilizing the phong illumination model, *Opt. Express* 26 (8) (2018) 10282–10298, <http://dx.doi.org/10.1364/OE.26.10282>, URL: <http://www.opticsexpress.org/abstract.cfm?URI=oe-26-8-10282>.
- [27] A. Ahar, M. Chlipala, T. Birnbaum, W. Zaperty, A. Symeonidou, T. Kozacki, M. Kujawinska, P. Schelkens, Suitability analysis of holographic vs light field and 2D displays for subjective quality assessment of fourier holograms, *Opt. Express* 28 (24) (2020) 37069–37091, <http://dx.doi.org/10.1364/OE.405984>, URL: <http://www.opticsexpress.org/abstract.cfm?URI=oe-28-24-37069>.
- [28] T. Kozacki, M. Chlipala, P.L. Makowski, Color fourier orthoscopic holography with laser capture and an led display, *Opt. Express* 26 (9) (2018) 12144–12158, <http://dx.doi.org/10.1364/OE.26.012144>, URL: <http://www.opticsexpress.org/abstract.cfm?URI=oe-26-9-12144>.
- [29] Holografika, Holovizio 722RC, 2019 (accessed Feb 15, 2019), <http://holografika.com/722rc/>.
- [30] EIZO, Coloredge CG318-4k, 2019 (accessed Feb 15, 2019), <https://www.eizoglobal.com/products/coloredge/cg318-4k/index.html>.
- [31] A. Ahar, M. Chlipala, T. Birnbaum, W. Zaperty, A. Symeonidou, T. Kozacki, M. Kujawinska, P. Schelkens, Digital holography database for multi-perspective multi-display subjective experiment(holodb), 2019, <http://ds.erc-interfere.eu/holodb/>.
- [32] J.W. Goodman, *Speckle Phenomena in Optics: Theory and Applications*, Roberts and Company Publishers, 2007.
- [33] V. Bianco, P. Memmolio, M. Leo, S. Montesor, C. Distante, M. Paturzo, P. Picart, B. Javidi, P. Ferraro, Strategies for reducing speckle noise in digital holography, *Light: Sci. Appl.* 7 (1) (2018) 48.
- [34] T. Birnbaum, A. Ahar, S. Montrésor, P. Picart, C. Schretter, P. Schelkens, Speckle denoising of computer-generated macroscopic holograms, in: *Digital Holography and Three-Dimensional Imaging 2019*, Optical Society of America, 2019, p. W3A.1, <http://dx.doi.org/10.1364/DH.2019.W3A.1>, URL: <http://www.osapublishing.org/abstract.cfm?URI=DH-2019-W3A.1>.
- [35] E. Fonseca, P.T. Fiadeiro, M.V. Bernardo, A. Pinheiro, M. Pereira, Assessment of speckle denoising filters for digital holography using subjective and objective evaluation models, *Appl. Opt.* 58 (34) (2019) G282–G292.
- [36] A. El Rhammad, P. Gioia, A. Gilles, A. Cagnazzo, B. Pesquet-Popescu, Color digital hologram compression based on matching pursuit, *Appl. Opt.* 57 (17) (2018) 4930–4942, <http://dx.doi.org/10.1364/AO.57.004930>, WOS:000434872300027.
- [37] J.P. Peixeiro, C. Brites, J. Ascenso, F. Pereira, Holographic data coding: Benchmarking and extending HEVC with adapted transforms, *IEEE Trans. Multimed.* 20 (2) (2018) 282–297, <http://dx.doi.org/10.1109/TMM.2017.2742701>.
- [38] M.V. Bernardo, P. Fernandes, A. Arrifano, M. Antonini, E. Fonseca, P.T. Fiadeiro, A.M.G. Pinheiro, M. Pereira, Holographic representation: Hologram plane vs. object plane, *Signal Process., Image Commun.* 68 (2018) 193–206, <http://dx.doi.org/10.1016/j.image.2018.08.006>, URL: <http://www.sciencedirect.com/science/article/pii/S0923596518300407>.
