Scientia Iranica D (2023) 30(2), 492–505

SCIENTIA IRANICA

Sharif University of Technology
Scientia Iranica
Transactions D: Computer Science & Engineering and Electrical Engineering
http://scientiairanica.sharif.edu

Multiple hallucinated deep network for image quality assessment

Z. Javidian, S. Hashemi*, and S.M. Hazrati Fard

Department of Computer Science and Engineering, Shiraz University, Molla Sadra Ave., Shiraz, Fars, Iran.

Received 13 October 2021; received in revised form 2 June 2022; accepted 21 November 2022

KEYWORDS
Image quality assessment;
Deep learning;
Generative adversarial network;
Distribution alignment.

Abstract. Image Quality Assessment (IQA) refers to quantitative evaluation of the human's perception of a distorted image quality. Blind IQA (BIQA) is a type of IQA that does not include any reference or information about the distortion. Since the human brain has no information about the distortion type, BIQA is more reliable and compatible with the real world. Traditional methods in this realm used an expert opinion, such as Natural Scene Statistics (NSS), to measure the distance of a distorted image from the distribution of pristine samples. In recent years, many deep learning-based IQA methods have been proposed to use the ability of deep models in automatic feature extraction. However, the main challenge of these models is the need for many annotated training samples. In this paper, through the inspiration of Human Visual System (HVS), a Generative Adversarial Network (GAN)-based approach was proposed to address this problem. To this end, multiple images were sampled from a submanifold of the pristine data manifold by conditioning the network on the corresponding distorted image. In addition, NSS features were employed to improve the network training and conduct the training process on the right track. The testing results of the proposed method on three datasets confirmed its superiority over other the state-of-the-art methods.

1. Introduction

Nowadays, the explosive growth of social networks has led to the production of a massive number of images. It should be noted that digital images are at the risk of distortion at any stage of their life cycle. For example, during photography, variations in focal length and camera angle can distort the real image. In addition, there is high possibility of compression, storage, transmission, and loss of received visual information. To obtain high-quality images, it is necessary that a reliable quality assessment metric be established. Evaluation of the image quality by experts can be reliable and accurate; yet, it is practically quite costly and time-consuming. Image Quality Assessment (IQA), as part of the quality of experience measures [1], is the automatic process of determining the level of accuracy and perceptual quality of an image. IQA is an applicable alternative to imitating the IQA by humans that have found extensive applications in several fields such as image retrieval [2] and restoration [3].

Depending on the amount of reference information required during quality evaluation, IQA methods are divided into three problem statements:

- Full-Reference IQA (FR-IQA) [4]: Here, the perfect quality reference image is fully available to predict

*. *Corresponding author.*
E-mail address: s_hashemi@shirzau.ac.ir (S. Hashemi)

doi: 10.24200/sci.2022.59243.6134

the quality score. Then, the distorted image can be compared with an undistorted version or the original image;

- Reduced-Reference IQA (RR-IQA) [5]: It aims to measure quality with part of the reference data accurately. A challenge faced during quality measurement is to deal with the effective representation of the visual content of images with limited data;
- No-Reference IQA (NR-IQA) [6]: It is defined as the quality measurement of distorted images with no reference to the original images.

Given that FR-IQA and RR-IQA use reference images for comparison in the quality prediction process, they are less challenging. Of note, despite reaching noticeable improvement in designing a metric for IQA, they are not practically applicable given their need for non-distorted reference images. In contrast, there is no need for non-distorted reference images in NR-IQA and for this reason, it has received significant attention in real-world applications. As NR-IQA takes distorted images to be assessed as input without any additional information, we face an ill-posed problem [7].

Some early NR-IQA approaches were concentrated on specific distortion types, e.g., blocking artifacts [8], blur, and ringing effects [9,10]. However, these methods can only deal with the problems containing one known type of distortion in the images. In contrast, in real-world applications, distortion types are unknown [6,11–14] that leads to the performance bottleneck for such problems.

In the absence of distorted form and its corresponding non-distorted reference image, the ill-posed nature of the underdetermined NR-IQA is highlighted. Designing powerful feature representation models can be a remedy to alleviate this dilemma. To tackle the real-world problems, general-purpose NR-IQA approaches attempt to characterize the general rules of image distortions to establish an image quality prediction model based on hand-crafted [11] or learned features [6].

In the last decade, Convolutional Neural Networks (CNNs) [15] have expanded to many fields of computer vision [16,17]. To learn the complex relationship between the image data and human perceived quality, CNNs need many trainable parameters to be effective. In addition, NR-IQA approaches based on CNNs [7,18,19] have achieved better performance than the traditional hand-crafted feature-based NR-IQA methods [12,13,20]. The main challenge is that training CNN-based IQA models requires a massive number of labeled samples. As annotating image quality by humans is extremely expensive and time-consuming, the lack of such public datasets significantly affects these methods.

Given that the scale of the existing annotated IQA databases [21,22] is usually limited, training deep IQA models using only these databases will lead to overfitting. Use of data augmentation strategies can be a solution to this issue. Another solution is using transfer learning and general image feature representations from a pre-trained model to quality prediction [23]. Existing works usually rely on the pre-trained network models in which an extensive training dataset is available, e.g., using ImageNet in image classification tasks [24,25]. The issue that leads to a reduction in the effectiveness of transfer learning is the different nature of NR-IQA and image classification tasks. These models cannot quickly adapt to new distortions, and the generalization performance is unsatisfactory in such cases.

Hallucinated-IQA model [7] is a deep network comprising a Quality network that works as a regression model. In this network, a score can be assigned to each input distorted image. To approximate the quality function of deep networks, a sufficient number of annotated samples are required. In the absence of such annotated instances, searching the hypothesis space does not support enough information. Since the final function contains high variance in the data sample, this is not robust. In case of no sufficient examples, Human Visual System (HVS) considers perceptual information between the distorted image and imagined undistorted image to have a more precise prediction. This undistorted image in the HVS can be named the hallucinated reference. To address the lack of data to train a deep model, it is common to simulate the behavior of HVS with a hallucinated model. To this end, this network tries to take a sample from solution space and make a hallucinated image [7].

In the hallucinated-IQA model [7], multiple objective functions are used to efficiently handle different aspects of the problem. These objective functions should be combined with hyperparameters involving a trade-off. Therefore, finding the optimized values of these hyper-parameters is a complex procedure. To avoid this complexity, the application of an objective function based on perceptual features was proposed. The hallucinated model uses a reinforced structure to overcome the complexity of the hyper-parameter optimization [7]. However, this may cause difficulty to the algorithm convergence. As the co-training of quality and hallucinated model (reinforced structure) can make it hard to train, using a unique objective function only based on perceptual features was recommended. Furthermore, using a unified objective function makes it feasible to separate the training procedure of the quality and hallucinated model. To address the problem of insufficient annotated examples, another suggestion is to reduce the model capacity. To this end, the hypothesis space was divided by sampling from a submanifold of data instead of the whole data manifold [26].

