RESEARCH

A Deep Superpixel-based Network for Blind Image Quality Assessment

Yunpeng Liu and Guangyi Yang

Abstract

The primary goal of blind image quality assessment (BIQA) is to automate the manual process of evaluating images. Although existing techniques have effectively identified degradations, they typically do not consider the full semantic content, resulting in distortions. To address this issue, we propose a deep adaptive superpixel-based network, namely DSN-IQA, to assess the quality of images using multi-scale and superpixel segmentation. The proposed DSN-IQA can adaptively accept images of arbitrary scale as input samples, in an assessment process similar to that of human perception. The network applies two models to extract multi-scale semantic features and generate a superpixel adjacency map. These two elements are then united and feature fusion is used to accurately predict image quality. Experimental results from multiple benchmark databases demonstrated that our method was highly competitive in assessing challenging authentic images. In addition, our model accurately assessed images exhibiting complicated distortions, much like the human eye.

Keywords: Image quality assessment; superpixel; multiscale features; semantic features; arbitrary scale input.

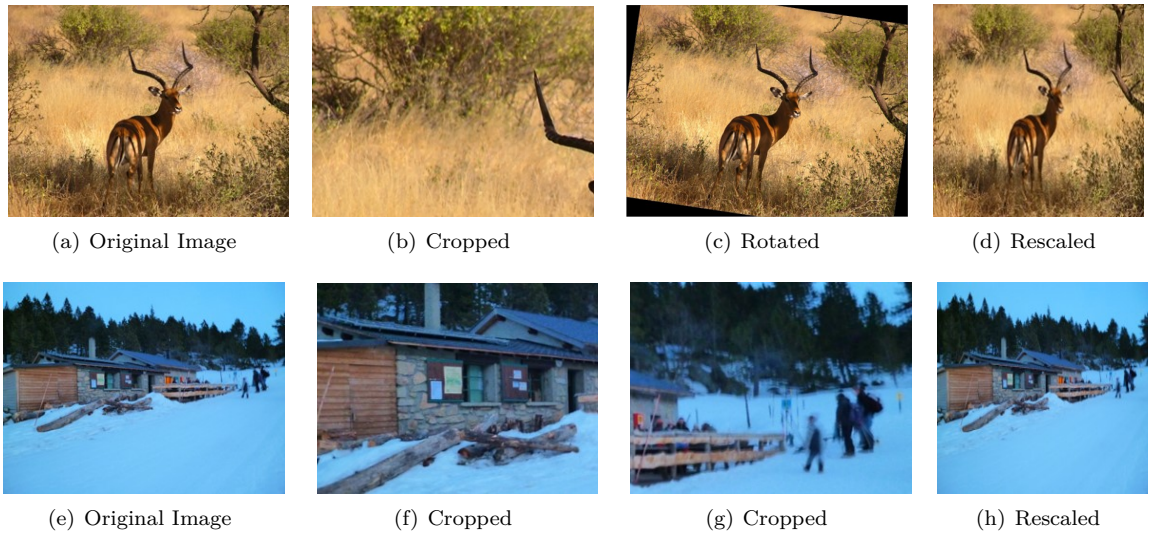
1 Introduction

The unprecedented development of communication technologies has underscored the role of images as the primary carrier of visual information [1]. In many cases, the quality of an image is correlated with content coherence, since distortions have a significant negative effect on the readability of an image. However, nearly any stage in common image acquisition, transmission, and storage processes can produce varying degrees of distortion [2]. Consequently, image quality assessment (IQA) is critical to assuring the reliability of image processing systems. As a result, research into novel IQA methods has received widespread attention in recent years.

IQA algorithms can be divided into two types: subjective and objective assessments, typically depending on whether the classification process requires manual intervention [3]. Subjective assessment utilizes intuitive human visual experience as the standard metric [4]. This is the most accurate type of evaluation, since perceptual image quality is determined by the human visual system (HVS). However, this process is also time-consuming, expensive, and highly labor-intensive, preventing routine implementation for most tasks. As such, objective assessments are more practical and are widely used for large sample sets, since image quality can be automatically predicted using mathematical models. Objective assessment is often divided into three categories, determined by the presence or absence of a reference image: full-reference IQA (FR-IQA), reduced-reference IQA (RR-IQA), and no-reference or blind IQA (BIQA) [3]. In practical applications, a reference image is often not provided, which impedes the potential scope of FR-IQA and RR-IQA. As such, BIQA has attracted increased interest among researchers [5].

Feature extraction algorithms have made BIQA models more widely applicable. However, current algorithms explicitly designed for this process exhibit certain limitations. For example, some BIQA techniques adopt low-level features and employ machine learning to assess quality. Learning-based regression models have also been employed, trained by a set of features extracted from images whose mean opinion scores (MOS) or different MOS (DMOS) metrics were acquired in subjective experiments. The resulting model can then be used to predict a ground truth MOS. BIQA models can also apply the principles of natural scene statistics (NSS) to successfully represent overall image quality. However, this approach is often ineffective for evaluating local distortions in an image. To address this issue, feature extractors based on deep convolutional neural networks (CNNs) have been proposed and are widely used among researchers [6, 7]. These algorithms automatically capture deep features used to represent degradations and, as a result, have been widely employed in BIQA tasks. However, one of the major issues with CNN-based image quality assessment is the attention of our eyes is not evenly distributed across different regions in an image. Ignoring these varying weights in different sections adds uncertainty to the assessment process, since HVS results differ from predictions generated by a CNN model.

IQA methods based on deep learning encounter two inevitable problems: a shortage of data and fixed-scale



(a) Original Image (b) Cropped (c) Rotated (d) Rescaled

(e) Original Image (f) Cropped (g) Cropped (h) Rescaled

Fig. 1 Examples of pre-processing steps that can affect the quality of an image, including cropping, rotation, and rescaling. (a), (e) The original images. (b), (f)-(g) Cropped images in which varying levels of distortion are evident and the content has visually changed. (c) A rotated image in which a non-physical black background is introduced to pad the edges. (d), (h) Rescaled images exhibiting conspicuous deformations.

input required by end-to-end models. As such, a variety of pre-processing techniques have been developed to address these issues. However, these methods often decrease consistency between images and their corresponding ground-truth scores, as shown in Fig. 1. Cropped images, like those shown in Figs. 1(b) and 1(f), include content clipped from the original samples. For instance, Fig. 1(b) displays grass while 1(a) contains both grass and an antelope. Two cropped images from the same original sample also exhibit differing levels and types of degradations, as seen in Figs. 1(f) and 1(g), where the latter includes more severe blurring. Fig. 1(c) demonstrates that rotation of an image may introduce unreal coloration, which can negatively affect image quality. Rescaling also deforms the subject of an image, thereby affecting both the semantic content and quality, as shown in Figs. 1(d) and 1(h). As a result, a ground-truth quality may no longer be suitable for pre-processed images used to train a neural network, which will inevitably lead to bias in predictions and a less objective classification result.

