



OPEN

## ConIQa: A deep learning method for perceptual image quality assessment with limited data

M. Hossein Eybposh<sup>1,2</sup>, Changjia Cai<sup>1,2</sup>, Aram Moossavi<sup>1</sup>, Jose Rodriguez-Romaguera<sup>1,3,4,5</sup>✉ & Nicolas C. Pégard<sup>1,2,5</sup>✉

Effectively assessing the realism and naturalness of images in virtual (VR) and augmented (AR) reality applications requires Full Reference Image Quality Assessment (FR-IQA) metrics that closely align with human perception. Deep learning-based IQAs that are trained on human-labeled data have recently shown promise in generic computer vision tasks. However, their performance decreases in applications where perfect matches between the reference and the distorted images should not be expected, or whenever distortion patterns are restricted to specific domains. Tackling this issue necessitates training a task-specific neural network, yet generating human-labeled FR-IQAs is costly, and deep learning typically demands substantial labeled data. To address these challenges, we developed ConIQa, a deep learning-based IQA that leverages consistency training and a novel data augmentation method to learn from both labeled and unlabeled data. This makes ConIQa well-suited for contexts with scarce labeled data. To validate ConIQa, we considered the example application of Computer-Generated Holography (CGH) where specific artifacts such as ringing, speckle, and quantization errors routinely occur, yet are not explicitly accounted for by existing IQAs. We developed a new dataset, HQA1k, that comprises 1000 natural images each paired with an image rendered using various popular CGH algorithms, and quality-rated by thirteen human participants. Our results show that ConIQa achieves superior Pearson (0.98), Spearman (0.965), and Kendall's tau (0.86) correlations over fifteen FR-IQA metrics by up to 5%, showcasing significant improvements in aligning with human perception on the HQA1k dataset.

Augmented reality (AR) and virtual reality (VR) are cutting-edge technologies that enable immersive experiences in a rapidly growing number of applications including healthcare<sup>1</sup>, video-gaming and entertainment<sup>2</sup>, education<sup>3</sup>, and manufacturing<sup>4</sup>. Computer-generated holography (CGH) has emerged as a promising technology for AR and VR displays due to its unique volume image rendering capabilities<sup>5–9</sup>. With CGH, users can experience a highly immersive environment where virtual objects and scenes are perceived with exceptional realism and accurate depth cues<sup>10</sup>. Additionally, by rendering true 3D objects to each eye, CGH eliminates the vergence-accommodation conflict, a common issue in AR and VR systems based on 2D near-eye display technology that leads to visual discomfort and eye strain<sup>5</sup>. CGH also has the advantage of enabling image renderings matched to the morphology of the human visual system, with both wide field capabilities needed to stimulate the peripheral vision, and high-resolution capabilities adjustable to match the human eye's fovea<sup>11</sup>. Finally, CGH can be implemented in a compact form factor that is practical for portable and wearable devices<sup>8</sup>.

The development of CGH-based displays has sparked significant research efforts to enhance physical models<sup>8,12</sup> and algorithms for image synthesis<sup>6,7,10</sup>. To complement these advances, there is currently a growing need for reliable metrics that quantify the realism and immersiveness of display technologies<sup>11,13,14</sup>. This necessitates the development of advanced Image Quality Assessment (IQA) metrics that align with human perception of image quality. Accurate metrics that can assess the quality of CGH renderings and CGH-specific noise patterns will facilitate the refinement of future CGH algorithms<sup>15</sup> and display technologies and enable more immersive and realistic AR and VR experiences.

<sup>1</sup>Department of Applied Physical Sciences, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA. <sup>2</sup>Joint Department of Biomedical Engineering, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA. <sup>3</sup>Department of Psychiatry, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA. <sup>4</sup>Department of Cell Biology and Physiology, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA. <sup>5</sup>UNC Neuroscience Center, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA. ✉email: jose\_rodriguezromaguera@med.unc.edu; pegard@unc.edu

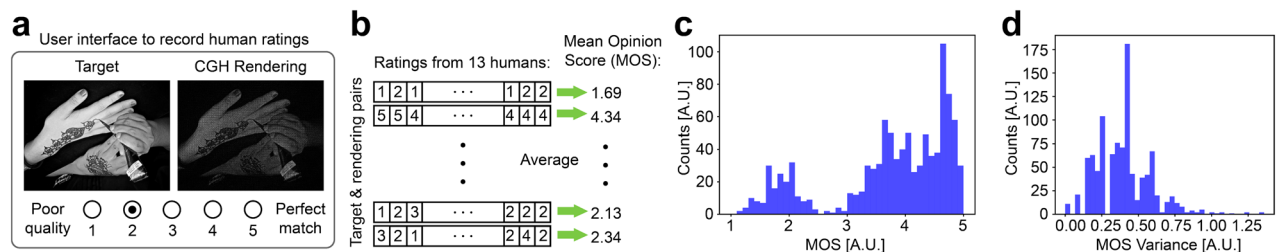
Multiple IQA metrics have already been developed to evaluate the quality of visual data across a broad range of computer vision applications, including image rendering<sup>6,10,13,16,17</sup>, image synthesis<sup>18,19</sup>, transfer<sup>20</sup>, and reconstruction<sup>21,22</sup>. There are three categories of IQA metrics, depending on how imaging data is assessed. The first category is Full Reference IQAs (FR-IQA), which compare a target image to a rendered or measured image. The second category is Reduced Reference IQAs (RR-IQA), which quantify image mismatches indirectly by comparing information derived from both the target and the rendered image. The third category is No Reference IQA (NR-IQA), which evaluates a single image for one or more intrinsic properties such as naturalness<sup>18,23</sup>, sharpness<sup>24</sup>, or contrast<sup>25</sup>.

In the context of CGH-based AR and VR displays, FR-IQAs are utilized to quantitatively rate the fidelity of renderings to the target images. Most of the currently available FR-IQA metrics do not incorporate human feedback data to calculate image mismatch scores. Furthermore, these metrics are primarily intended for evaluating images in general computer vision applications, and their sensitivity to distortions that are specific to any particular image-rendering technology is not well understood. For example, images generated using CGH often display ringing patterns, quantization noise, and speckle artifacts that are direct consequences of the wave-like nature of light, or common limitations of the light modulation hardware. The impact of these specific artifacts on human perception of quality, as well as how various FR-IQA methods measure and evaluate them, remains unclear.

Another challenge FR-IQAs face in the context of CGH is that exact matches between target and rendered images are highly unlikely. CGH-rendered images must be feasible solutions to the wave propagation equation, while target images typically do not meet this requirement. As a result, CGH algorithms only aim to identify the best possible *approximation*. Oppositely, the majority of FR-IQAs are designed to expect exact matches (e.g., mean squared error) or exact correlations (e.g., cosine similarity), and systematically penalize any mismatch, even if they are imperceptible to the human eye. For a precise evaluation of image quality in specialized applications like CGH, FR-IQAs should prioritize penalizing noticeable disparities between target and rendered images while being more tolerant towards mismatches that are less detectable by the human eye.