Previous methods aiming to divide the data manifold only use mean statistics such as Mean Square Error (MSE) to minimize the reconstruction difference [27]. To obtain better features, utilization of an inference network was suggested to align the generated perceptual features. The inference network can bring back data from the perceptual space to the image ambient space. The small reconstruction difference between the perceptual features and original distribution witnesses the quality of the proposed network.

In the literature, such as the proposed methods by Lin and Wang [7] and Gu et al. [28], only one sample was generated for each distorted image that cannot bring about a significant level of estimation. To address this problem, we generate multiple hallucinated images conditioned on each distorted image. On the contrary, multiple hallucinated images reflect different aspects of the original image. Finally, the Quality network can use the best parts of images to reach a reliable score. Further in this study, use of Natural Scene Statistics (NSS) compiled by an expert was proposed as a prior to restrict the hypothesis space. In the initial training steps, this prior knowledge can guide the steps of the algorithm in the right region of the hypothesis space.

To recap the presented model, a deep Quality network was developed that received a distorted image and returned its score. The score reflects the similarity of the distorted image to its corresponding pristine image. The main contribution of this paper is its conversion of a problem that has been always solved in a multi-objective manner into a single objective by learning perceptual features through distribution alignment. To this end, instead of using a sample in each training iteration, using multiple examples to reach a better estimation was suggested.

The rest of this paper is organized as follows: Section 2 reviews the related works. Section 3, firstly, examines some required research backgrounds and then, elaborates on the proposed algorithm in detail. Section 4 mentions the datasets and compares the experimental results with those from the rival methods. Finally, Section 5 concludes the study with a summary of the proposed work and discussions.

2. Related work

NR-IQA can be classified into distortion-specific [8,10] and blind methods [6,13,14,20]. While in the former, the image quality is evaluated by extracting the features of a determined distortion, no distortion is assumed in the latter during the training of the Quality network. To address this problem, some feature-based models have been developed. These features can be based on the experts' experience if available; otherwise, they should be learned. To this end, blind NR-IQA methods can be divided into the NSS- [12,29] and learning-based approaches [6,30,31].

The NSS-based methods assume that natural images have certain statistical characteristics, which will be changed under different distortions. Moorthy and Bouik [32] proposed to extract the NSS features from Discrete Wavelet Transform (DWT) domain for blind IQA. Saad et al. [29] leveraged the statistical features of Discrete Cosine Transform (DCT) to estimate the image quality. Mittal et al. [12] proposed a general-purpose NR-IQA metric by extracting the NSS features in the spatial domain and achieved promising performance. In addition to the NSS-based approaches, learning-based approaches have also been developed. For example, the codebook representation approaches [6,30] were proposed to predict subjective image quality scores by Support Machine Regression (SVR) model. Zhang et al. [31] combined the semantic-level features affecting the HVS with local features for image quality estimation.

In recent years, the deep learning-based general-purpose NR-IQA methods have exhibited superior prediction performance over traditional methods [7,19,24,33]. One key issue in deep learning is that it requires abundant labeled data, while IQA is a typical small sample problem. Bianco et al. trained a deep model on a large-scale database for image classification task and then, fine-tuned it for NR-IQA task [24]. Talebi and Milanfar [25] used a model based on Deep Convolutional Neural Network (DCNN). They proposed their model by predicting the perceptual distribution of subjective quality opinion scores. The model parameters were also initialized by pre-training on ImageNet database [25]. Lin and Wang suggested the application of a Generative Adversarial Network (GAN) to generate some hallucinated images to overcome the problem of annotated data availability [7]. However, the GAN-based IQA approaches are subject to some shortcomings. The most dominant challenge is to use several loss functions; consequently, many hyperparameters should be trained. Each loss function needs a hyper-parameter that is a part of the perception-distortion trade-off. Given that the parameters should be estimated through cross-validation, finding an equilibrium point in these trade-offs can be quite challenging [27]. As another problem, in the GAN-based Blind IQA (BIQA) systems, despite the Generator, the Discriminator is not conditioned on the distorted images. Then, instead of concentrating on a specific distortion, the Generator must estimate the whole pristine manifold. Intrinsically, GAN has instability issues, i.e., mode collapse and lack of equilibrium and convergence. Using reinforced structure in the GAN-based IQA approaches [7,28] can make them more susceptible to divergence from the stable path.

Table 1 presents the specifications of the selected

Table 1. Summary of the major state-of-the-art in IQA and their specifications.

Model	Features	Quality estimation	Time complexity
BRISQUE	Mean Subtracted Contrast Normalized (MSCN) coefficient statistics	SVR	O(d2N) d: Window size
BLIINDS-II	Features based on DCT coefficient	MVG	O((N/d2)log(N/d2)) d: Window size
ILNIQE	DCT coefficient pooling Normalized luminance MSCN products Gradient statistics Log Gabor filter responses Color statistics	Pooling + MVG	O(d2N) d: Window size
Hallucinated IQA	Deep features	NN	—

state-of-the-art methods. Since these are base methods while other new models are somehow the improved versions of them, most of the findings in the field of IQA have been compared in the followings.

3. Proposed method

Using reinforced structure in Hallucination-based papers [7] does not comply with the NR-IQA definition. During the learning process, Hallucinated Generator in [7] and its Quality network can reinforce each other. Thus, this may lead to much more instability in GAN-based models. To resolve this issue, the hallucinated and quality models were separated. The contributions of the proposed methods are as follows: (1) using multiple samples around a submanifold of pristine images conditioned on the distorted image; (2) using the NSS feature as a prior to weighted hypothesis space; and (3) instead of using multiple loss functions, using a perceptual space as a unique space which represents the required information to generate multiple samples. Learning in this space needs alignment and accordingly, an inference network is used to learn these features, which can be considered a divide-and-conquer approach. To elaborate our contributions, the Quality and Hallucinated models will be explained in detail in the following subsections:

3.1. Quality model

The ultimate goal of the BIQA method is to reach a regression model which receives a distorted image and returns a score proportional to its quality. To this end, a deep structure was employed. Given that preparation of a multitude of labeled samples is quite costly, consideration of the prior knowledge is suggested instead to train the quality model. Accordingly, a hallucinated section is used for sampling around pristine manifold to help the regression model. This regression model is called the Quality network. Figure 1 shows the schematic view of the proposed solution.