In this paper, we propose a deep superpixel-based network for BIQA (DSN-IQA), which is more consistent with human visual perception. This approach builds on previous work by developing a CNN-based network that extracts multi-scale semantic features, which are then fused with an adjacency map acquired from a superpixel segmentation model [8, 9]. When humans evaluate an image, they simultaneously pay attention to semantic information and local details in the image, to determine its quality. Our proposed network mimics this assessment technique. Specifically, superpixel segmentation was included to focus the network on local adjacency information. In this process, input images are segmented into superpixels (perceptually meaningful blocks comprised of spatially neighboring groups). These pixels share similar local colors and serve as low-dimensional image representations. As such, they offer detailed information that can be used for subsequent quality prediction. Consequently, our approach simultaneously utilizes local superpixels and multi-scale semantic features to ensure the image evaluation process more closely resembles the HVS. This technique also addresses pre-processing problems by accepting images of arbitrary size as input, since humans assess whole pictures.

The proposed IQA method was evaluated using multiple databases. In these experiments, test images remained at their original sizes to ensure quality scores accurately represented the true quality. These tests included individual database, cross-database, individual distortion type, and ablation experiments. Results demonstrated the proposed adaptive model could process complicated distortions, achieving high accuracy in the prediction of image quality. The novelty of this technique lies in the utilization of superpixel-based information extraction and sufficient multi-scale features. The primary contributions of our study can be summarized as follows:

- To the best of our knowledge, the proposed method is the first to apply superpixel segmen-

tation to BIQA, in the extraction of local features and multi-scale semantic features. Experiments demonstrated these features are highly consistent with the HVS.
- The influence of training image cropping on the full image evaluation process was analyzed. Pooling layers were also adjusted to design a method that can overcome problems associated with image size.
- The results of validation experiments indicated our method outperforms comparable techniques, as measured by quality prediction, successfully processing images with complicated distortions.

The remainder of this paper is organized as follows. Section 2 reviews the development of superpixel segmentation for IQA and CNN-based BIQA methods. Section 3 describes the development of the DSN-IQA model. Section 4 provides extensive experimental results and a comparative analysis of the proposed technique. Section 5 summarizes our work and draws conclusions.

2 Related Work

2.1 Superpixel segmentation for IQA

Superpixels, as defined by Ren *et al.* [10] in 2003, refer to irregular pixel blocks exhibiting certain visual significance. These structures are composed of adjacent pixels with similar texture, color, brightness, and other characteristics. Superpixel segmentation involves a small number of visually meaningful superpixels used to reduce data volume. Conventional segmentation algorithms do not utilize CNNs, instead relying on statistical models that can be divided into two groups: graph-based and gradient-ascent-based techniques. Graph methods, which utilize data structures containing vertices and weighted edges, segment images by minimizing a cost function [11, 12]. Common examples include normalized cuts [13], graph cuts [14], and entropy rate superpixel segmentation algorithms [15]. Gradient ascent is an iterative process that clusters pixels by relying on shifts between groups with similar values [12]. A number of these are currently in use, including watershed [16], mean shift [17], quick shift [18], Turbopixel [19], and simple linear iterative clustering (SLIC) [12]. SLIC, specifically, performs segmentation by clustering pixels using color and distance similarities. Most superpixel segmentation algorithms are unsupervised and produce superpixels of uniform size with regular shapes. These structures exhibit a clear visual interpretation and are widely used in computer vision preprocessing.

IQA involves two primary modifications that have improved superpixel assessment accuracy. First, superpixels can reduce pixel redundancy, allowing the automated assessment process to be more perceptive. Many common IQA methods use square convolution kernels to ensure that a sufficient number of features are extracted from images used for quality prediction [6, 20, 21]. However, a 3×3 square kernel, for example, concentrates only on a single small region at any given time, thus losing visual meaning [22]. As such, square kernels do not exploit connections between adjacent pixels and thereby contribute to information redundancy. In contrast, superpixel segmentation functions more like the human visual system. When humans observe and assess an image, similar adjacent pixels are recognized and gathered into a single local region [23]. Superpixels also allow IQA methods to assess regional differences. The superpixel-based similarity index (SPSIM) [22], proposed by Sun *et al.*, demonstrates that different region types exhibit distinct noise responses. For example, textured areas are more resistant to Gaussian noise than flat areas, while the situation is reversed for image blurring. If the extracting network ignores these meaningful effects in certain areas, it can lose some common local details and the predicted outcome may deviate from a subjective manual score as a result. As such, superpixels are a vital tool for improving IQA.

Other superpixel-based segmentation methods have been developed in recent years. For instance, SPSIM, a full-reference IQA model based on SLIC, segments reference images without distortions. Distorted targets are then segmented into visually meaningful regions using the generated superpixels. Mean values for the intensity and chrominance components are accordingly extracted within each superpixel and compared to reference and target images used to precisely describe a local similarity index. Frackiewicz *et al.* developed an improved SPSIM index for IQA [24]. Their approach revised SPSIM in two ways: a new color space replaced the YUV convention, and a novel calculation method was used to define a mean deviation similarity index (MDSI) [25]. Fang *et al.* [26] used SLIC to distinguish between fused and exposed images. In this way, quality maps were calculated from a Laplacian pyramid that separately described regions exhibiting large and small changes, employing different regional strategies in each. All of these techniques implement regional solutions to IQA by deploying superpixels.

While superpixels increase consistency between segmentation methods and the HVS, other issues remain. For instance, it is difficult to combine non-CNN segmentation algorithms with CNN-based models, since neural networks involve dimensional feature tensors with visual meaning. As such, these tensors cannot simply be combined with labeled superpixels generated by non-learning algorithms. In addition, since

CNNs utilize back propagation in the training step, these non-differentiable models prevent training [27]. However, using CNNs to directly segment superpixels can overcome this limitation. After comparing several machine learning models, we opted for the superpixel segmentation via CNN (SSVCNN) technique proposed by Teppei [28]. This unsupervised algorithm optimizes a randomly initialized network, making it easier to integrate with existing image quality models. The output is then a clear and meaningful probabilistic map representing the degree of belonging for individual pixels.

2.2 CNN-based BIQA

Insufficient computational power prevented early CNN-based IQA models from predicting quality with a single network. These techniques initially extracted only hierarchical features, as subsequent operations were used to calculate quality from feature sets [29], an approach that differs from end-to-end models [7]. Tang *et al.* [30] proposed a non-end-to-end approach that utilized a radical basis function to pre-train a deep belief network using unlabeled data, later fine-tuning with labeled data. In addition, Bianco *et al.* [20] adopted CNN features pre-trained on image classification tasks as input, establishing a quality evaluator that employed support vector regression (SVR) [31] to generate quality scores from feature sets. In this process, mean opinion scores (MOS) were quantified into five categories. Pre-trained features were then fine-tuned using multi-category classification settings and fed to the SVR model. However, this approach cannot be optimized in an end-to-end process, since several manual parameter adjustments are necessary.