We developed ConIQA, a new deep learning-based IQA metric, to specifically address these outstanding issues. ConIQA aims to accurately represent human perception and account for domain-specific artifacts with only a limited amount of human-labeled data. We built a Siamese-like network structure and introduced semi-supervised training via consistency training<sup>26</sup> to enhance generalization and prevent overfitting. Our model uses both labeled and unlabeled data for loss computation, with the unsupervised loss calculated by measuring the consistency of network predictions. This is achieved by comparing the network predictions before and after random image transformations are applied to both the target and rendering. Since not all image transformations preserve the perceptual quality, we limited our use to horizontal flipping and introduced a new transformation, weighted averaging. We generated a weighted average between the rendering and the target. We assume that it approximates a CGH solution for which the associated perceptual quality is likely an equally weighted average between a hypothetical perfect match and the original perceptual quality. This weighted averaging method can be used in consistency training to enhance generalization.

We also created a publicly available labeled dataset (Fig. 1), evaluating the human perception of images rendered with CGH. We selected 1000 high-quality natural images from the KonIQ-10k dataset<sup>23</sup> and we generated a CGH rendering for each *target* image by randomly selecting a CGH algorithm from Gerchberg–Saxton algorithm (GS)<sup>27</sup>, NOVO-CGH<sup>28</sup>, and DeepCGH<sup>6</sup>. We then computed Mean Opinion Score (MOS) values for each pair by asking thirteen human volunteers to rate each pair of images for perceived quality on a scale from 1 to 5 (Fig. 1a), and by averaging these ratings for each pair (Fig. 1b,c,d). The resulting hologram quality assessment dataset, HQA1k, is available on a public repository<sup>29</sup>. We also assembled a larger unlabeled dataset based on images selected from the KonIQ-10k dataset. We performed image transformations to augment the dataset, then, we paired each image with a rendering from a randomly selected CGH algorithm. The result is an unlabeled dataset of 33,000 target-rendering pairs.



**Fig. 1.** Synthesis and characterization of the HQA1k dataset. **(a)** Our graphical user interface presented pairs of target-rendering images side by side. Users were asked to rate the quality of images with a score ranging from 1 (poor quality) to 5 (perfect match). **(b)** We recorded ratings across our entire dataset of 1000 image pairs obtained from thirteen participants. For each target-rendering pair of images, we averaged the 13 ratings to yield a mean opinion score (MOS). The resulting dataset, with 1000 rated image pairs, and associated MOS values is named HQA1k. **(c)** Histogram of MOS values in the HQA1k dataset. **(d)** Histogram of per-sample variance of the ratings obtained from thirteen participants evaluating individual image-rendering pairs in the HQA1k dataset.

## Related prior work

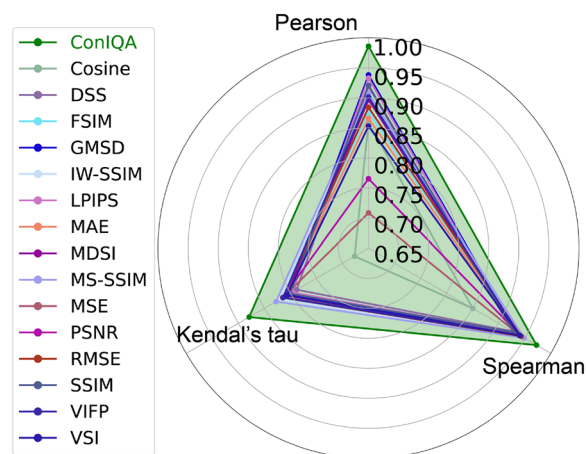
The majority of FR-IQA metrics currently in use are analytical, i.e. they quantify the discrepancies between reference and distorted images through mathematical norms and distances. Key examples of these include Mean Squared Error (MSE)<sup>30</sup>, Mean Absolute Error (MAE)<sup>10,30</sup>, Structural Similarity Index Measure (SSIM)<sup>31,32</sup>, Visual Information Fidelity (VIF)<sup>33</sup>, and Feature Similarity Index Measure (FSIM)<sup>34</sup>. Despite MSE and MAE's ability to measure pixel-level disparities between images, their representation of perceptual quality is not always accurate<sup>13</sup>. Comparing common FR-IQAs to human perception has shown that SSIM is one of the best analytical FR-IQAs in generic computer vision applications<sup>13,35,36</sup>. However, none of the advanced analytical metrics, including SSIM, directly incorporate any human-perceived quality data, and the correlations between these metrics and human ratings remain limited<sup>13,36</sup> (Fig. 2).

Advancements in deep learning<sup>13,20,37</sup> have made it possible to directly incorporate human-perceived quality data into IQA score computations. Deep learning models are especially adept at handling complex non-linear mappings. Consequently, when these models are trained with IQA datasets scored by humans, they can effectively estimate the perceived quality as experienced by humans<sup>13,23,36,38</sup>. However, the *supervised* training of a deep learning FR-IQA model requires a sizable labeled IQA dataset to prevent overfitting<sup>22,39</sup>. The amount of human input needed for creating such datasets can be prohibitively time-consuming and expensive. A variety of approaches have been explored to tackle this problem and train deep-learning FR-IQA models using smaller IQA datasets<sup>20,23,36,38,40–42,42–44</sup>.

The first and most commonly used approach<sup>38,41</sup> is “transfer learning”. This technique involves training a neural network on a proxy task, e.g. image recognition. The trained model is used to compute FR-IQA metrics. In these pre-trained models, convolutional layers serve as generic feature extractors that closely resemble human perception of images<sup>41,45</sup>. Johnson et al.<sup>41</sup> introduced “perceptual loss”, where the Euclidean distance between extracted features from the target and distorted image is computed. Zhang et al.<sup>38</sup> proposed “Learned Perceptual Image Patch Similarity” (LPIPS). LPIPS further expanded the perceptual loss by calculating the Cosine similarity between these features at all convolutional layers. Both studies showed the benefits of using pre-trained networks for image quality assessment tasks by simply comparing the extracted features. However, since the extracted features are based on an unrelated proxy task, human input is not incorporated directly, and alternative methods must be explored to build human-centric FR-IQA models.