The Quality network (Q) takes a distorted image and its perceptual representation as a prior in the middle of the network. Perceptual representation is generated based on the distorted images and reflects the information of its corresponding pristine image. To find the perceptual representations, a GAN-based structure was employed, to be discussed in the next section. We aim to find the perceptual representation of some examples of the estimated pristine manifold. Using these examples, we can inject the prior information to lead the search algorithm in the parametric space. The first part of the Quality network (Q_1) extracts a feature representation for a score prediction while the perceptual part (P) contains the information due to the manifold structure. The integration of features from the first part of the Quality network and perceptual part can feed the second part of the Quality network (Q_2):

$$\mathcal{Q} = Q_2(P(I_d, G(I_d)) \otimes (Q_1(I_d)), \tag{1}$$

where I_d is the distorted image, and $\mathcal{Q}$ the overall quality. The generator and perceptual parts can be separately trained through a Hallucinated network to boost the Quality network.

The Quality network consists of some convolution and ReLU layers alongside the downsampling module. Downsampling in this network was constructed based on a pooling mechanism. Further, skip connection or residual network was used to automatically adjust the

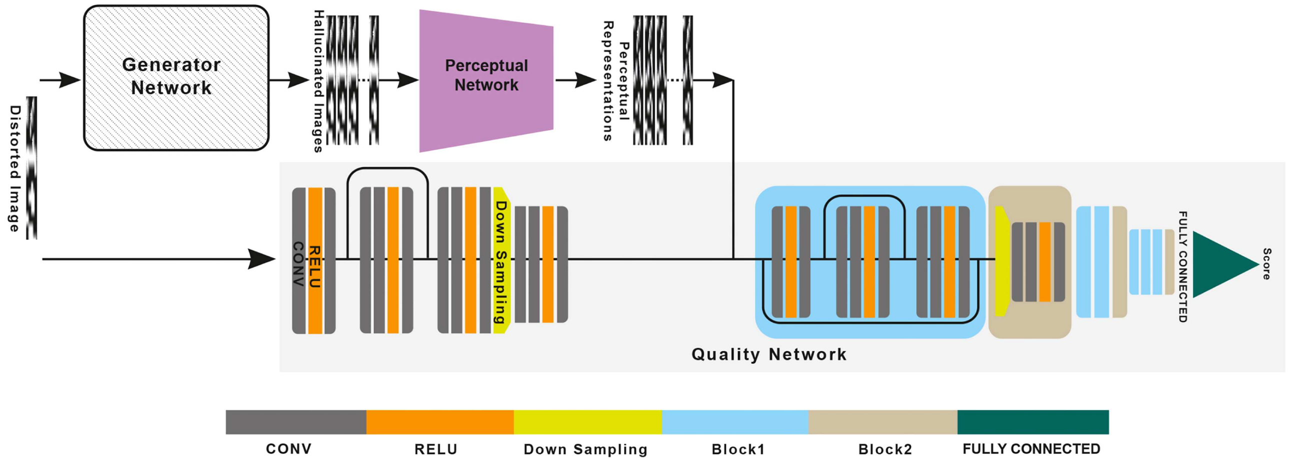

Figure 1. The Quality network (the bottom network) receives the distorted image as an input and returns the score in its output. Also, the extracted features from the perceptual network can be considered as a prior to support training phase with less samples.

deepness of this network regarding the nature of data. Finally, the linear regression loss function is used after fully connecting layers to train this network.

3.2. Hallucinated model

The proposed Hallucinated model consists of four networks: Generator, Perceptual, Judge, and Alignment. Our Hallucinated model was inspired by the main idea behind the GAN's Generative network to generate samples from the desired submanifold of pristine images. In the basic GAN, Generator, and Discriminator contest with each other in a zero-sum game, and the gain of the Generator would be the loss of the Discriminator, and vice versa. The Generator learns to map from a latent space to a data distribution of interest and generates candidates that will be evaluated by Discriminator. Given a training set, the Generator learns to generate new samples with the same statistics as the training set. Conditional GAN is a variation of the basic GAN, which divides data manifold based on a variable and considers a GAN for each part. Indeed, each part of the data manifold is considered conditioning on a discrete random variable. Thus, conditional GAN estimates each part separately instead of estimating the whole distribution of the data manifold.

As we face a continuous space of samples in images, continuous conditional GAN [27] is the best match for our proposal. Despite the discrete conditional GAN where the input was a discrete random variable, the continuous conditional GAN had an input that was a continuous random variable. Therefore, it can fragment the data manifold continuously using a GAN for each part. Sampling from the whole pristine manifold is a crucial task. Since the structure of this manifold is unknown and complicated, finding a map between the Euclidean space and whole manifold needs a high-capacity hypothesis space, huge amount of data, and suitable priors. Moreover, a high-dimensional manifold of pristine images for each distorted image is required, which makes the problem highly ill-posed and undetermined. A distorted image has the potential to be related to several pristine images. In this regard, these variations can be shown by several parameters, implying that the whole pristine manifold is characterized by a complicated structure that is hard to learn. To overcome this dilemma, the authors in this study suggested dividing the whole manifold into corresponding submanifolds and sample from each of them by a GAN fitted on the related submanifold.

To represent such a model, we employed a deep network as a Generator. The input of this network is a distorted image, and the output (hallucinated image) is supposed to be a sample from the pristine manifold corresponding to the available distorted input image. The Generator receives a distorted image as the input and produces a sample from a submanifold that emerges from pristine manifold conditioning on the distorted images. To create more than one image, we can add a noise vector besides the distorted image. Consequently, the network can produce several hallucinated images due to a distorted image. The schematic architecture of the Generator network is shown in Figure 2.

The Generator function takes a pair of distorted image and noise, (I_d, z) and predicts a hallucinated image, I_h:

$$G(I_d, z; \theta) : \mathbb{R}^{d_a} \times \mathbb{R}^{d_z} \to \mathbb{R}^{d_a}, \tag{2}$$

where d_a and d_z are the dimensions of the ambient space and noise, respectively, and θ is the Generator network parameter. This network contains some Convolutional and ReLU layers that facilitate proper feature extraction. Inspired by the study of Berthelot et al. [27], we succeeded in achieving better results