A variety of end-to-end BIQA models have since been developed, aided by increased computational power and deep CNNs [32]. Kang *et al.* proposed the CNNIQA [6] algorithm, which accepts image patches as input and employs back propagation for training. Since feature extraction and regression are integrated into the CNN, the depth of the neural network can be increased to improve learning ability. Kang *et al.* also proposed CNNIQA++ [33], which featured an increased number of convolutional layers for simultaneous estimation of image quality and distortion type. Zhang *et al.* proposed the deep bilinear CNN (DBCNN) method [34] based on VGG-16 [35], which was initially designed for image recognition but could be fine-tuned to assess image quality. This algorithm includes two deep CNN branches, specializing in separately assessing distortion scenarios (both synthetically and authentically) using bilinear pooling to fuse the network branches. However, this approach cannot accurately predict authentic databases containing images of complicated objects.

In recent years, semantic feature-based BIQA models have become an active area of research because of their ability to perceive semantic information and more accurately identify authentic image databases. For example, Kim *et al.* [36] introduced ResNet [37], a deep semantic CNN trained with classification databases to assist in improving IQA accuracy. Hosu *et al.* tested several deep CNNs and confirmed the advantages of semantic features for processing authentic IQA samples [38]. The semantic feature aggregation (SFA) method developed by Li *et al.* applies statistics from ResNet-50 multi-patch features for quality prediction [39]. The authors also suggested the content of images has an impact on quality prediction, by demonstrating that people will score an image of a clear blue sky as high quality, while a traditional prediction model mistakenly recognizes it as blur noise. This effect can be explained by semantic losses incurred during feature extraction. Considering that content varied from image to image, Su *et al.* proposed Hyper-IQA [8], which separates the IQA process into three stages: content understanding, perception rule learning, and quality prediction. A hyper network connection was included to mimic this mapping from image content to quality perception. Thus, employing semantic features in IQA allows the automated assessment process to be more like a human evaluation of semantic content and image quality. However, existing semantic models often ignore local visual details in images. Superpixels, as illustrated in Section 2.1, can overcome this inherent limitation. Our proposed adaptive method not only employs multi-scale semantic features, it also extracts superpixel information to imitate the HVS. These two innovations improve the consistency of the assessment process between our method and the human eye.

3 Proposed Method

3.1 Algorithm framework

The proposed technique includes two components: a multi-scale feature extraction module and a superpixel network. In the first step, a backbone network is employed to extract multi-scale features, including semantic information. In the second step, CNN-based superpixel segmentation is implemented to make the deep neural network aware of local superpixel regions. In this way, we ensure that fused features contain visual information from adjacent regions, provided by the superpixels. The resulting prediction process involves three steps:

(1) Extracting features with semantic information,
(2) Generating a superpixel adjacency map,
(3) Predicting a quality score.

The structure of the proposed network is illustrated in Fig. 2. In this process, an input image is fed into

both the backbone network (to extract semantic features) and a CNN-based superpixel segmentation network (to produce a superpixel probability map). Further consideration of feature aggregation and the refinement of crucial components then required designing a map generation network to produce superpixel adjacency maps after segmentation. Once generated, these maps were fused with semantic features and input to a prediction network used to determine the final quality score. These mixed features, comprised of both semantic and local adjacent information, are highly comprehensive and represent the image quality with high precision. Consequently, our approach is consistent with manual assessment, in which people concentrate on the semantic meaning and local details in an image simultaneously.

3.2 Semantic feature extraction

A pre-trained ResNet-50 was deployed as the backbone network used to extract more comprehensive semantic features. This network is aware of both semantic content and image quality. Inspired by earlier studies [8, 9], we applied a multi-scale feature extraction model in the backbone network. In this way, local content and distortions were extracted more completely. Multi-scale feature extraction also strengthens the effects of superpixel adjacency maps by conducting broader information fusion. Fig. 2 demonstrates the application of this multi-scale extraction model, in which local features are acquired from key points in the backbone network. In order to maintain principal components and a fast calculation speed, a 1×1 convolutional layer, an average pooling layer, and a full-connection layer were applied to refine multi-scale features. The introduction of this step allows the network to be represented as follows:

$$\boldsymbol{V}_{ms} = [\boldsymbol{L}_1, \boldsymbol{L}_2, \boldsymbol{L}_3, \boldsymbol{F}] = \varphi(\boldsymbol{x}), \tag{1}$$

where $\boldsymbol{x}$ represents input images, φ denotes the extraction model, $\boldsymbol{L}_i (i = 1, 2, 3)$ is the i-th local feature, $\boldsymbol{F}$ is a holistic feature, and $\boldsymbol{V}_{ms}$ indicates multi-scale features.

In addition, since this model is based on a consideration of semantic features and the HVS, input images should incorporate all of the original content. This requires that the pre-processing of input samples does not modify primary content or the quality of the images themselves. However, various augmented approaches have been applied previously, including cropping the images into smaller patches, resizing the images, or padding their edges. Each of these steps will affect the content and quality, thereby reducing adjacent information and making MOS/DMOS labels unsuitable for the processed data. As a result, since size can vary from picture to picture, our model must be capable of processing images of arbitrary size without affecting quality. Although convolutional layers can accept images of arbitrary size as input, FC layers only allow fixed vectors, making the entire network accessible solely by fixed-size samples. One common tactic is to use global average pooling (GAP) [40, 41] or global maximum pooling (GMP) [42] to regularize features. Although these techniques aim to establish a relationship between scalar quantities in features and channel quantities for features, they lose too much information in the process. For this reason, average pooling and maximum pooling steps were replaced with adaptive pooling, in which GAP and GMP can be considered special cases of adaptive pooling. In this way, the size of features can be adjusted to preserve the most useful information.

3.3 Superpixel adjacency map generation

3.3.1 Superpixel segmentation

SSVCNN represents superpixel segmentation as an N-category classification problem. The network is comprised of a five-layer CNN and the segmentation process can be defined as:

$$\boldsymbol{P} = S(\boldsymbol{x}), \tag{2}$$

where $\boldsymbol{P} \in \mathbb{R}_{+}^{H \times W \times N}$, $\sum_n \boldsymbol{P}_{h,w,n} = 1$ is a probabilistic representation map for superpixels, $S(\boldsymbol{x})$ denotes the whole network, and $\boldsymbol{x}$ represents an input image of size $H \times W$. We then removed the $\arg\max_n \boldsymbol{P}_{h,w,n}$ operation used to transform the probability map into visible superpixels for visual appreciation. In practical applications, these visible superpixels are not necessary and the probability map can be used directly. During implementation, we mostly employed the default parameters set by the author. Specifically, we set the maximum number $N(n)$ of superpixels to 100 to increase the processing speed.