Zhang et al.<sup>38</sup> further explored the transfer learning approach by fine-tuning the parameters of the pre-trained neural network. They then used a small fully connected network to predict the final perceptual quality. They also employed data augmentation to prevent overfitting risks associated with the smaller training dataset. This approach has been applied to video quality assessment<sup>44</sup> and NR-IQA tasks<sup>23</sup>, resulting in robust model performance. Despite promising results, feature extractors in these approaches are mainly trained on proxy tasks that are unrelated to IQA, and the final predictions of the network remain imperfect estimations of how humans perceive image quality<sup>13,20,37</sup>.

The second category of techniques trains an entire network from scratch to take full advantage of the deep neural networks' nonlinear mapping abilities. However, since available IQA datasets are usually small, additional strategies must be implemented to prevent overfitting. This includes extensive data augmentation techniques<sup>38,42,46</sup>, reducing the size of the neural networks<sup>38,41,46</sup>, and/or sharing parameters across Siamese-like<sup>38,47</sup> neural network architectures<sup>38,41,42,46</sup>. For instance, Bosse et al.<sup>42</sup> proposed a Siamese-like CNN<sup>48</sup> for both FR-IQA and NR-IQA tasks, using publicly available IQA datasets<sup>31,49,50</sup> to perform fully supervised training. They augmented the training data, by dividing input images into smaller overlapping patches for individual processing, and the final MOS was obtained from the predictions of each patch. This approach has also been



**Fig. 2.** Comparison of ConIQa to existing methods. We used 80% of the HQA1k dataset to train ConIQa. Then we used the remaining 20% to compare ConIQa to other IQA methods on previously unseen data. The radar plot compares the Pearson, Spearman, and Kendall's tau-b correlation coefficients between ground truth MOS values, and FR-IQA predictions. Higher values indicate a stronger correlation with manually scored image pairs.

explored in video quality assessment<sup>43</sup> by segmenting videos into temporal blocks to further increase the size of the available training dataset.

Despite their promising results, these approaches often rely on image transformations that do not always preserve human-perceived quality in FR-IQA applications. Certain operations like cropping, shearing, and rotation can modify crucial global or local information, affecting perceived image quality. For instance, removing segments with high-frequency patterns via cropping will affect human perception to a greater extent than removing uniform areas devoid of high-frequency content. Moreover, high-frequency speckle might go unnoticed if it overlaps with similar patterns in the target image, and may inadvertently inflate the perceptual quality in that part of the image. Unwanted effects such as the inverted face effect<sup>51</sup> or the Thatcher effect<sup>52</sup> can occur when applying transformations like rotating or vertically flipping facial images. Shearing can either suppress or magnify image distortions. These specificities of the human visual system limit the validity of patch-based data augmentation techniques<sup>42</sup>. In high-performance IQAs, global image quality can no longer be deduced from local features within smaller patches. As a result, data augmentation techniques cannot be implemented without first considering their potential impact on FR-IQA scores. Relying solely on these methods during initial neural network training may lead to drift, inaccurate IQA predictions, and ultimately, subpar model performance.

Self-supervised learning<sup>53</sup> is appealing for applications such as FR-IQA, where unlabeled data is easy to obtain, but access to labeled data is limited. Recently, Madhusudana et al.<sup>40</sup> proposed applying self-supervised contrastive learning for NR-IQA. The principle is to train a network to perform an auxiliary task, which shares similarities with the final task, NR-IQA, but is not equal to the task. After initial training, the pre-trained network operates as a feature extractor for the NR-IQA task, similar to the transfer learning approaches discussed earlier. Even though this approach is more closely aligned with the desired NR-IQA task, the outcomes only showed performance that was either similar or slightly superior to existing methods. Our interpretation of these findings is that it is impossible to achieve a high-performance NR-IQA by exclusively employing unlabeled data to train the convolutional layers (or feature extractors). Also, these results show that defining an auxiliary task that closely corresponds to a NR or a FR-IQA task is challenging. Overall, these results indicate that training CNNs using semi-supervised techniques benefits from utilizing both labeled and unlabeled data.

While significant progress has been made, there are currently no methodologies that effectively utilize unlabeled data without compromising the quality of IQA predictions. Even the best existing self-supervised learning approaches are only marginally better than existing methods. Hence, the next logical step is to devise strategies that maintain high IQA prediction accuracy and leverage both a small amount of labeled data, and a large amount of easily available unlabeled data during training.

## Results

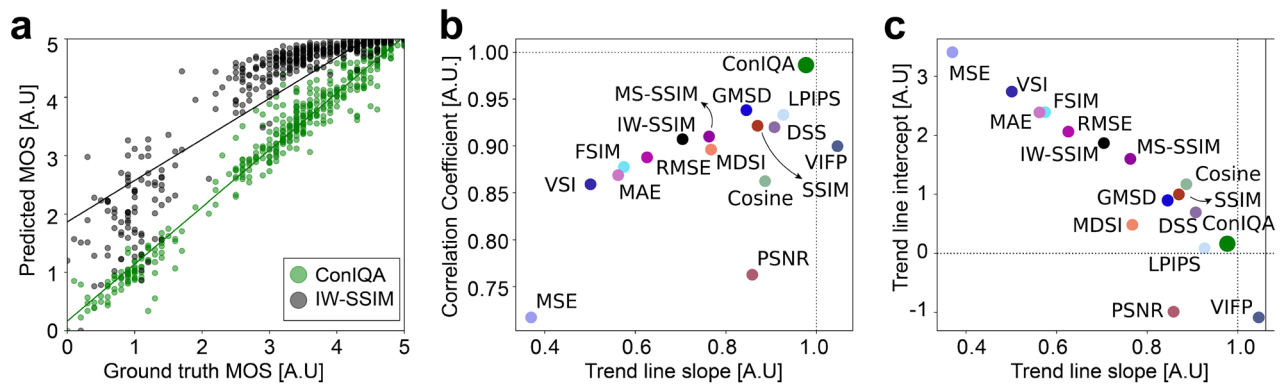
We validated our ConIQA model in a FR-IQA task on images generated with Computer Generated Holography. For this, we first trained our model using 80% of the HQA1k dataset and an additional 33,000 unlabeled pairs of target images matched with their CGH renderings. Our model was trained and tested on a RTX 4090 GPU, with a desktop computer operating a Intel Core i9 4900K CPU, with 64GB of Random Access memory. We trained our model through 4000 epochs which required a total time of 45 hours.