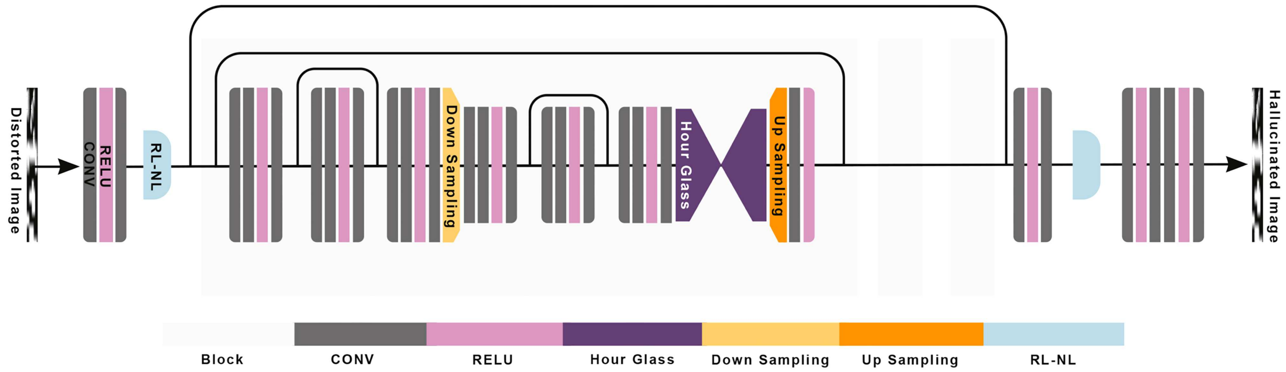

Figure 2. Generator network receives a noise besides the distorted image, generates a hallucinated image using some layers and blocks, and passes it to the perceptual network.

in this network using the Wasserstein GAN (WGAN) objective function.

The convolutional approach is based on the locality concept which reduces the complexity of the problem. The locality is used as a transformed function from the local area of the nonlinear structure. Learning a more complicated function on the whole data structure can be replaced by learning a function on the local part of that structure to solve a problem. This simplification may neglect some features of the complex problems. A review of the super-resolution literature revealed that using only local layers for the Generator network proved to be insufficient [35]. As a result, based on the traditional feature map concept, non-local transformations were introduced. Non-locality differs from the holistic approaches that can be considered the locality in another space, i.e., feature space. The non-local neural network was proposed in [35] to reveal the non-local dependence through the entire image. However, non-local operations at the global level were limited for reasons: First, when we face a large feature size, it imposes a high computational burden on the global-level non-local operations. Second, for low-level tasks, it is better to use non-local operations at a proper neighborhood size [36]. Thus, it is natural to perform Region-Level Non-Local (RL-NL) operations for features with higher spatial resolution or degradation.

Since some perceptual and traditional features were used in the quality network, employing RL-NL features without access to the reference model might lead to difficulty in learning. According to Dai et al. [37], the larger number of distorted images than that of the benign ones at the training phase of the quality network might cause the model to be biased and hence, the non-locality is found while training. To address this issue, in the proposed quality network, we trained an RL-NL layer in the generator network in the first part, which could indirectly bring its advantages to the quality network. In addition, we proposed dividing the feature map into a grid of regions. The $k \times k$ box of RL-NL indicates that the input feature is first divided into a grid of k blocks with the same size and then, each one is processed by the subsequent layers. The RL-NL reveals the forgotten structure of the distorted image features and self-similarities in pristine images.

After non-local operations, the feature representation is non-locally enhanced and fed into the subsequent layers by exploiting the spatial correlations of the features. Moreover, the Generator contains some simplified residual blocks with local-source skip connection, followed by an hourglass stack network to exploit feature inter-dependencies. It was verified that stacking residual blocks would be helpful to form a deep CNN [35,38]. To this end, a stacked hourglass [39] was adopted in the proposed generative network.

The ultimate goal is to use the perceptual interpretation of this submanifold to feed the quality network as a prior. For this reason, our adversarial game was developed in the perceptual space conditioned on a distorted image rather than in the ambient space of the image. To this end, a neural network architecture was used for perceptual mapping from the ambient to perceptual space. In addition, both referenced and hallucinated images were transformed by perceptual network conditioned on the distorted image. Finally, the Judge network decides whether its perceptual input features corresponding to the real or hallucinated image. The perceptual features can be informative whether or not the distribution of the pristine images is reconstructable with a low error rate. Further, an alignment network was employed to align the reconstructed distribution with the distribution of the generated images. Figure 3 shows the overall structure of such a network.

To generate more reliable hallucinated images from a human's perspective, we need to generate some perceptual features based on the distribution of the reconstructed images. To this end, the Perceptual network receives a hallucinated/pristine image conditioned on a distorted image and maps the image submanifold to the corresponding perceptual features. The perceptual network P, parameterized by ϕ, can be

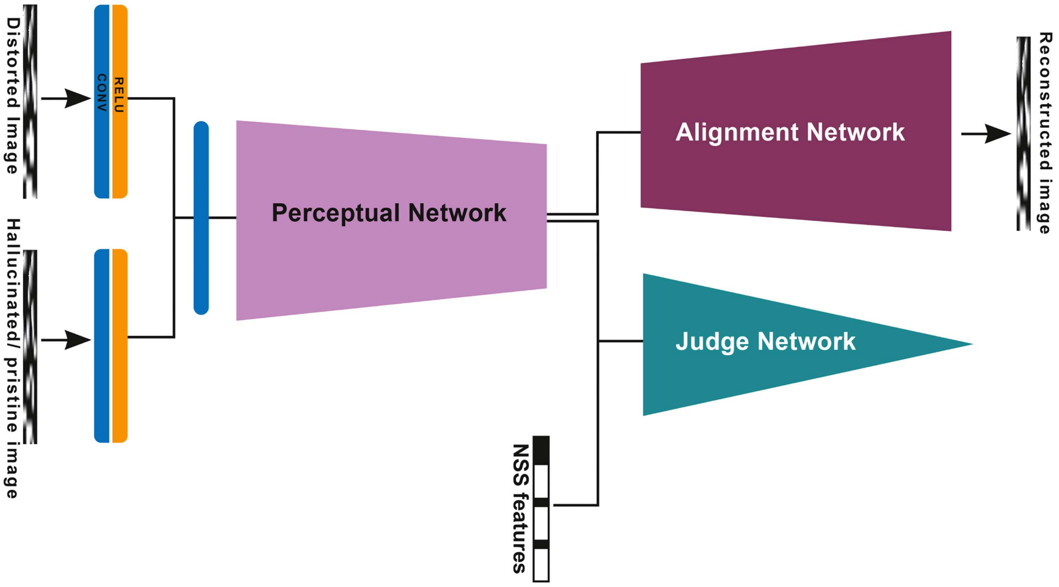

Figure 3. The perceptual network receives hallucinated images conditioned on a distorted image and then, maps the image submanifold to the corresponding perceptual features. The Alignment network will reconstruct the distribution of the mentioned submanifold to align it with the original distribution. Also, The Judge network can discriminate between hallucinated and real images conditioned on the distorted image.