Fig. 3 shows two examples of images processed using our superpixel segmentation model. These pictures were divided into 100 superpixels and processed by the argmax function to produce a visually crisper and clearer result. Adjacent pixels exhibiting the same type of features were grouped together in each image. These pixels were similar to each other in color and intensity and always belonged to a single semantic object. For instance, it is evident the area marked by the red frame in Fig. 3(a) includes semantic information describing a window. Also, the black shadow cast by the roof and the body of the window are clearly separated. Similarly, the area marked by the green frame in Fig. 3(b) represents the roof, while the region marked by the yellow frame is part of a tree. The textured tree and the

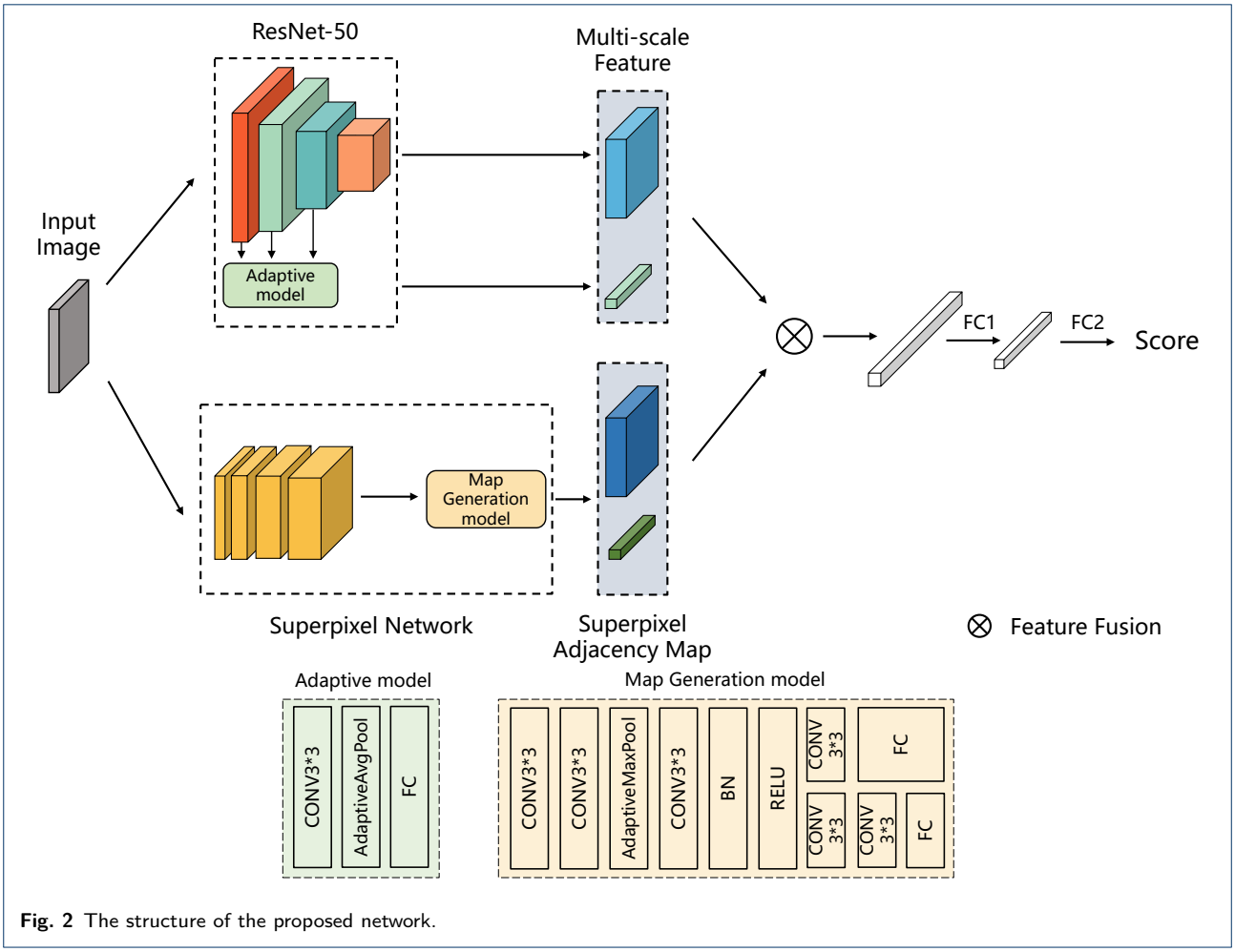


Fig. 2 The structure of the proposed network.

flat roof were both segmented successfully, with varying resistance to blur noise [39]. As such, the information present in those superpixels enabled further IQA processing for enhanced visual results. This step allowed feature extraction to correlate more closely with image quality. In this way, combining the results of superpixel segmentation with a CNN could also compensate for the inability of existing methods to exploit semantic image content.

3.3.2 Adjacency map generation

The proposed superpixel segmentation outputs a probabilistic representation map $\boldsymbol{P}$ of size $N \times H \times W$. A map generation model was also included to produce adjacency maps for future aggregation and redundancy elimination. The structure of this network is illustrated in Fig. 2, where several 3×3 convolution layers were used to acquire meaningful features and then apply adaptive maximum pooling, making the network suitable for arbitrary input sizes. This feature representation implies two different branches are required to fit both local and holistic features. Multi-scale features and the superpixel adjacency map are then integrated using direct multiplication and mixed features are fed to FC layers used to predict the final score.

3.4 Model training

3.4.1 Implementation details

The proposed technique was implemented in PyTorch 1.7.1 [43] and training and testing were conducted using 16G NVIDIA Tesla V100 GPUs. As described in Section 3.2, we tested images of varying sizes, based on their original dimensions in the database, while all training images were the same size. This is one of the requirements of mini batch-based training, which stabilizes loss and increases generalizability. The Adam [44] optimizer was utilized with a weight decay rate of $\lambda = 0.0005$, which helped to prevent overfitting. This step can be represented as follows:

$$\mathcal{L} = \mathcal{L}_0 + \lambda \sum_{\omega} \omega^2, \tag{3}$$

(a) Buildings (b) Church and Capitol

Fig. 3 An example of superpixel segmentation applied to two different images. In (a), the window and roof shadow are separated in the red frame, demonstrating that superpixels always belong to a single semantic object. In (b), the gathered pixels near the cupola share smoothness and flatness characteristics. Clustered pixels in the trees also share a similar texture. These images exhibit a unique collective perception to the human eye.

where ω represents all training patches and $\mathcal{L}$ and $\mathcal{L}_0$ denote the patch loss and original patch loss, respectively. The initial learning rate was set to 10^{-3} and dynamic adjustments were applied.

3.4.2 Loss function

Stochastic gradient descent and backward propagation are widely used in convolutional neural networks to calculate gradients and update learning parameters. Specifically, the loss function serves as an index for the entire network. In this study, we minimized $L1$ loss, which describes the absolute error between predicted and ground-truth scores in a training set. This term can be defined as follows:

$$\ell = \frac{1}{M}\sum_{i}^{M} |\Phi(\varphi(\boldsymbol{x}_i), S(\boldsymbol{x}_i)) - Q_i|, \tag{4}$$

where $\boldsymbol{x}_i$ and Q_i represent the i-th training patch and its corresponding ground truth score, respectively, Φ denotes the prediction model, and M is the number of input samples.

4 Results and Discussion

4.1 Databases and criteria

In the experimental stage, five databases were selected for training and testing, including KonIQ-10K [38], LIVE in the wild image quality challenge (LIVEC) [45], LIVE Facebook (FLIVE) [46], LIVE [47], and CSIQ [48]. Detailed descriptions of each are provided in Table 1.