- [39] A. Symeonidou, D. Blinder, A. Munteanu, P. Schelkens, Computer-generated holograms by multiple wavefront recording plane method with occlusion culling, *Opt. Express* 23 (17) (2015) 22149–22161, <http://dx.doi.org/10.1364/OE.23.022149>, URL: <http://www.opticsexpress.org/abstract.cfm?URI=oe-23-17-22149>.
- [40] J.-H. Park, Recent progress in computer-generated holography for three-dimensional scenes, *J. Inform. Display* 18 (1) (2017) 1–12, <http://dx.doi.org/10.1080/15980316.2016.1255672>, <https://doi.org/10.1080/15980316.2016.1255672>.
- [41] Y. Pan, J. Liu, X. Li, Y. Wang, A review of dynamic holographic three-dimensional display: Algorithms, devices, and systems, *IEEE Trans. Ind. Inf.* 12 (4) (2016) 1599–1610, <http://dx.doi.org/10.1109/TII.2015.2496304>.
- [42] T. Sugie, T. Akamatsu, T. Nishitsuji, R. Hirayama, N. Masuda, H. Nakayama, Y. Ichihashi, A. Shiraki, M. Oikawa, N. Takada, Y. Endo, T. Kakue, T. Shimobaba, T. Ito, High-performance parallel computing for next-generation holographic imaging, *Nat. Electron.* 1 (4) (2018) 254, <http://dx.doi.org/10.1038/s41928-018-0057-5>, URL: <https://www.nature.com/articles/s41928-018-0057-5>.
- [43] Shimobaba, Tomoyoshi, Ito, Tomoyoshi, *Computer Holography: Acceleration Algorithms and Hardware Implementations*, CRC Press, 2019.
- [44] T. Nishitsuji, T. Shimobaba, T. Kakue, T. Ito, Review of fast calculation techniques for computer-generated holograms with the point light-source-based model, *IEEE Trans. Ind. Inf.* 13 (5) (2017) 2447–2454.
- [45] R. Corda, C. Perra, Hologram domain data compression: Performance of standard codecs and image quality assessment at different distances and perspectives, *IEEE Trans. Broadcast.* 66 (2) (2019) 292–309.
- [46] A.W. Lohmann, R.G. Dorsch, D. Mendlovic, Z. Zalevsky, C. Ferreira, Space-bandwidth product of optical signals and systems, *J. Opt. Soc. Amer. A* 13 (3) (1996) 470–473, <http://dx.doi.org/10.1364/JOSAA.13.000470>.
- [47] J.W. Goodman (Ed.), *Introduction to Fourier Optics*, Roberts & Company, 2004.
- [48] A. Golos, W. Zaperty, G. Finke, P. Makowski, T. Kozacki, Fourier RGB synthetic aperture color holographic capture for wide angle holographic display, in: *Proc. SPIE*, Vol. 9970, 2016, pp. 296–305.
- [49] G. Stroke, Lensless fourier-transform method for optical holography, *Appl. Phys. Lett.* 10 (6) (1965) 201–203.
- [50] D.S. Taubman, M.W. Marcellin, JPEG2000: standard for interactive imaging, *Proc. IEEE* 90 (8) (2002) 1336–1357, <http://dx.doi.org/10.1109/JPROC.2002.800725>.
- [51] P. Schelkens, A. Skodras, T. Ebrahimi, *The JPEG 2000 Suite*, 15, John Wiley & Sons, 2009.
- [52] G.J. Sullivan, J.-R. Ohm, W.-J. Han, T. Wiegand, Overview of the high efficiency video coding (HEVC) standard, *IEEE Trans. Circuits Syst. Video Technol.* 22 (12) (2012) 1649–1668, <http://dx.doi.org/10.1109/TCSVT.2012.2221191>, URL: <http://ieeexplore.ieee.org/document/6316136/>.
- [53] T. Birnbaum, A. Ahar, D. Blinder, C. Schretter, T. Kozacki, P. Schelkens, Wave atoms for lossy compression of digital holograms, in: 2019 Data Compression Conference (DCC), 2019, pp. 398–407, <http://dx.doi.org/10.1109/DCC.2019.00048>.
- [54] T. Birnbaum, A. Ahar, D. Blinder, R.M. Kizhakkumkara, C. Schretter, P. Schelkens, WG1M88009: Wavelet Compression in the Hologram Plane and Associated Aliasing Artifacts After Reconstruction in Object Plane, *Tech. rep.*, JPEG Pleno, 2020.