described as:

$$P(I, I_d; \phi) : \mathbb{R}^{d_a} \times \mathbb{R}^{d_a} \rightarrow \mathscr{P}, \tag{3}$$

where d_a is the dimension of the ambient space, ϕ the perceptual network parameter, and $\mathscr{P}$ the perceptual space. To train the perceptual network, it is required that Alignment and Judge networks be used. Some distributional alignments are required to generate meaningful perceptual features that represent the distribution of images as much as possible. The Alignment network brings back the generated perceptual features to the ambient space, aiming to generate the same distribution as the input images. Incorporation of this constraint helps the Judge and Generator networks generate more meaningful hallucinated images. The Alignment network will reconstruct the distribution of the mentioned submanifold to align it with the original distribution. Although there are two samples from the original and reconstructed distributions (two-sample problem), KL-divergence is used to assess the distance of distributions between the mentioned samples. This function can be expressed as:

$$A(P_I; \psi) : \mathscr{P} \rightarrow \mathbb{R}^{d_a}, \tag{4}$$

where A stands for the Alignment network, P_I represents the perceptual feature of specific image I, and ψ contains the Alignment network's parameters. In this study, the KL-divergence minimization between two distributions of the original and reconstructed images was taken into consideration. This KL-divergence constraint can be considered a regularization R of the Generator, Perceptual, and Alignment networks in the objective function. Put $I' = A(P_I)$ and consider $f(.)$ as the corresponding probability density for a random variable related to (.) variable. Here, $\mathbb{E}(.)$ is the expectation of a random variable that can be interpreted in the same way:

$$D_{KL}(f_I || f_{I'}) = \mathbb{E}_I[\ln f_I - \ln f_{I'}]. \tag{5}$$

The objective here is to find a descriptive perceptual feature set that collects the most relevant information. Then, D_{KL} should be minimized related to these perceptual features:

$$D_{KL}(f_I || f_{I'}) = \mathbb{E}_I[\ln f_I - \ln \sum_{P_I} f_{I'|P_I} f_{P_I}]. \tag{6}$$

In order to add this constraint to our GAN-based objective function, it should be approximated through Monte Carlo simulation. To this end, it should be reformulated using Jensen's inequality:

$$D_{KL}(f_I || f_{I'}) \leq \mathbb{E}_I[\ln f_I - \mathbb{E}_{P_I}[\ln f_{I'|P_I}]]. \tag{7}$$

It is enough to minimize the cross-entropy in the upper bound based on P_I. Therefore, the regularization term denoted by $R(G, P, A)$ can be formulated as:

$$R(G, P, A) = -\mathbb{E}_{P(G(I_d,z))}$$
$$[\mathbb{E}_{G(I_d,z)}[\ln f_{A(P(G(I_d,z)))|P(G(I_d,z))}]], \tag{8}$$

where G denotes the Generator network. We can minimize it by considering alignment and perceptual networks. Of note, Judge network J can discriminate between the hallucinated and real images conditioned on the distorted images. The Judge network is similar to the Discriminator of GANs; however, in this case, the perceptual space instead of the ambient space was used.

Non-convergence is an inherent problem with GAN models, especially when addressing a high-dimensional problem. In addition, in the problem at hand, estimation of the parameters is highly related to each other. In the traditional problems, Expectation-Maximization (EM) algorithm was used to address this dilemma. GAN leverages the adversarial approach to solve the problem of the related parameters. Since Bayesian EM utilizes the prior knowledge to overcome the instability problem of the EM algorithm, the NSS features were used in the proposed method as the prior for conveying the search steps.

In order to obtain sufficient knowledge to find a satisfactory transformation, the NSS features were introduced to the Judge network as a prior. One of the contributions of this paper was its application of the Judge network to the Cartesian product of the perceptual and NSS space instead of the ambient space. Use of the NSS features as a prior for the hallucinated model leads to its enhanced performance due to small available data. As presented in several sources [40,41], the NSS features created based on the experts' experiences can be very effective in the IQA tasks. The parameters are randomly initialized to prevent the divergence of learning at the beginning of the training phase. Therefore, NSS features can be considered as the initial knowledge that confine the search space. The Judge Network is parameterized by ψ:

$$J(p, NSS; \psi) : \mathbb{R}^{d_p} \times \mathbb{R}^{d_{NSS}} \rightarrow \mathbb{R}, \qquad (9)$$

where d_p and d_{NSS} are the dimensions of the perceptual and NSS spaces, respectively. The Perceptual network $\mathscr{P}$ projects an image to a latent space $\mathscr{P}$, and J maps the Cartesian product of this latent space and NSS space to $\mathbb{R}$. $\mathscr{P}$ is referred to as the latent perceptual space.

The four networks under study, i.e., G, P, A, and J, are linked together via a game-based objective function. The perceptual representation of a high-resolution image I is $P(I)$ that must be aligned with each submanifold. The constructed hallucinated image is compared to the original one by KL-divergence to ensure equality for the two samples generated in each direction. If the samples follow a Gaussian distribution, minimization of the KL-divergence can be the same as the use of the mean difference to extract the informative features. Since the data distribution of the submanifold of pristine images is far from the Gaussian distribution, the mean difference cannot yield a desirable result; therefore, it is required that the KL-divergence minimization be substitutes. In this case, there are not enough annotated samples to train the model and the mean is not an effective parameter.

For the corresponding input image I_d, the probability distribution is $P(G(I_d, z))$. To learn this part of the network, the following objective function can be used:

$$\begin{aligned} V(G,P,A,J) = \mathbb{E}_{f_{I_d}}[&\mathbb{E}_{I \sim f_I}[\ln J(P(I, I_d), NSS(I))] \\ &+ \mathbb{E}_{z \sim f_Z}[\ln(1 - J(P(G(I_d, z), I_d), \\ &NSS(G(I_d, z)))]]. \end{aligned} \qquad (10)$$

The overall objective function is:

$$\min_{G,P,A} \max_{J} V(G,P,A,J) + \lambda R(G,P,A), \qquad (11)$$

where λ is the only regularization control parameter that can be estimated by cross-validation. The results of the mentioned structure will be investigated in the next section.

4. Experimental results

In this section, three state-of-the-art datasets used to train and evaluate the proposed model are first introduced. Then, the experimental settings are explained, and the evaluation metrics are introduced. Finally, the results of the proposed method are compared with those of some state-of-the-art methods in this realm.

4.1. Datasets

Table 2 introduces three benchmark datasets that have been widely used in the IQA field. To apply

Table 2. The specifications of used dataset for our evaluation and comparison.

Databases	#Ref. images	#Dist. images	#Dist. type	Score range
TID2013 [42]	25	3000	24	[0,9]
TID2008 [43]	25	1700	17	[0,9]
LIVE [44]	29	779	5	[1,100]

the rating results, TID2013 and TID2008 use Mean Opinion Scores (MOS). Instead of directly applying rating results, LIVE uses differences in quality between images as Difference Mean Opinion Scores (DMOS).