We then compared the performances of ConIQA to fifteen other FR-IQA metrics using the remaining 20% of the HQA1k dataset that was not available during training. Inference time for individual frames, after model optimization was 1 millisecond with the same computer and GPU. Batch inference can further reduce this time by at least one order of magnitude. We considered popular and high-performance FR-IQA metrics: LPIPS<sup>38</sup>, Mean Square Error (MSE)<sup>31</sup>, Mean Absolute Error (MAE)<sup>31</sup>, Root Mean Squared Error (RMSE)<sup>31</sup>, Peak Signal-to-Noise Ratio (PSNR)<sup>31</sup>, Cosine similarity (Cosine)<sup>28</sup>, Structural Similarity Index Measure (SSIM)<sup>32,54</sup>, Multi-Scale SSIM (MS-SSIM)<sup>55</sup>, Information content Weighted SSIM (IW-SSIM)<sup>56</sup>, Visual Information Fidelity (VIFp)<sup>33</sup>, Feature Similarity Index Measure (FSIM)<sup>34</sup>, Gradient Magnitude Similarity Deviation (GMSD)<sup>57</sup>, Visual Saliency-induced Index (VSI)<sup>58</sup>, DCT Subband Similarity Index (DSS)<sup>59</sup>, and Mean Deviation Similarity Index (MDSI)<sup>60</sup>.

To compare the performance of these metrics, we first measured the correlation between the ground truth MOS values and the predictions from our ConIQA model on previously unseen data from our HQA1k dataset. We calculated the same correlations for fifteen other FR-IQAs. Our results (Fig. 2) show that the Pearson, Spearman, and Kendall's tau correlations are consistently higher with ConIQA (0.98, 0.965, and 0.86, respectively) than with any other FR-IQAs. Correlation coefficients measure the degree of linear relationship between the ground truth MOS and the predictions from various IQA metrics. A higher correlation coefficient with human ratings indicates that the FR-IQA being tested precisely predicts human perception. Our results indicate that ConIQA outperforms all existing metrics.

Imperfect correlations between the scores predicted by an FR-IQA and the ground truth MOS values do not always represent uniform, noise-like, prediction errors across the range of possible scores (0-5). Another possibility is that they represent systemic non-linearities in their predictions. For instance, when comparing the predictions of ConIQA and IW-SSIM (Fig. 3a) we find that although the correlation coefficient between IW-SSIM predictions and ground truth MOS is relatively high,  $r = 0.92$ , there's a clear non-linear pattern in the predictions of the model and the overall trend does not follow an ideal identity function. This non-linearity can be indirectly reflected in the slope and intercept of the fitted trend lines. Therefore we fitted similar linear models to all the metrics and we compared the Pearson correlation coefficient to the slope and intercept of the trend lines (Fig. 3b,c). In this context, the correlation coefficient indicates the precision of predictions and the trend line's slope and intercept indicate the prediction accuracy. For instance, we find that VIFp, Cosine, and PSNR are precise, but not accurate, whereas LPIPS is more accurate but not as precise as ConIQA.

In summary, our results show that ConIQA consistently outperforms other metrics with improved correlation, precision, and accuracy. ConIQA predictions compared to actual human predictions follow a linear trend



**Fig. 3.** Comparing the slope and intercept of the fitted trend line for predictions by different metrics as a function of ground truth MOS values. The linear trend is defined as  $MOS' = aMOS + b$  where  $a$ , and  $b$  are slope, and intercept respectively. **(a)** The scatter plot compares the linearity of predictions by ConIQa and IW-SSIM as a function of ground truth MOS values. **(b)** Pearson correlation coefficient vs. slope of the fitted lines for all metrics. The dotted black lines represent the ideal values for the correlation coefficient,  $r = 1$ , and the slope,  $a = 1$ . **(c)** Intercept and slope of the fitted trend lines. The dotted black lines represent the ideal values for the correlation coefficient,  $r = 1$ , and the intercept,  $b = 0$ .

(Fig. 3a). These results show that ConIQa can reliably infer human perceptions of image quality in the context for which it was trained, with only 800 labeled data points. Since the computation of ConIQa does not depend on the image synthesis technique used to generate the training data, our method can be implemented with any other type of image quality assessment task for which labeled data is difficult to obtain.

## Discussion

We have developed and validated ConIQa, a deep learning-based FR-IQA metric that leverages both labeled and unlabeled data for training. ConIQa is specifically designed for image rendering applications where access to labeled image quality assessment data is difficult, expensive, or time-consuming. We employed consistency training to train the model simultaneously with both labeled and unlabeled data. We also introduced an image transformation technique based on weighted averaging for data augmentation in FR-IQA applications.

To validate our approach, and demonstrate ConIQa's robustness, we considered the case of images rendered with computer-generated holography. For this purpose, we developed a labeled dataset, called HQA1k, consisting of 1000 target and rendering pairs with human-rated quality. ConIQa was trained on 80% of this labeled dataset, and with an additional large unlabeled dataset. Our results indicate that ConIQa consistently outperforms other FR-IQA methods, by reliably predicting MOS values that correlate well with human perception and are accurate, and precise.

Although we validated ConIQa in the context of images rendered with CGH, ConIQa is designed to assess image quality across a broad spectrum of image synthesis, rendering, or manipulation methods and the computation of ConIQa does not set constraints on how the training data is obtained. Our approach is particularly beneficial in contexts where the unique, or experimental nature of the rendering technique makes obtaining large labeled datasets challenging. Thus, the principles of ConIQa extend far beyond the specific application presented in this work, to any other image synthesis technique.

The first and most obvious limitation of our technique is the sample size and the number of human participants who volunteered to manually review our dataset of 1000 image pairs. Recently, crowd-sourcing has enabled the creation of large human-labeled datasets where thousands of humans can collectively rate larger IQA datasets. However, we envision that labeling samples with a smaller number of participants can be sufficient. One way to test this hypothesis would be to train and test ConIQa on a dataset that is labeled by a smaller group of human participants and compare performance with our current results. Beyond CGH applications, further testing of ConIQa in other domains would provide a more comprehensive understanding of the model's performance, which could be the subject of future studies.

HQA1k and ConIQa evaluations are performed on simulated CGH renderings, which are different from experimentally recorded CGH renderings. While further training and testing on experimentally recorded holograms may help improve our proposed technique, it should be noted that the absence of standardized CGH-based displays introduces challenges to this approach. The imperfections and limitations of the experimental setup that physically renders the hologram to the users will introduce a considerable amount of *display-specific* distortions to the rendering. This considerably affects the perceptual quality of renderings yet is fundamentally separate from the CGH algorithm and is unique to the display setup being used. However, ConIQa can be trained specifically on experimentally recorded renderings and can be used to evaluate both the CGH algorithm and the specific display together. For instance, ConIQa can be implemented to evaluate the performance of different CGH algorithms with a particular type of near-eye display hardware.

Future research could also explore extending ConIQa to other applications such as image and video quality assessment, where accurate and reliable FR-IQA metrics are also in demand. For instance, ConIQa could help develop more efficient image and video compression algorithms that prioritize the storage of information that

human observers consider important. We expect that the versatility of ConIQA will similarly benefit many other research fields, beyond display technologies and computer vision.