The first three databases listed in italic text include authentic distorted images, which mimics the distribution of real images. These distortion types are more complex and composite than the other two synthetic databases, which makes image quality prediction more difficult. The KonIQ-10K database is comprised of $10,073$ distorted images of sizes 512×384 and 1024×768. These samples were selected from YFCC100m [49], a large public database. Their MOS values, ranging from $[0, 100]$, were provided as a ground truth. The LIVEC database contains $1,162$ images of size 500×500, acquired by cameras in real settings. Corresponding MOS values range from $[0, 100]$ and were collected using an online platform. The LIVE Facebook database includes $39,810$ images selected from several databases. It was released and implemented on the Amazon mechanical truck (AMT) crowd sourcing system, used to gather individual MOS values ranging from $[0, 100]$.

We also used synthetic data sets to test our methodology. The LIVE database contains 779 synthetically distorted images with provided DMOS values. CSIQ includes 866 images and uses DMOS values ranging from $[0, 1]$ as ground truth scores. Two common assessment criteria, namely Spearman's rank order correlation coefficient (SRCC) and Pearson's linear correlation coefficient (PLCC), were selected to evaluate the proposed technique. SRCC describes the monotonicity of an algorithm, while PLCC represents the accuracy of prediction. Both metrics range from -1 to 1, where a higher positive value indicates the result is more consistent with manual evaluation. The following splitting steps were involved in several tests, including individual database, individual distortion type, ablation, and sub-image size experiments. We randomly selected 80% of the images from the authentic distorted

Table 1 A comparison of experimental databases.

Database	Unique Content	Resolution	Distortions	No. of Distorted Images	Distortion Type	Quality Index
LIVEC	1,162	500 × 500 ∼ 640 × 960	-	1,162	In-The-Wild	MOS
KonIQ-10k	10,073	512 × 384 and 1,024 × 768	-	10,073	In-The-Wild	MOS
FLIVE	39,810	160 × 186 ∼ 1,200 × 660	-	39,810	In-The-Wild	MOS
LIVE	29	480 × 720 ∼ 768 × 512	5	779	Synthetic	DMOS
CSIQ	30	512 × 512	6	866	Synthetic	DMOS

image databases (KonIQ-10k and LIVEC) to form the training set, while 20% of the images comprised the test set. This 8 : 2 train-test ratio was also applied to reference images in the synthetic databases, which contributed to independence between the training and test sets.

4.2 Performance for individual databases

The proposed algorithm was trained and tested with the same database split using the technique described in Section 4.1. The splitting and train-test steps were repeated 10 times and the median results were calculated, thereby decreasing non-representative outcomes and avoiding dependence on a specific training set. We compared our model with 10 other algorithms, including:

- Full reference-based: PSNR and SSIM [50].
- Hand crafted feature-based BIQA: BRISQUE [51] and BMPRI [52].
- Deep learning-based BIQA: CNN-IQA [6], WaDIQaM-NR [53], SFA [39], DIQA [21], HFANet [41], and DBCNN [34].

It can be difficult to reproduce the results achieved with certain learning-based methods, due to the unavailability of code and related parameters. As such, some metrics were extracted from corresponding papers [34, 41, 54, 55] for comparative analysis. These results are provided in Table 2, where the best performance achieved for each database is shown in bold.

As seen in the table, our method outperformed other models when applied to authentic image databases. These results confirmed that visual system-based algorithms can process complex structures in authentic images. In addition, compared to synthetic databases, authentic data contain more varied semantic information and thus exhibit less repetition. This suggests these methods are more effective with increased data driving, thereby avoiding over-fitting issues. Deep Bilinear CNN (DBCNN) also outperformed other models applied to the FLIVE database, since it included two branches for addressing both authentic and synthetic images. One of these branches was pre-trained with the PASCAL VOC 2012 database [56], which was also partially included in the FLIVE dataset. As a result, certain advantages were provided to the DBCNN during training and testing with FLIVE. Although this technique was not intended for synthetic data, it still achieved a top three SRCC score for CSIQ and was above average for LIVE. This illustrates our method correctly assigns high scores to high quality images and low scores to low quality images. However, predicted scores were not as accurate as labeled ground-truth scores.

Table 2 Overall performance for individual databases.

SRCC	LIVEC	KonIQ-10k	FLIVE	LIVE	CSIQ
PSNR	-	-	-	0.876	0.806
SSIM	-	-	-	0.913	0.834
BRISQUE	0.601	0.715	0.320	0.942	0.698
BMPRI	0.487	0.658	0.274	0.931	0.908
CNN-IQA	0.627	0.685	0.306	0.955	0.683
WaDIQaM-NR	0.692	0.710	0.452	0.954	-
SFA	0.804	0.888	0.542	0.883	0.796
DIQA	0.703	-	-	**0.970**	0.844
HFANet	0.754	-	-	0.950	0.913
DBCNN	0.851	0.875	**0.554**	0.968	**0.946**
DSN-IQA	**0.854**	**0.913**	0.526	0.954	0.921
PLCC	**LIVEC**	**KonIQ-10k**	**FLIVE**	**LIVE**	**CSIQ**
PSNR	-	-	-	0.872	0.800
SSIM	-	-	-	0.945	0.861
BRISQUE	0.621	0.702	0.356	0.935	0.829
BMPRI	0.523	0.655	0.315	0.933	0.934
CNN-IQA	0.601	0.684	0.285	0.953	0.754
WaDIQaM-NR	0.730	0.738	0.433	0.963	-
SFA	0.821	0.897	0.626	0.895	0.818
DIQA	0.704	-	-	**0.972**	0.880
HFANet	0.766	-	-	0.963	0.918
DBCNN	0.869	0.884	**0.652**	0.971	**0.959**
DSN-IQA	**0.880**	**0.932**	0.623	0.954	0.910

4.3 Cross-database performance

A cross-database experiment was performed to test the generalizability of our technique. In this process, the algorithm was trained on one database and tested on another independent data set. Robust algorithms not only perform effectively on training data, but also on other samples. This cross-database test can be separated into two components, an authentic step and a synthetic step. It is worth noting the processes of acquiring images and ground truth distributions are unique to each database, due to the varied distortion types. As a result, evaluation metric scores were lower, especially when the type of data varied.

In the case of authentic distorted databases, we selected the most competitive algorithm (DBCNN) for

comparison purposes. We also tested each method using two additional authentic databases, which provided an evaluation of model generalizability. In the case of FLIVE, we followed the process described in [46] when establishing the test set. This approach was excluded from the training set because of its specific pre-processing requirements. Table 3 shows SRCC results for authentic samples. Values on the left were achieved using our technique and values on the right were calculated with DBCNN. The higher index is highlighted in bold text in each case. Our model outperformed a comparable algorithm in 3 out of 4 experiments, which illustrates our trained method is applicable to a more diverse set of images. In the case of training with two authentic databases, comparing the FLIVE test results suggests this approach assesses more perceptively and broadly if more training data are available. Thus, generalizability can be improved by ensuring sufficient training image quantities.

Table 3 Cross-database results comparing SRCC with DBCNN for authentic samples.