4.2. Evaluation metrics

Like the other research studies in this field, this study also employed Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank Order Correlation Coefficient (SROCC) to evaluate the performance of the proposed method. While PLCC measures the linear correlation between the predicted score and ground-truth, SROCC considers the monotonic relationship between the predicted score and ground-truth [19,45]. PLCC can be calculated through:

$$PLCC = \frac{\sum_{i=1}^{N}(s_i - \mu_{s_i})(\hat{s}_i - \mu_{\hat{s}_i})}{\sqrt{\sum_{i=1}^{N}(s_i - \mu_{s_i})^2 \sum_{i=1}^{N}(\hat{s}_i - \mu_{\hat{s}_i})^2}}, \quad (12)$$

where N is the number of test samples, s_i and $\hat{s}_i$ are the scores of ground truth and predicted quality of the ith image, and μ_{si} and $\mu_{\hat{s}_i}$ are their average values, respectively. SROCC can be calculated through:

$$SROCC = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)}, \quad (13)$$

where d_i is the difference between the ranks of the ith test image in predicted quality scores and ground truth. Both PLCC and SROCC vary from -1 to 1, and the higher the absolute value, the better the prediction performance.

4.3. Experimental settings

The input shape for the proposed network is 256×256. Therefore, image patches in this size should be sampled from the original images in each dataset. Then, a data augmentation with the flip and random rotation of (±20) degrees was used to collect more input samples to train the model. The reason for choosing these augmentation techniques is that they do not change the quality of images and can be considered part of each image. Finally, the model is trained from scratch with a mini-batch size of 32.

In order to train the Quality model, Stochastic Gradient Descent (SGD) optimization function was utilized. It was trained with an initial learning rate of $\alpha_1 = 10^{-2}$ and a drop of 0.1 for every 10000 iterations. A weight decay of 0.0005 and a momentum of 0.9 were considered in the training procedure.

The Hallucinated model is trained using Adam optimization procedure. The corresponding parameters were $\alpha = 10^{-4}$, $\beta_1 = 0$, and $\beta_2 = 0.9$. Despite using a batch-normalization, a weight decay of $\lambda = 10^{-3}$ was applied. The mentioned parameters were suggested to optimize the WGAN network [46]. The overlapped image patches at a fixed stride from each image were extracted to train and test the model. The average of all predicted scores was used to calculate the image quality score due to each image. The code was implemented in the PyTorch environment and run by an NVIDIA GTX 1080 Ti GPU and 128 GB of RAM.

4.4. Results

The results obtained from the proposed model on the mentioned datasets were compared with those from four state-of-the-art general-purpose NR-IQA methods, i.e., BLIINDS-II [47], BRISQUE [48], ILNIQE [41], and Hallucinated IQA [7].

In order to train TID2013 [42] and TID2008 [43], Leave-One Distortion-Out cross-validation was performed. To implement this method, the samples with 23 types of distortions from TID2013 [42] were used for training, and the samples were left to test the model due to the remaining distortion. For TID2008 [43], the samples were considered as training due to 16 distortions, and the rest of them were performed in testing. In the proposed approach, first, the MOS scores of the images should be normalized. To ensure a fair comparison, the results of the rival methods released by original authors were brought in the tables under the same subsampling strategy. Tables 3 and 4 exhibit the performance evaluation of the proposed method, compared with the aforementioned state of the art due to TID2013 [42] and TID2008 [43], respectively. In addition, SROCC was used for evaluation.

As observed in Table 3, on average, the proposed method for TID2013 outperforms the state-of-the-art methods. The achieved improvement in the average is 0.031 compared to the rival method, i.e., hallucinated IQA. In 13 out of 24 distortions available in the TID2013, the proposed method obtained better SROCC than others. While the Hallucinated model, ILNIQE, and BlindsII were the best in two cases, Brisque was the best in five items. The best result in each row is given in bold. In each case, the samples due to the mentioned distortion were set aside for testing, while all the other samples were used for model training.

As shown in Table 4, the average result of the proposed method for TID2008 is 0.098 more than that of the best rival method, i.e., hallucinated IQA. Further, in 10 distortions out of 17 in this dataset, the proposed method performs the best SROCC score, which is indicated in bold. The related samples were set aside for testing for each distortion, while all the other samples were used for model training.

To evaluate the methods on the LIVE dataset, all images, regardless of their distortion type, were randomly divided into 80% and 20% due to training and testing samples, respectively. To avoid the bias of randomness, experiments were repeated 10 times.

Table 3. SROCC in the TID2013 database. Comparison of the results of the proposed method with the rival methods in this realm.

Dist. type	**BLIINDS-II [47]**	**BRISQUE [48]**	**ILNIQE [41]**	**Hallucinated IQA [7]**	**Our method**
AGN	0.765	**0.932**	0.856	0.923	0.912
ANC	0.832	0.806	0.811	0.880	**0.892**
SCN	0.639	0.538	0.889	0.945	**0.948**
MN	0.203	0.573	0.509	0.673	**0.705**
HFN	0.710	0.888	0.812	**0.955**	0.942
IN	0.503	0.642	0.743	0.810	**0.827**
QN	0.291	0.645	**0.869**	0.855	0.859
GB	**0.899**	0.850	0.776	0.832	0.842
DEN	0.741	0.609	0.744	0.957	**0.960**
JPEG	0.751	0.493	0.834	0.914	**0.919**
JP2K	0.821	0.759	**0.857**	0.624	0.824
JGTE	0.409	**0.558**	0.282	0.460	0.515
J2TE	0.721	0.712	0.521	0.782	**0.803**
NEPN	0.116	0.307	–0.093	0.664	**0.669**
Block	**0.270**	0.224	–0.131	0.122	0.247
MS	0.093	**0.191**	0.184	0.182	0.186
CTC	0.311	0.015	0.014	0.376	**0.398**
CCS	0.039	0.208	–0.160	0.156	**0.224**
MGN	0.719	**0.865**	0.651	0.850	0.857
CN	0.078	0.466	0.331	**0.614**	0.522
LCNI	0.416	0.818	0.828	0.852	**0.870**
CQD	0.736	0.476	0.748	0.911	**0.916**
CHA	0.532	**0.747**	0.672	0.381	0.449
SSR	0.737	0.772	0.862	0.616	**0.869**
Average	0.521	0.597	0.551	0.683	**0.714**

Finally, the average results of considering SROCC and PLCC are presented in Tables 5 and 6, respectively.

In Table 5, in three distortions out of five available ones in the LIVE dataset, our proposed method performs the best SROCC score, which is shown bold. In addition, our average SROCC was more satisfactory than all others. After the proposed method, in the case of the two items hallucinated, IQA obtained the best result.