## Methods

### HQA1k: a labeled dataset for perceptual hologram quality assessment

To validate ConIQA, we considered the use case of Computer-Generated Holography. We developed a labeled dataset, HQA1k, to train, test, and compare ConIQA to other FR-IQA methods. We created HQA1k by first gathering 1000 sample images from the KonIQ-10K dataset<sup>23</sup>. We specifically selected images that human observers consider to be of high quality according to the KonIQ-10k dataset's recorded MOS values. This was done to ensure that human evaluations are minimally biased by variations in the inherent quality of the target images. We then randomly selected a CGH algorithm for each image and we calculated a potential CGH rendering. The CGH algorithms we used include Gerchberg–Saxton<sup>27</sup>, NOVO-CGH<sup>28</sup>, and DeepCGH<sup>6,16</sup>. For iterative algorithms, we randomly selected a number of iterations between 10 and 100 to ensure that HQA1k would contain a broad distribution of examples with various levels of mismatches. All samples rated by users exhibited a certain level of correspondence between the target and the rendering, excluding cases where the iterative algorithms diverged.

We displayed each pair of target-rendering in HQA1k on a 1080p display using a custom-designed graphical user interface (see Fig. 1a). Thirteen participants manually reviewed and rated their perception of the degree of the mismatch between the target images and their matched renderings with a grade ranging from 1 to 5. Before scoring image pairs, participants were advised that a grade of '1' signified a poor match while a '5' indicated a perfect match. The image pairs were presented randomly to eliminate bias from previous observations, and to mitigate the effects of experience, and drifts of grading between images that were scored early and late in the sequence. The mean opinion score (MOS) for each sample was calculated by averaging the ratings of all participants (see Fig. 1b). The result is a dataset of 1000 target-rendering pairs matched to a human label representing a suitable FR-IQA score for human perception.

Figure 1c shows the distribution of MOS values across the dataset. We observe that MOS skew toward high values corresponding to high-quality renderings. This uneven distribution is because even a low iteration count can still produce visually accurate solutions. Also, the convergence speed of iterative CGH methods is generally unpredictable and depends both on the type of algorithm and on the target image.

Figure 1d shows the variance of the 13 recorded scores associated with each MOS element in the dataset. Individual perceptions of mismatches between the target and the rendered image do not substantially vary from person to person, indicating that human observers were overall in agreement with each other in their individual perceptions of image quality.

In most Full-Reference Image Quality Assessment (FR-IQA) applications, renderings may be entirely devoid of content. For instance, when initialization effects remain entirely visible after only one iteration of a CGH algorithm. Since our HQA1k dataset does not include these low-quality samples, we incorporated an additional 180 lowest-quality image pairs into the ConIQA's training set. We paired target images from the KonIQ-10k dataset with renderings made of uniform zeros, uniform ones, or random patterns, that correspond to commonly used initialization values for CGH algorithms, and we assigned to these image pairs a bottom-value MOS score of 0.

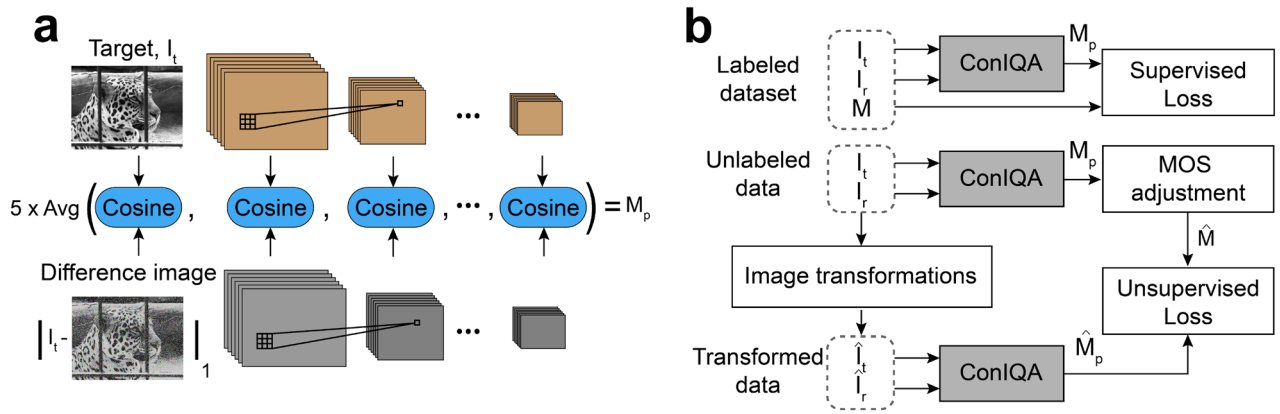
### ConIQA and consistency training

The notations used in the following sections are listed below along with their descriptions:

$I_t$	Target image.
$I_r$	Rendered image.
$\hat{I}_t$	Transformed target image, for consistency training.
$\hat{I}_r$	Transformed rendered image, for consistency training.
$M$	Ground truth MOS value associated with the original rendering.
$M_p$	Predicted MOS, the output of ConIQA.
$\hat{M}$	Expected MOS value after the transformation.
$\hat{M}_p$	Predicted MOS, the output of ConIQA for transformed images.
$L$	Number of convolutional layers in the Network.
$H_l$	Height of the feature maps at Level $l$ .
$W_l$	Width of the feature maps at Level $l$ .
$\mathbf{f}_l^t(i, j)$	Feature vectors at pixel $(i, j)$ in the target image at level $l$ .
$\mathbf{f}_l^d(i, j)$	Feature vectors at pixel $(i, j)$ in the difference image at level $l$ .
$\lambda$	Weight used in transforming the image and MOS value.

### ConIQA network structure

Figure 4a shows ConIQA's network structure. ConIQA is a Siamese-style network where two parallel branches with shared parameters process both the target and the *difference image*. The *difference image* is calculated as the absolute value of the difference between the rendered image and the target. Each branch consists of five convolutional blocks, where each block consists of two convolutional layers, each followed by batch normalization. Each convolutional block is followed by a max pooling layer that halves the spatial dimensions of the feature maps. The number of feature maps in the convolutional blocks is doubled after each max-pooling halves the spatial dimensions, a common practice in deep learning model design. Although the parameters of the convolutional layers are shared between the two Siamese branches, the batch normalization layers have distinct parameters to allow for the learning of different statistics in each branch. The final MOS is obtained by calculating the cosine similarity between the extracted feature vectors at each pixel in the target and the difference images, and by