- [55] R. Corda, C. Perra, A dataset of hologram reconstructions at different distances and viewpoints for quality evaluation, in: 2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX), IEEE, 2019, pp. 1–3.
- [56] P.S. Tobias Birnbaum, Providing a visual understanding of holography through phase space representations, *Appl. Sci.* 10 (14) (4766) (2020) 1–27.
- [57] T. Kozacki, K. Falaggis, Angular spectrum method with compact space-bandwidth: generalization and full-field accuracy, *Appl. Opt.* 55 (19) (2016) 5014–5024, <http://dx.doi.org/10.1364/AO.55.005014>, URL: <http://ao.osa.org/abstract.cfm?URI=ao-55-19-5014>.
- [58] A. Ahar, A. Barri, P. Schelkens, From sparse coding significance to perceptual quality: A new approach for image quality assessment, *IEEE Trans. Image Process.* 27 (2) (2017) 879–893.
- [59] A. Ahar, T. Birnbaum, D. Blinder, A. Symeonidou, P. Schelkens, Performance evaluation of sparseness significance ranking measure (SSRM) on holographic content, in: *Laser Applications To Chemical, Security and Environmental Analysis*, Optical Society of America, 2018, pp. JTU4A–10.
- [60] L. Zhang, L. Zhang, X. Mou, D. Zhang, FSIM: A feature similarity index for image quality assessment, *IEEE Trans. Image Process.* 20 (8) (2011) 2378–2386.
- [61] Z. Wang, Q. Li, Information content weighting for perceptual image quality assessment, *IEEE Trans. Image Process.* 20 (5) (2010) 1185–1198.
- [62] Z. Wang, E.P. Simoncelli, A.C. Bovik, Multiscale structural similarity for image quality assessment, in: *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers*, 2003, 2, Ieee, 2003, pp. 1398–1402.
- [63] H.R. Sheikh, A.C. Bovik, Image information and visual quality, *IEEE Trans. Image Process.* 15 (2) (2006) 430–444.
- [64] V. Laparra, J. Ballé, A. Berardino, E.P. Simoncelli, Perceptual image quality assessment using a normalized Laplacian pyramid, *Electron. Imaging* 2016 (16) (2016) 1–6.
- [65] W. Xue, L. Zhang, X. Mou, A.C. Bovik, Gradient magnitude similarity deviation: A highly efficient perceptual image quality index, *IEEE Trans. Image Process.* 23 (2) (2013) 684–695.

- [66] Z. Wang, A.C. Bovik, A universal image quality index, *IEEE Signal Process. Lett.* 9 (3) (2002) 81–84.
- [67] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, et al., Image quality assessment: from error visibility to structural similarity, *IEEE Trans. Image Process.* 13 (4) (2004) 600–612.
- [68] Y. Wu, Y. Luo, G. Chaudhari, Y. Rivenson, A. Calis, K. de Haan, A. Ozcan, Bright-field holography: cross-modality deep learning enables snapshot 3D imaging with bright-field contrast using a single hologram, *Light: Sci. Appl.* 8 (1) (2019) 25.
- [69] X.-N. Pang, S.-J. Jiang, J.-W. Dong, Dynamic holographic imaging of real-life scene, *Opt. Laser Technol.* 119 (2019) 105590, <http://dx.doi.org/10.1016/j.optlastec.2019.105590>, URL: <http://www.sciencedirect.com/science/article/pii/S0030399219304384>.
- [70] A. Ahar, D. Blinder, T. Bruylants, C. Schretter, A. Munteanu, P. Schelkens, Subjective quality assessment of numerically reconstructed compressed holograms, in: *Applications of Digital Image Processing XXXVIII*, 9599, International Society for Optics and Photonics, 2015, p. 95990K.
- [71] V.Q.E. Group, et al., Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II, 2003, 2003 VQEG.
- [72] P. Hanhart, P. Korshunov, T. Ebrahimi, Benchmarking of quality metrics on ultra-high definition video sequences, in: *2013 18th International Conference on Digital Signal Processing (DSP)*, IEEE, 2013, pp. 1–8.