Table 6, similar to Table 5, shows the evaluation of methods in the LIVE dataset and PLCC is used for this evaluation. Our proposed method outperforms all others in 3 out of 5 available distortions in the LIVE dataset. The hallucinated IQA was the best in one item while the two mentioned methods were commonly the best in one distortion. However, the average of PLCC for our method was better than all while being very close to that of the hallucinated IQA.

5. Conclusion and future work

In this paper, a deep-based method was proposed to assign a score to a distorted image to solve the NR-IQA problem. This score showed the distance of a distorted image from the corresponding pristine image without reference distortion. In this study, the simulation of the ability of the human visual system was taken into account to overcome the problem due to the insufficiency of the available samples. For this reason, the pristine data manifold was divided into some submanifolds corresponding to each distorted image. Then, multiple hallucinated images were sampled from each submanifold and transformed into a perceptual space through distributional alignment. Finally, the model was examined in three benchmark datasets, which yielded significant results, compared to the rival methods in this realm.

Table 4. SROCC in the TID2008 database. Comparison of the results of the proposed method with the rival methods in this realm.

Dist. type	BLIINDS-II	BRISQUE	Hallucinated IQA	Our method
AGN	0.567	0.660	**0.927**	0.916
ANC	0.488	0.317	0.898	**0.903**
SCN	0.823	0.799	**0.940**	0.922
MN	0.344	0.220	0.747	**0.774**
HFN	0.803	0.841	**0.967**	0.823
IN	0.760	0.830	0.940	**0.952**
QN	0.673	0.690	0.714	**0.720**
GB	0.544	**0.810**	0.618	0.623
DEN	0.599	0.445	0.917	**0.928**
JPEG	0.808	0.821	**0.937**	0.891
JP2K	**0.772**	0.745	0.610	0.622
JGTE	0.321	0.279	0.628	**0.643**
J2TE	0.597	**0.740**	0.381	0.466
NEPN	0.388	0.130	0.733	**0.740**
Block	0.302	0.316	0.331	**0.367**
MS	0.265	**0.305**	0.275	0.267
CTC	0.272	0.091	0.357	**0.391**
Average	0.564	0.521	0.742	**0.840**

Table 5. Comparing the results of the proposed method considering SROCC with the rival methods in this realm using LIVE database.

Dist. type	BLIINDS-II	BRISQUE	ILNIQE	Hallucinated IQA	Our method
JP2K	0.928	0.914	0.893	**0.983**	0.942
JPEG	0.942	0.965	0.941	0.961	**0.973**
WN	0.969	0.979	0.980	0.984	**0.987**
BLUR	0.923	0.951	0.915	0.983	**0.986**
FF	0.889	0.877	0.832	**0.989**	0.923
Average	0.930	0.937	0.912	0.980	**0.984**

Table 6. Comparing the results of the proposed method considering PLCC with the rival methods in this realm using LIVE database.

Dist. type	BLIINDS-II	BRISQUE	Hallucinated IQA	Our method
JP2K	0.935	0.923	0.977	**0.978**
JPEG	0.968	0.973	0.984	**0.986**
WN	0.981	0.985	**0.993**	0.988
BLUR	0.939	0.951	**0.990**	**0.990**
FF	0.895	0.903	0.960	**0.966**
All	0.943	0.942	0.982	**0.984**

To regularize the model, especially at the beginning of the training phase, the Natural Scene Statistics (NSS) features were used as prior knowledge to initiate the learning process in the proposed model that helped overcome the problem of divergence and drive the learning process on the right track. For the future studies, we recommend developing a pre-trained model instead of using NSS features to conduct the learning process in the right track.

References

1. Kim, J., Zeng, H., Ghadiyaram, D., et al. "Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment", *IEEE Signal Processing Magazine*, **34**(6), pp. 130–141 (2017).
2. Guo, Y., Ding, G., and Han, J. "Robust quantization for general similarity search", *IEEE Transactions on Image Processing*, **27**(2), pp. 949–963 (2017).
3. Dong, C., Loy, C.C., He, K., et al. "Image super-resolution using deep convolutional networks", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **38**(2), pp. 295–307 (2015).
4. Kim, J. and 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*, pp. 1676–1684 (2017).
5. Golestaneh, S. and Karam, L.J. "Reduced-reference quality assessment based on the entropy of dwt coefficients of locallyweighted gradient magnitudes", *IEEE Transactions on Image Processing*, **25**(11), pp. 5293–5303 (2016).
6. Ye, P., Kumar, J., Kang, L., et al. "Unsupervised feature learning framework for no-reference image quality assessment", In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1098–1105, IEEE (2012).
7. Lin, K.-Y. and Wang, G. "Hallucinated-iqa: No-reference image quality assessment via adversarial learning", In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 732–741 (2018).
8. Li, L., Zhu, H., Yang, G., et al. "Referenceless measure of blocking artifacts by tchebichef kernel analysis", *IEEE Signal Processing Letters*, **21**(1), pp. 122–125 (2013).
9. Li, L., Lin, W., Wang, X., et al. "No-reference image blur assessment based on discrete orthogonal moments", *IEEE Transactions on Cybernetics*, **46**(1), pp. 39–50 (2015).
10. Liu, H., Klomp, N., and Heynderickx, I. "A no-reference metric for perceived ringing artifacts in images", *IEEE Transactions on Circuits and Systems for Video Technology*, **20**(4), pp. 529–539 (2009).
11. Saad, M.A., Bovik, A.C., and Charrier, C. "Blind image quality assessment: A natural scene statistics approach in the DCT domain", *IEEE Transactions on Image Processing*, **21**(8), pp. 3339–3352 (2012a).
12. Mittal, A., Moorthy, A.K., and Bovik, A.C. "No-reference image quality assessment in the spatial domain", *IEEE Transactions on Image Processing*, **21**(12), pp. 4695–4708 (2012a).
13. Xue, W., Zhang, L., and Mou, X. "Learning without human scores for blind image quality assessment", In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 995–1002 (2013).
14. Ghadiyaram, D. and Bovik, A.C. "Perceptual quality prediction on authentically distorted images using a bag of features approach", *Journal of Vision*, **17**(1), pp. 32–32 (2017).
15. He, K., Zhang, X., Ren, S., et al. "Deep residual learning for image recognition", In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016).
16. Ding, G., Chen, W., Zhao, S., et al. "Real-time scalable visual tracking via quadrangle kernelized correlation filters", *IEEE Transactions on Intelligent Transportation Systems*, **19**(1), pp. 140–150 (2017).
17. Ding, G., Guo, Y., Chen, K., et al. "DECODE: Deep confidence network for robust image classification", *IEEE Transactions on Image Processing*, **28**(8), pp. 3752–3765 (2019).
18. Kang, L., Ye, P., Li, Y., et al. "Convolutional neural networks for noreference image quality assessment", In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1733–1740 (2014).
19. Yan, B., Bare, B., and Tan, W. "Naturalness-aware deep noreference image quality assessment", *IEEE Transactions on Multimedia*, **21**(10), pp. 2603–2615 (2019).
20. Gu, K., Zhai, G., Yang, X., et al. "Using free energy principle for blind image quality assessment", *IEEE Transactions on Multimedia*, **17**(1), pp. 50–63 (2014).
21. Sheikh, H. "LIVE image quality assessment database release 2", http://live. ece. utexas. edu/research/quality (2005).
22. Ponomarenko, N., Jin, L., Ieremeiev, O., et al. "Image database TID2013: Peculiarities, results and perspectives", *Signal Processing: Image Communication*, **30**, pp. 57–77 (2015).
23. Deng, J., Dong, W., Socher, R., et al. "Imagenet: A large-scale hierarchical image database", In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, *Ieee* (2009).
24. Bianco, S., Celona, L., Napoletano, P., et al. "On the use of deep learning for blind image quality assessment", *Signal, Image and Video Processing*, **12**(2), pp. 355–362 (2018).