**Fig. 4.** ConIQa network structure and the consistency training paradigm. (a) ConIQa is a Siamese-style network where the same network is applied to both the target (top) and the difference (bottom) images. The difference image is defined as  $|I_t - I_r|$ . The final estimated perceptual quality,  $M_p$ , is calculated by averaging the cosine similarity of feature maps in the two branches. (b) To train ConIQa, we compute a supervised loss, calculated over the HQA1k dataset, as well as an unsupervised loss which is calculated over a large unlabeled dataset of target-rendering pairs. The unsupervised loss is calculated by comparing the network predictions before and after image transformations are applied to each unlabeled sample.

averaging their values for all positions across the spatial dimension. While LPIPS includes normalization steps in this process, we chose to omit this for simplicity. The network-estimated MOS,  $M_p$ , is obtained by averaging the average cosine similarities at each level, scaled to match the (0-5) range of HQA1k. The output of our ConIQa model, the predicted MOS value,  $M_p$ , can be expressed as:

$$M_p = \sum_{l=1}^L \frac{5}{LH_lW_l} \sum_{i=1}^{H_l} \sum_{j=1}^{W_l} \frac{\mathbf{f}_l^t(i,j) \cdot \mathbf{f}_l^d(i,j)}{\|\mathbf{f}_l^t(i,j)\|_2 \|\mathbf{f}_l^d(i,j)\|_2} \quad (1)$$

### Consistency training

Consistency training is a highly advantageous semi-supervised training strategy, especially in scenarios where unlabeled data are abundant. Unlike traditional approaches that heavily rely on labeled data, consistency training leverages both labeled and unlabeled data to calculate both a supervised and an unsupervised loss (Fig. 4b). In our implementation, the supervised loss compares predictions of our ConIQa network to the ground truth labels in HQA1k, using a weighted average ( $\alpha = 0.5$ ), of MSE and MAE given by:

$$\mathcal{L}_{sup} = \alpha \text{MSE}(M, M_p) + (1 - \alpha) \text{MAE}(M, M_p) \quad (2)$$

To compute the unsupervised loss, we first pass the unlabeled data ( $I_t, I_r$ ) through our ConIQa network and record the corresponding Mean Opinion Score (MOS) predictions,  $M_p$  (Fig. 5a). Subsequently, we apply random transformations to the unlabeled data to generate transformed image pairs ( $\hat{I}_t, \hat{I}_r$ ) and we record the MOS predictions,  $\hat{M}_p$ , of these transformed images from the network. We use the network's MOS values,  $M_p$ , for the original images to independently predict the expected MOS values,  $\hat{M}$ , for the transformed images. We performed consistency training using only two image transformations, horizontal flip, and a novel technique called image augmentation by weighted averaging. When only horizontal flipping is applied (Fig. 5b), the expected MOS remains equal to the original MOS,  $\hat{M} = M_p$ .

The principle for weighted averaging is shown in Fig. 5c. The transformed rendered image,  $\hat{I}_r$ , is a weighted average of the target and rendered images given by:

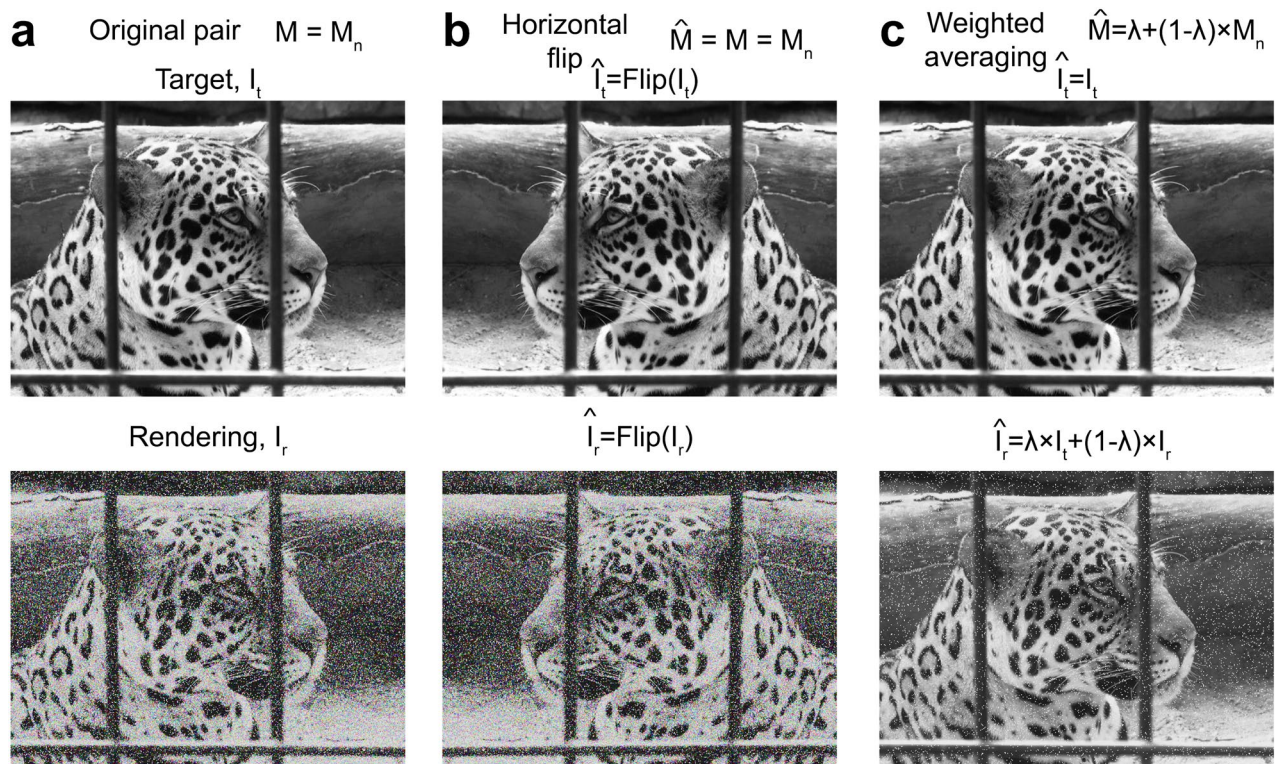
$$\hat{I}_r = \lambda \times I_t + (1 - \lambda) \times I_r \quad (3)$$

For labeled data, the MOS value of the transformed rendering,  $\hat{M}$ , is approximated as the weighted average of the ground truth MOS,  $M$ , and 1:

$$\hat{M} = \lambda \times 1 + (1 - \lambda) \times M \quad (4)$$

For unlabeled data, there is no ground truth MOS value associated with the original rendering. Therefore this will be replaced by the MOS value that the ConIQa predicts for the original pair,  $M_p$ . In our experiments, we choose  $\lambda = 0.05$ . It is essential to note that during this process, the network isn't subjected to gradient calculations, instead it serves as an inference tool to estimate the expected MOS for the unlabeled data.