Train \ Test	KonIQ-10k	LIVEC	FLIVE_test
KonIQ-10k	**0.913**/0.875	**0.780**/0.755	**0.478**/0.470
LIVEC	0.724/**0.754**	**0.854**/0.851	**0.448**/0.405

One of the synthetic databases was used as a training set, while the other was used for testing. We also included LIVEC as an extended test set. A clearer comparison was made by selecting 6 outstanding methods, including BLIINDS-II [57], CNN-IQA, BIECON [58], PQR [59], WaDIQaM-NR, and TTL-IQA [55]. Corresponding SRCC results are shown in Table 4, where the top two generalizability indicators are listed in bold. The results of comparatively more difficult experiments, involving the LIVEC database for training and testing, indicated that our approach was superior in each case. This verifies that our method offers high generalizability, since the LIVEC data differ significantly from the two synthetic databases. These results also suggest that algorithms trained and tested with the same data type achieve more precise performance. Since these data distributions are similar, the included models can be more easily generalized.

4.4 Performance by distortion type

After experimenting with a holistic database, we analyzed the results of each distortion type separately, which involved measuring an ability to assess the quality of specific distortions. We only conducted these experiments using synthetic data, since distortion types are extremely complex and not tagged in authentic databases. These algorithms were trained using samples with various types of image distortions and tested by specific distortion type. Table 5 shows the results for the LIVE and CSIQ databases, with the top three performance values displayed in bold. It is evident from the table that our method is among the top two performing models, providing a significant advantage for addressing typical distortions. Specifically, our technique outperformed other models in the cases of Gaussian blur and fast fading. Superpixel segmentation also allows for a more adequate extraction of adjacency information, although the resulting content suffers from severe distortions. Thus, our approach has the ability to extract semantic features and accurately evaluate these two distortion types. It is worth noting our model did not perform as well as others for JPEG and JP2K images, each of which severely contaminates surrounding pixels and has a negative impact on superpixel segmentation. However, our approach is still viable, due to the inclusion of semantic and multi-scale features.

4.5 Ablation experiment

Four ablation experiments were conducted to determine the impact of our proposed methodology, in which ResNet50 was selected as a baseline. Resnet50-ft accepts images of fixed size (224×224) as input and was fine-tuned for testing on samples of the same size. A simple adaptive model was included in Resnet50-arbitrary, tested with arbitrarily sized images. Results in the third row were achieved using a multi-scale extraction model and are indicative of whether our approach improved the baseline network. Finally, Resnet50+Multi+SP was assessed to demonstrate that a combination of two sub-models is appropriate for the whole network. Results for all SRCC and PLCC experiments are provided in Table 6. The best results are listed in bold and the letters behind the database name indicate the SRCC (S) or PLCC (P) metric.

Table 6 demonstrates the function of each specific module. Results in the first row, taken from [60], suggest the inclusion of a semantic network outperforms many hand-crafted and CNN-based techniques. Unlike Resnet50-ft, Resnet50-arbitrary accepts intact images and retains semantic information during the assessment process. Corresponding results demonstrated a 2-3% improvement for LIVEC after allowing images of arbitrary size. However, the performance of this arbitrary model was limited for LIVE, due to special components and a requirement of equal image size in each mini batch. The results shown on rows 3 and 4 imply the sub-models are compatible, allowing the network to achieve better performance. In addition, multi-scale feature extraction, with its awareness of

Table 4 Synthetic cross-database SRCC results.

Train	Test	**DSN-IQA**	*BLIINDS-II*	*CNN-IQA*	*BIECON*	*PQR*	*WaDIQaM-NR*	*TTL-IQA*
LIVEC	LIVE	**0.603**	0.228	0.427	0.435	0.440	-	**0.676**
	CSIQ	**0.623**	0.305	0.239	0.412	0.538	-	**0.606**
LIVE	LIVEC	**0.520**	0.369	0.400	0.427	**0.547**	-	0.490
	CSIQ	**0.742**	0.601	0.723	0.719	0.717	0.704	**0.807**
CSIQ	LIVE	0.896	0.894	0.854	**0.922**	**0.930**	-	0.840
	LIVEC	**0.436**	0.276	0.300	0.317	**0.479**	-	0.298

Table 5 SRCC results for specific distortion types in the LIVE and CSIQ databases.

Database	LIVE						CSIQ						
Type	**JP2K**	**JPEG**	**WN**	**GB**	**FF**	**ALL**	**JP2K**	**JPEG**	**WN**	**GB**	**PN**	**CC**	**ALL**
BRISQUE	0.929	0.965	**0.982**	**0.964**	0.828	0.939	0.840	0.806	0.723	0.820	0.378	0.804	0.746
BLIINDS-II	0.930	0.950	0.947	0.915	0.871	0.912	0.850	0.846	0.702	0.880	0.812	0.336	0.780
BIECON	0.952	**0.974**	0.980	0.956	0.923	0.961	**0.954**	**0.942**	0.902	**0.946**	0.884	0.523	0.815
PQR	0.953	0.965	0.981	0.944	0.921	**0.965**	**0.955**	**0.934**	0.915	0.921	0.926	0.837	0.873
BMPRI	0.939	0.967	**0.986**	0.918	0.827	0.931	0.900	0.918	**0.928**	0.918	-	-	0.909
DIQA	**0.961**	**0.976**	**0.988**	**0.962**	0.912	**0.975**	0.927	0.931	0.835	0.870	0.893	0.718	0.884
DBCNN	**0.955**	**0.972**	0.980	0.935	**0.930**	**0.968**	0.953	**0.940**	**0.948**	**0.947**	**0.940**	**0.870**	**0.946**
HyperIQA	0.949	0.961	**0.982**	0.926	**0.934**	0.962	**0.960**	0.934	**0.927**	0.915	**0.931**	**0.874**	**0.923**
DSN-IQA	**0.956**	0.933	0.975	**0.967**	**0.955**	0.954	0.936	0.931	0.886	**0.944**	**0.929**	**0.869**	**0.921**

Table 6 The results of an ablation experiment.

Method	**LIVEC_S**	**LIVEC_P**	**LIVE_S**	**LIVE_P**
Resnet50-ft	0.819	0.849	0.954	0.950
Resnet50-arbitrary	0.848	0.869	0.946	0.939
Resnet50+Multi	0.850	0.872	0.952	0.946
Resnet50+Multi+SP	**0.855**	**0.880**	**0.954**	**0.954**

local and global features, was highly consistent with manual evaluation. Furthermore, superpixel segmentation permits the extraction of regional content and simulates the human visual system. Corresponding results illustrate the consolidation of multi-scale and superpixel segmentation as a feasible and effective way to extract accurate representations of image quality.

4.6 The effects of sub-image size

Data augmentation, including cropping an image to a specific size, can increase the number of samples available for training with small databases. However, structural distributions within an image can be regional and uneven as a result. Every random cropping in an image will create sub-images that vary in quality, which requires the selection of standardized sub-image dimensions that minimize distortions. We designed the following experiment to determine optimal sub-image sizes, in which the LIVEC and KonIQ-10k databases were selected for testing, due to their uniform image dimensions. The training set was generated by randomly cropping samples into varying sizes, ranging from 32×32 to full size [43]. The total number of epochs was determined by image dimensions, to ensure sufficient training time. All experiments involving the test set were conducted using original size samples. Corresponding LIVEC results are shown in Fig. 4, while KonIQ-10k values are provided in Fig. 5. These results indicate that as sample size grows larger, the monotonicity and accuracy simultaneously increase. Thus, ensuring consistency and conformity between training and test sets is necessary for a network to be fully trained and process complex images. This effect demonstrates that our method can correctly perceive quality and precisely evaluate images of arbitrary size, while preserving training content.