25. Talebi, H. and Milanfar, P. "NIMA: Neural image assessment", *IEEE Transactions on Image Processing*, **27**(8), pp. 3998–4011 (2018).

26. Hazrati Fard, S.M. and Hashemi, S. "Proposing a sparse representational based face verification system to run in a shortage of memory", *Multimedia Tools and Applications*, **79**(3), pp. 2965–2985 (2020).

27. Berthelot, D., Milanfar, P., and Goodfellow, I. "Creating high resolution images with a latent adversarial generator", arXiv preprint arXiv:2003.02365 (2020).

28. Gu, S., Bao, J., Chen, D., et al. "GIQA: Generated image quality assessment", In *European Conference on Computer Vision*, pp. 369–385, Springer (2020).

29. Saad, M.A., Bovik, A.C., and Charrier, C. "Blind image quality assessment: A natural scene statistics approach in the DCT domain", *IEEE Transactions on Image Processing*, **21**(8), pp. 3339–3352 (2012b).

30. Ye, P. and Doermann, D. "No-reference image quality assessment using visual codebooks", *IEEE Transactions on Image Processing*, **21**(7), pp. 3129–3138 (2012).

31. Zhang, P., Zhou, W., Wu, L., et al. "SOM: Semantic obviousness metric for image quality assessment", In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2394–2402 (2015a).

32. Moorthy, A.K. and Bovik, A.C. "A two-step framework for constructing blind image quality indices", *IEEE Signal Processing Letters*, **17**(5), pp. 513–516 (2010).

33. Zhang, W., Zhai, K., Zhai, G., et al. "Learning to blindly assess image quality in the laboratory and wild", In *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 111–115, IEEE (2020).

34. Arjovsky, M. and Chintala, S. "Wasserstein GAN", arXiv preprint arXiv:1701.07875 (2017).

35. Zhang, Y., Tian, Y., Kong, Y., et al. "Residual dense network for image super-resolution", In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2472–2481 (2018).

36. Liu, D., Wen, B., Fan, Y., et al. "Non-local recurrent network for image restoration", In *Advances in Neural Information Processing Systems*, pp. 1673–1682 (2018).

37. Dai, T., Cai, J., Zhang, Y., et al. "Second-order attention network for single image super-resolution", In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11065–11074 (2019).

38. Lim, B., Son, S., Kim, H., et al. "Enhanced deep residual networks for single image super-resolution", In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 136–144 (2017).

39. Newell, A., Yang, K., and Deng, J. "Stacked hourglass networks for human pose estimation", In *European Conference on Computer Vision*, pp. 483–499, Springer (2016).

40. Mittal, A., Soundararajan, R., and Bovik, A.C. "Making a completely blind image quality analyzer", *IEEE Signal Process. Lett.*, **20**(3), pp. 209–212 (2013).

41. Zhang, L., Zhang, L., and Bovik, A.C. "A feature-enriched completely blind image quality evaluator", *IEEE Transactions on Image Processing*, **24**(8), pp. 2579–2591 (2015b).

42. Ponomarenko, N., Ieremeiev, O., Lukin, V., et al. "Color image database TID2013: Peculiarities and preliminary results", In *European Workshop on Visual Information Processing (EUVIP)*, pp. 106–111, IEEE (2013).

43. Ponomarenko, N., Lukin, V., Zelensky, A., et al. "TID2008-a database for evaluation of full-reference visual quality assessment metrics", *Advances of Modern Radioelectronics*, **10**(4), pp. 30–45 (2009).

44. Sheikh, H.R., Sabir, M.F., and Bovik, A.C. "A statistical evaluation of recent full reference image quality assessment algorithms", *IEEE Transactions on Image Processing*, **15**(11), pp. 3440–3451 (2006).

45. Bosse, S., Maniry, D., Müller, K.-R., et al. "Deep neural networks for noreference and full-reference image quality assessment", *IEEE Transactions on Image Processing*, **27**(1), pp. 206–219 (2017).

46. Gulrajani, I., Ahmed, F., Arjovsky, M., et al. "Improved training of wasserstein GANs", arXiv preprint arXiv:1704.00028 (2017).

47. Saad, M.A., Bovik, A.C., and Charrier, C. "A DCT statistics-based blind image quality index", *IEEE Signal Processing Letters*, **17**(6), pp. 583–586 (2010).

48. Mittal, A., Moorthy, A.K., and Bovik, A.C. "No-reference image quality assessment in the spatial domain", *IEEE Transactions on Image Processing*, **21**(12), pp. 4695–4708 (2012b).

Biographies

Zahra Javidian earned her BSc degree in 2005 from Islamic Azad University Meybod Branch, Yazd and her MSc degree in 2013 from Tarbiat Modares University, Tehran. Now, she is a PhD student at the International University of Shiraz and is about to defend her thesis. Her main interest is in machine learning applications in image processing and vision.

Sattar Hashemi received a PhD in Computer Science from the Iran University of Science and Technology in conjunction with Monash University, Australia in 2008. Following academic appointments at Shiraz University, he is currently an Associate Professor at the Computer Department at Electrical and Computer Engineering School, Shiraz University, Shiraz, Iran. He is recognized for his contributions to machine learning and data mining. He has published many refereed papers and book chapters on data stream classification, social networks, database intrusion detection, and computer security.

Seyed Mehdi Hazrati Fard received his MSc and PhD in Artificial Intelligence from Shiraz University. He had his first postdoc at the University of Guelph and currently, he is a postdoc researcher at the University of Victoria. His primary interest is in machine learning applications in fields such as providing security and privacy in smart environments and using machine learning in forecasting and classification. His positions as R&D Head in companies such as Irancell and Soshianest Co. helped him build a strong background in pragmatic applications of machine learning.