We then equate the anticipated MOS values,  $\hat{M}$ , with those generated by the ConIQa network for the transformed images,  $\hat{M}_p$ . Using a weighted average of MAE and MSE as our loss function, we calculate the unsupervised loss,  $\mathcal{L}_{unsup}$  given by:



**Fig. 5.** Image transformations used for consistency training in ConIQa. Two types of transformations were used to augment (a) samples from our HQA1k dataset. We considered (b) horizontal flipping, and (c) a novel augmentation method based on weighted averaging. The latter transformation generates a new sample from an existing pair, wherein the target is fixed and the rendering is replaced by a weighted average of the target and the rendering. The weight,  $\lambda$ , is a small random number drawn from a uniform distribution in the  $[0, 0.05]$  range.

$$\mathcal{L}_{\text{unsup}} = \alpha \text{MSE}(\hat{M}, \hat{M}_p) + (1 - \alpha) \text{MAE}(\hat{M}, \hat{M}_p) \quad (5)$$

The overall loss is the sum of the supervised and unsupervised losses. With this methodology, our model is capable of learning more robust data representations, which minimizes the risk of overfitting to the labeled data, and improves its generalization potential.

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \mathcal{L}_{\text{unsup}} \quad (6)$$

### Data availability

The HQA1k Hologram Perceptual Quality Assessment Dataset is available at: <https://doi.org/10.17615/mz1n-te79>.

### Code availability

A curated version of the code is available on our repository. <https://github.com/UNC-optics/>.

Received: 2 February 2024; Accepted: 16 August 2024

Published online: 29 August 2024

### References

- Pillai, A. S. & Mathew, P. S. Impact of virtual reality in healthcare: a review. *Virtual and augmented reality in mental health treatment* 17–31 (2019).
- Bates, J. Virtual reality, art, and entertainment. *Presence: Teleoperators & Virtual Environments* **1**, 133–138 (1992).
- Radianti, J., Majchrzak, T. A., Fromm, J. & Wohlgenannt, I. A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. *Computers & Education* **147**, 103778 (2020).
- Malik, A. A., Masood, T. & Bilberg, A. Virtual reality in manufacturing: immersive and collaborative artificial-reality in design of human-robot workspace. *International Journal of Computer Integrated Manufacturing* **33**, 22–37 (2020).
- Eybposh, M. H., Curtis, V. R., Moossavi, A. & Pégard, N. C. Dynamic computer generated holography for virtual reality displays. In *Frontiers in Optics, FM3B-3* (Optical Society of America, 2021).
- Eybposh, M. H., Caira, N. W., Atisa, M., Chakravarthula, P. & Pégard, N. C. Deepcgh: 3d computer-generated holography using deep learning. *Optics Express* **28**, 26636–26650 (2020).
- Eybposh, M. H., Moossavi, A., Curtis, V. R. & Pégard, N. C. Optimization of time-multiplexed computer-generated holograms with surrogate gradients. In *Emerging Digital Micromirror Device Based Systems and Applications XIV, PC1201406* (SPIE, 2022).
- Choi, S., Gopakumar, M., Peng, Y., Kim, J. & Wetzstein, G. Neural 3d holography: Learning accurate wave propagation models for 3d holographic virtual and augmented reality displays. *ACM Transactions on Graphics (TOG)* **40**, 1–12 (2021).