5 Conclusion

In this paper, we proposed a BIQA method based on the extraction of multi-scale semantic features and the use of superpixels. Our improved pooling technique avoided making changes to the quality of input images, caused by pre-processing. As a result, images of arbitrary scale can be accepted by the proposed model, while the original information and quality are preserved. In this process, multi-scale features containing semantic and quality information are acquired by a backbone network. These multi-scale features mimic information produced by the human visual system when people assess images, leading to credible prediction results. Furthermore, since adjacent pixels share several similar attributes and have a certain impact on perception, we implemented a superpixel model to extract this neighboring information. These structures also contain semantic details that complemented the

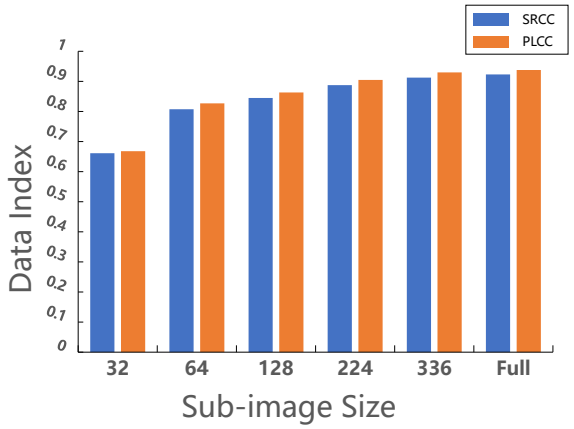


Fig. 4 A comparison histogram for varying sub-image sizes in LIVEC.

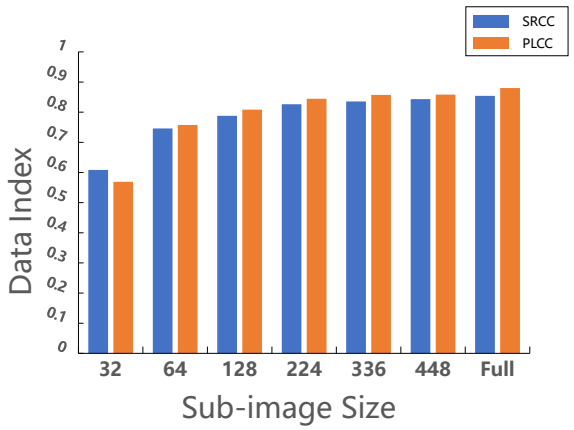


Fig. 5 A comparison histogram for varying sub-image sizes in KonIQ-10k.

multi-scale information. As a result, the fusion of these two elements allowed the prediction model to be highly consistent with human perception, especially for complex images. The proposed methodology can address complicated authentic images and accepts samples of arbitrary size as input during the testing period. In a future study, we will investigate new ways to accept images of any size during training and explore more efficient options for exploiting the information contained in superpixels.

Abbreviations

AMT: Amazon mechanical truck; BIQA: Blind image quality assessment; CC: Contrast noise; CNN: Convolutional neural network; CSIQ: Categorical subjective image quality; DMOS: Differential mean opinion score; DSN: Deep superpixel-based network; FC: Fully connected; FF: Fast fading; FLIVE: LIVE facebook; FR-IQA: Full-reference IQA; ft: Fine-tune; GAP: Global average pooling; GB: Gaussian blur; GMP: Global maximum pooling; HVS: Human visual system; IQA: Image quality assessment; JP2K: JPEG 2000; JPEG: Joint photographic experts group; LIVE: Laboratory for image and video engineering; LIVEC: LIVE in the wild image quality challenge; MOS: Mean opinion scores; MS: Multi-scale; NR-IQA: No-reference IQA; NSS: Natural scene statistics; PLCC: Pearson's linear correlation coefficient; PN: Pink noise; ResNet: Residual network; RR-IQA: Reduced-reference IQA; SFA: Semantic feature aggregation; SLIC: Simple linear iterative clustering; SPSIM: Superpixel-based similarity index; SRCC: Spearman's rank order correlation coefficient; SSVCNN: Superpixel segmentation via CNN; SVR: Support vector regression; WN: White noise.

Acknowledgments

The numerical calculations involved in this study were performed on the supercomputing system at the Supercomputing Center of Wuhan University.

Author Contributions

ZY implemented the core method, performed a statistical analysis, and drafted the manuscript. YX conducted validation experiments and drafted the manuscript. GY designed the methodology. All authors read and approved the final manuscript.

Funding

This study is partially supported by the National Natural Science Foundation of China (NSFC) (No. 61871298, 42071322) and the National Key Research and Development Program of China (No. $2018YFB0504501$).

Author Information

The authors are with.

Availability of Data and Materials

The Python source code for DSN-IQA can be downloaded at https://github.com/SN-F-QR/DSN-IQA for public use and evaluation. You may modify this program as you like and use it anywhere, but please refer to its original source.

Competing Interests

The authors declare they have no competing interests.

Author details

, , .

References

1. Yan, C., Shao, B., Zhao, H., Ning, R., Zhang, Y., Xu, F.: 3D room layout estimation from a single RGB image. IEEE Transactions on Multimedia **22**(11), 3014–3024 (2020)
2. Gu, K., Zhai, G., Yang, X., Zhang, W., Liu, M.: Subjective and objective quality assessment for images with contrast change. In: 2013 IEEE International Conference on Image Processing, pp. 383–387 (2013)
3. Wang, Z., Bovik, A.: Modern image quality assessment. 2006. Morgan & Claypool Publishers
4. Zhang, F., Xu, Y.: Image quality evaluation based on human visual perception. In: 2009 Chinese Control and Decision Conference, pp. 1487–1490 (2009). IEEE
5. Wang, Z., Bovik, A.C.: Reduced-and no-reference image quality assessment. IEEE Signal Processing Magazine **28**(6), 29–40 (2011)
6. Kang, L., Ye, P., Li, Y., Doermann, D.: Convolutional neural networks for no-reference image quality assessment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1733–1740 (2014)
7. Ma, K., Liu, W., Zhang, K., Duanmu, Z., Wang, Z., Zuo, W.: End-to-end blind image quality assessment using deep neural networks. IEEE Transactions on Image Processing **27**(3), 1202–1213 (2017)
8. Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., Zhang, Y.: Blindly assess image quality in the wild guided by a self-adaptive hyper network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3667–3676 (2020)
9. Lu, Y., Li, W., Ning, X., Dong, X., Zhang, L., Sun, L., Cheng, C.: Blind image quality assessment based on the multiscale and dual-domains features fusion. Concurrency and Computation: Practice and Experience, 6177 (2021)
10. Ren, Malik: Learning a classification model for segmentation. In: Proceedings Ninth IEEE International Conference on Computer Vision, pp. 10–171 (2003)
11. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. International journal of computer vision **59**(2), 167–181 (2004)

12. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**, 2274–2282 (2012)
13. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**, 888–905 (2000)
14. Moore, A.P., Prince, S.J.D., Warrell, J., Mohammed, U., Jones, G.: Superpixel lattices. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
15. Liu, M.-Y., Tuzel, O., Ramalingam, S., Chellappa, R.: Entropy rate superpixel segmentation. In: CVPR 2011, pp. 2097–2104 (2011). IEEE
16. Vincent, L., Soille, P.: Watersheds in digital spaces: an efficient algorithm based on immersion simulations. IEEE Transactions on Pattern Analysis and Machine Intelligence **13**, 583–598 (1991)
17. Comaniciu, D., Meer, P.: Mean shift: a robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**, 603–619 (2002)
18. Vedaldi, A., Soatto, S.: Quick shift and kernel methods for mode seeking. In: Computer Vision – ECCV 2008, pp. 705–718 (2008)
19. Levinshtein, A., Stere, A., Kutulakos, K.N., Fleet, D.J., Dickinson, S.J., Siddiqi, K.: Turbopixels: Fast superpixels using geometric flows. IEEE Transactions on Pattern Analysis and Machine Intelligence **31**, 2290–2297 (2009)
20. Bianco, S., Celona, L., Napoletano, P., Schettini, R.: On the use of deep learning for blind image quality assessment. Signal, Image and Video Processing **12**(2), 355–362 (2018)
21. Kim, J., Nguyen, A.-D., Lee, S.: Deep cnn-based blind image quality predictor. IEEE transactions on neural networks and learning systems **30**(1), 11–24 (2018)
22. Sun, W., Liao, Q., Xue, J.-H., Zhou, F.: SPSIM: A superpixel-based similarity index for full-reference image quality assessment. IEEE Transactions on Image Processing **27**(9), 4232–4244 (2018)
23. Shotton, J., Winn, J., Rother, C., Criminisi, A.: Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. International journal of computer vision **81**(1), 2–23 (2009)
24. Frackiewicz, M., Szolc, G., Palus, H.: An improved spsim index for image quality assessment. Symmetry **13**(3) (2021)
25. Nafchi, H.Z., Shahkolaei, A., Hedjam, R., Cheriet, M.: Mean deviation similarity index: Efficient and reliable full-reference image quality evaluator. Ieee Access **4**, 5579–5590 (2016)
26. Fang, Y., Zeng, Y., Jiang, W., Zhu, H., Yan, J.: Superpixel-based quality assessment of multi-exposure image fusion for both static and dynamic scenes. IEEE Transactions on Image Processing **30**, 2526–2537 (2021)
27. Jampani, V., Sun, D., Liu, M.-Y., Yang, M.-H., Kautz, J.: Superpixel sampling networks. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 352–368 (2018)
28. Suzuki, T.: Superpixel segmentation via convolutional neural networks with regularized information maximization. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2573–2577 (2020). IEEE
29. Yang, G., Ding, X., Huang, T., Cheng, K., Jin, W.: Explicit-implicit dual stream network for image quality assessment. EURASIP Journal on Image and Video Processing **2020**(1), 1–13 (2020)
30. Tang, H., Joshi, N., Kapoor, A.: Blind image quality assessment using semi-supervised rectifier networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
31. Drucker, H., Burges, C.J., Kaufman, L., Smola, A., Vapnik, V., *et al.*: Support vector regression machines. Advances in neural information processing systems **9**, 155–161 (1997)
32. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
33. Kang, L., Ye, P., Li, Y., Doermann, D.: Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks. In: 2015 IEEE International Conference on Image Processing (ICIP), pp. 2791–2795 (2015)
34. Zhang, W., Ma, K., Yan, J., Deng, D., Wang, Z.: Blind image quality assessment using a deep bilinear convolutional neural network. IEEE Transactions on Circuits and Systems for Video Technology **30**(1), 36–47 (2018)
35. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
36. Kim, J., Zeng, H., Ghadiyaram, D., Lee, S., Zhang, L., Bovik, A.C.: Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment. IEEE Signal processing magazine **34**(6), 130–141 (2017)
37. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
38. 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)
39. Li, D., Jiang, T., Lin, W., Jiang, M.: Which has better visual quality: The clear blue sky or a blurry animal? IEEE Transactions on Multimedia **21**(5), 1221–1234 (2018)
40. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
41. Wu, J., Yang, W., Li, L., Dong, W., Shi, G., Lin, W.: Blind image quality prediction with hierarchical feature aggregation. Information Sciences **552**, 167–182 (2021)
42. Su, Y., Korhonen, J.: Blind natural image quality prediction using convolutional neural networks and weighted spatial pooling. In: 2020 IEEE International Conference on Image Processing (ICIP), pp. 191–195 (2020)
43. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., *et al.*: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32**, 8026–8037 (2019)
44. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
45. Ghadiyaram, D., Bovik, A.C.: Massive online crowdsourced study of subjective and objective picture quality. IEEE Transactions on Image Processing **25**(1), 372–387 (2015)
46. Ying, Z., Niu, H., Gupta, P., Mahajan, D., Ghadiyaram, D., Bovik, A.: From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3575–3585 (2020)
47. 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**(11), 3440–3451 (2006)
48. Larson, E.C., Chandler, D.M.: Most apparent distortion: full-reference image quality assessment and the role of strategy. Journal of electronic imaging **19**(1), 011006 (2010)
49. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.-J.: Yfcc100m: The new data in multimedia research. Communications of the ACM **59**(2), 64–73 (2016)
50. 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**(4), 600–612 (2004)
51. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on image processing **21**(12), 4695–4708 (2012)
52. Min, X., Zhai, G., Gu, K., Liu, Y., Yang, X.: Blind image quality estimation via distortion aggravation. IEEE Transactions on Broadcasting **64**(2), 508–517 (2018)
53. 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**(1), 206–219 (2017)
54. Sun, W., Min, X., Zhai, G., Ma, S.: Blind quality assessment for in-the-wild images via hierarchical feature fusion and iterative mixed database training. arXiv preprint arXiv:2105.14550 (2021)
55. Yang, X., Li, F., Liu, H.: Ttl-iqa: Transitive transfer learning based no-reference image quality assessment. IEEE Transactions on Multimedia, 1–1 (2020)

56. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)
57. Saad, M.A., Bovik, A.C., Charrier, C.: Blind image quality assessment: A natural scene statistics approach in the dct domain. IEEE transactions on Image Processing **21**(8), 3339–3352 (2012)
58. Kim, J., Lee, S.: Fully deep blind image quality predictor. IEEE Journal of selected topics in signal processing **11**(1), 206–220 (2016)
59. Zeng, H., Zhang, L., Bovik, A.C.: A probabilistic quality representation approach to deep blind image quality prediction. arXiv preprint arXiv:1708.08190 (2017)
60. Kim, J., Zeng, H., Ghadiyaram, D., Lee, S., Zhang, L., Bovik, A.C.: Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment. IEEE Signal Processing Magazine **34**(6), 130–141 (2017)