9. Shi, L., Li, B., Kim, C., Kellnhofer, P. & Matusik, W. Towards real-time photorealistic 3d holography with deep neural networks. *Nature* **591**, 234–239 (2021).
10. Peng, Y., Choi, S., Padmanaban, N. & Wetzstein, G. Neural holography with camera-in-the-loop training. *ACM Transactions on Graphics (TOG)* **39**, 1–14 (2020).
11. Georgiou, A., Kollin, J., Hewitt, C., Chakravarthula, P. & Guenter, B. Visual perception of noise in a simulated holographic display-a user study. *Displays* **76**, 102333 (2023).
12. Chakravarthula, P., Tseng, E., Srivastava, T., Fuchs, H. & Heide, F. Learned hardware-in-the-loop phase retrieval for holographic near-eye displays. *ACM Transactions on Graphics (TOG)* **39**, 1–18 (2020).
13. Eybposh, M. H., Moossavi, A., Xu, J., Caira, N. W. & Pégard, N. C. Perceptual quality assessment in holographic displays with a semi-supervised neural network. In *Digital Holography and Three-Dimensional Imaging, Th1A-6* (Optica Publishing Group, 2022).
14. Ahar, A. et al. Comprehensive performance analysis of objective quality metrics for digital holography. *Signal Processing: Image Communication* **97**, 116361. <https://doi.org/10.1016/j.image.2021.116361> (2021).
15. Yang, F. et al. Perceptually motivated loss functions for computer generated holographic displays. *Scientific Reports* **12**, 7709–7709 (2022).
16. Eybposh, M. H., Caira, N. W., Chakravarthula, P., Atisa, M. & Pégard, N. C. High-speed computer-generated holography using convolutional neural networks. In *Optics and the Brain, BTu2C-2* (Optical Society of America, 2020).
17. Eybposh, M. H., Curtis, V. R., Rodriguez-Romaguera, J. & Pégard, N. C. Advances in computer-generated holography for targeted neuronal modulation. *Neurophotonics* **9**, 041409–041409 (2022).
18. Creswell, A. et al. Generative adversarial networks: An overview. *IEEE signal processing magazine* **35**, 53–65 (2018).
19. Croitoru, F.-A., Hondru, V., Ionescu, R. T. & Shah, M. Diffusion models in vision: A survey. arXiv preprint [arXiv:2209.04747](https://arxiv.org/abs/2209.04747) (2022).
20. Zhai, G. & Min, X. Perceptual image quality assessment: a survey. *Science China Information Sciences* **63**, 1–52 (2020).
21. Tatsugami, F. et al. Deep learning-based image restoration algorithm for coronary ct angiography. *European radiology* **29**, 5322–5329 (2019).
22. Eybposh, M. H., Ebrahim-Abadi, M. H., Jalilpour-Monesi, M. & Saboksayr, S. S. Segmentation and classification of cine-mr images using fully convolutional networks and handcrafted features. arXiv preprint [arXiv:1709.02565](https://arxiv.org/abs/1709.02565) (2017).
23. Hosu, V., Lin, H., Sziranyi, T. & Saupe, D. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing* **29**, 4041–4056 (2020).
24. Ferzli, R. & Karam, L. J. A no-reference objective image sharpness metric based on the notion of just noticeable blur (jnb). *IEEE transactions on image processing* **18**, 717–728 (2009).
25. Jaya, V. & Gopikakumari, R. Iem: a new image enhancement metric for contrast and sharpness measurements. *International Journal of Computer Applications* **79** (2013).
26. Xie, Q., Dai, Z., Hovy, E., Luong, T. & Le, Q. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems* **33**, 6256–6268 (2020).
27. Gerchberg, R. W. A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik* **35**, 237–246 (1972).
28. Zhang, J., Pégard, N., Zhong, J., Adesnik, H. & Waller, L. 3d computer-generated holography by non-convex optimization. *Optica* **4**, 1306–1313 (2017).
29. Eybposh, M., Cai, C., Moossavi, A., Rodriguez Romaguera, J. & Pégard, N. Hq1k hologram perceptual quality assessment dataset. *Carolina Digital Repository/UNC Libraries* (2023).
30. Wang, Z. & Bovik, A. C. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE signal processing magazine* **26**, 98–117 (2009).
31. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **13**, 600–612 (2004).
32. Eybposh, M. H., Turani, Z., Mehregan, D. & Nasirivanaki, M. Cluster-based filtering framework for speckle reduction in oct images. *Biomedical optics express* **9**, 6359–6373 (2018).
33. Sheikh, H. R. & Bovik, A. C. A visual information fidelity approach to video quality assessment. In *The first international workshop on video processing and quality metrics for consumer electronics*, 2117–2128 (sn, 2005).
34. Zhang, L., Zhang, L., Mou, X. & Zhang, D. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on image processing* **20**, 2378–2386 (2011).
35. Sara, U., Akter, M. & Uddin, M. S. Image quality assessment through fsim, ssim, mse and psnr-a comparative study. *Journal of Computer and Communications* **7**, 8–18 (2019).
36. Jinjin, G. et al. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI* **16**, 633–651 (Springer, 2020).
37. Chow, L. S. & Paramesran, R. Review of medical image quality assessment. *Biomedical signal processing and control* **27**, 145–154 (2016).
38. Zhang, R., Isola, P., Efros, A. A., Shechtman, E. & Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595 (2018).
39. LeCun, Y., Bengio, Y. & Hinton, G. *Deep learning*. *nature* **521**, 436–444 (2015).
40. Madhusudana, P. C., Birkbeck, N., Wang, Y., Adsumilli, B. & Bovik, A. C. Image quality assessment using contrastive learning. *IEEE Transactions on Image Processing* **31**, 4149–4161 (2022).
41. Johnson, J., Alahi, A. & Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, 694–711 (Springer, 2016).
42. Bosse, S., Maniry, D., Müller, K.-R., Wiegand, T. & Samek, W. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on image processing* **27**, 206–219 (2017).
43. Zhang, Y., Gao, X., He, L., Lu, W. & He, R. Objective video quality assessment combining transfer learning with cnn. *IEEE transactions on neural networks and learning systems* **31**, 2716–2730 (2019).
44. Sun, W., Wang, T., Min, X., Yi, F. & Zhai, G. Deep learning based full-reference and no-reference quality assessment models for compressed ugc videos. In *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 1–6 (IEEE, 2021).
45. Nguyen, A., Yosinski, J. & Clune, J. Understanding neural networks via feature visualization: A survey. *Explainable AI: interpreting, explaining and visualizing deep learning* 55–76 (2019).
46. Kim, J. & Lee, S. Deep learning of human visual sensitivity in image quality assessment framework. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).
47. Koch, G., Zemel, R., Salakhutdinov, R. et al. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop* (Lille, 2015).
48. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2014).
49. Sheikh, H. Live image quality assessment database release 2. <http://live.ece.utexas.edu/research/quality> (2005).
50. Sheikh, H. R., Sabir, M. F. & Bovik, A. C. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing* **15**, 3440–3451 (2006).

51. Civile, C., McLaren, R. P. & McLaren, I. P. The face inversion effect—parts and wholes: Individual features and their configuration. *The Quarterly Journal of Experimental Psychology* **67**, 728–746 (2014).
52. Dahl, C. D., Logothetis, N. K., Bülthoff, H. H. & Wallraven, C. The thatcher illusion in humans and monkeys. *Proceedings of the Royal Society B: Biological Sciences* **277**, 2973–2981 (2010).
53. Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D. & Makedon, F. A survey on contrastive self-supervised learning. *Technologies* **9**, 2 (2020).
54. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **13**, 600–612 (2004).
55. Wang, Z., Simoncelli, E. & Bovik, A. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2, 1398–1402 Vol.2, doi: 10.1109/ACSSC.2003.1292216 (2003).
56. Wang, Z. & Li, Q. Information content weighting for perceptual image quality assessment. *IEEE Transactions on Image Processing* **20**, 1185–1198. <https://doi.org/10.1109/TIP.2010.2092435> (2011).
57. Xue, W., Zhang, L., Mou, X. & Bovik, A. C. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *CoRR abs/1308.3052* (2013). 1308.3052.
58. Zhang, L., Shen, Y. & Li, H. Vsi: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image Processing* **23**, 4270–4281. <https://doi.org/10.1109/TIP.2014.2346028> (2014).
59. Balanov, A., Schwartz, A., Moshe, Y. & Peleg, N. Image quality assessment based on dct subband similarity. In *2015 IEEE International Conference on Image Processing (ICIP)*, 2105–2109, doi: 10.1109/ICIP.2015.7351172 (2015).
60. Nafchi, H. Z., Shahkolaei, A., Hedjam, R. & Cheriet, M. Mean deviation similarity index: Efficient and reliable full-reference image quality evaluator. *CoRR abs/1608.07433* (2016). 1608.07433.

## Acknowledgements

The authors would like to thank Nickolas Caira, Maria Magdalena Ortiz, Jovan Tormes Vaquerano, Harshith Manepali, and Xiyuan Du for their contributions and discussions regarding the concept of ConIQA and the creation of HQA1k.

## Author contributions

M.H.E. developed ConIQA with major contributions to the code and HQA1k dataset development. He conceptualized the study, simulation and experiments, and was the main writer of the manuscript. C.C. performed data analysis, and result interpretation, and contributed to writing the manuscript. A.M. developed the methodology and wrote part of the code, contributed to writing the manuscript. J.R.R. provided guidance mentorship, and funding, and helped prepare the manuscript. N.C.P., Scientific advisor, provided guidance, mentorship, and funding, and helped prepare the manuscript.

## Funding

This study was funded by Foundation of Hope for Research and Treatment of Mental Illness, Brain and Behavior Research Foundation, Whitehall Foundation, National Institutes of Health (No. R01MH132073), North Carolina Translational and Clinical Sciences Institute, University of North Carolina at Chapel Hill, Kavli Foundation, Burroughs Wellcome Fund, Arnold and Mabel Beckman Foundation and Alfred P. Sloan Foundation.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to J.R.-R. or N.C.P.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024