

RESEARCH

Deep Superpixel-based Network for Blind Image Quality Assessment

Guangyi Yang^{*}, Yang Zhan and Yuxuan Wang

Abstract

The goal in a blind image quality assessment (BIQA) model is to simulate the process of evaluating images by human eyes and accurately assess the quality of the image. Although many approaches effectively identify degradation, they do not fully consider the semantic content in images resulting in distortion. In order to fill this gap, we propose a deep adaptive superpixel-based network, namely DSN-IQA, to assess the quality of image based on multi-scale and superpixel segmentation. The DSN-IQA can adaptively accept arbitrary scale images as input images, making the assessment process similar to human perception. The network uses two models to extract multi-scale semantic features and generate a superpixel adjacency map. These two elements are united together via feature fusion to accurately predict image quality. Experimental results on different benchmark databases demonstrate that our algorithm is highly competitive with other approaches when assessing challenging authentic image databases. Also, due to adaptive deep superpixel-based network, our model accurately assesses images with complicated distortion, 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 main carrier of visual information [1]. In many situations, the quality of an image is correlated to the coherence of the content since distortions have a significant negative impact on the readability of an image. Almost every stage in image acquisition, transmission, and storage could cause different degrees of distortion [2]. Consequently, image quality assessment (IQA) is necessary for monitoring the quality of images and thus assuring the reliability of image processing systems. As a result, research on IQA has received wide attention.

IQA methods can be divided into two types: subjective assessment and objective assessment, depending whether they need human eyes for classification [3]. Subjective assessment evaluates images using human eyes [4] and based on intuitive visual experience is the standard. It is the most accurate metric because the perceptual image quality is in the final instance, judged by the human visual system (HVS). However, it is time-consuming, expensive and the practical tasks are too laborious for routine implementation. So, objective

assessment is more practical and widely used because a machine can automatically predict the quality of an image using mathematical models. Objective assessment is usually divided into three categories on account of 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 application processes, a reference image is often not given, which impedes the application scope of FR-IQA and RR-IQA. BIQA therefore, has attracted growing interest among researchers [5].

Feature extraction algorithms make the BIQA model more widely applicable; however, the current algorithms explicitly designed for the model have weakness. Some BIQA models adopt low-level features and employ machine learning to assess quality: using a learning-based regression model trained by a set of features extracted from training images whose mean opinion scores (MOS) or Different MOS (DMOS) already gained via subjective experiments. This regression model is then used to predict the ground truth MOS. The BIQA models apply the principles of natural scene statistics (NSS) and can successfully represent the overall quality of the image. However, they are not effective for evaluating local distortions in an image. To resolve this issue feature extractors based on deep convolutional neural networks (CNN) [6, 7] have

^{*}Correspondence: ygy@whu.edu.cn

School of Electronic Information, Wuhan University, 430072 Wuhan, China
Full list of author information is available at the end of the article

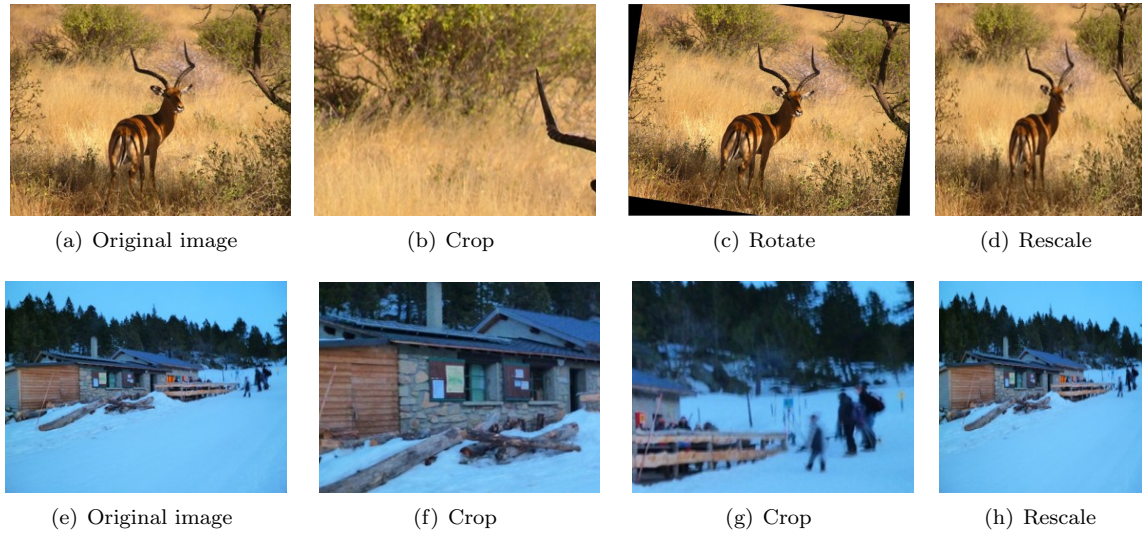


Figure 1 Pre-processing ways change the quality of image, including cropping, rotation, rescaling. (a), (e)Original images. (b), (f)-(g)Cropped images. Levels of the distortion and contents visually change due to the cropping. (c)Rotated image. The unreal black is introduced to pad the edges. (d), (h)Rescaled images, which contain conspicuous deformation.

been proposed and are widely used among researchers. They automatically capture deep features to represent degradation and they have been widely used in BIQA tasks. One of the major problems with CNN-based image quality assessment however, is that the attention of our eyes to the image is not evenly distributed across different regions. Ignoring the different weights of the different regions will add uncertainty to the quality assessment of an image, since HVSs are different from the prediction processes generated by a CNN model.

Moreover, another two inevitable problems that arise in deep learning solutions are the shortage of data and the fixed-scale input required by an end-to-end model. Many pre-processing methods are deployed to solve these problems. However, these arbitrary ways will decrease the consistency between the images and their ground-truth scores, as shown in Figure 1. The cropped images, like those in Figure 1(b) and (f), show content clipped from original images. In detail, Figure 1(b) includes the grass while the (a) contains both grass and an antelope. Also, the two cropped images from the same original image contain different levels and types of degradation, as shown in Figure 1(f) and (g). Figure 1(g) includes more severe blur than (f). Figure 1(c) indicates that the rotation of one image will introduce unreal colorations and exerts a negative influence on the image quality. Rescaling deforms the subject of images thus affecting both the semantic content and quality of images, Figure 1(d) and (h). As a result, the ground-truth quality is no longer suitable for the pre-processed images. Thus, using these pre-

processed images to train a model will definitely cause bias in prediction and a less objective model.

In order to design a BIQA method that is more consistent with human perception, we propose a deep BIQA network based on superpixels (DSN-IQA). Following the previous work [8, 9], we develop our CNN based network that extracts multi-scale semantic features, and fuse these features with the superpixel adjacency map obtained from superpixel segmentation model. Because when human eyes evaluate an image, they will pay attention to the semantic information and local details of the image at the same time and finally get the quality. Our network mimics this assessment process. In detail, we added superpixel segmentation to the method to help the network aware of local adjacency information. The images are segmented into superpixels, which are perceptually meaningful blocks, comprised of spatial neighboring pixels and generally given as a group of pixels. These pixels share similar local color and serve as low-dimensional representations of the images. Thus, they offer the detail information for further quality prediction. Consequently, our method simultaneously uses the local superpixels and multi-scale semantic features to ensure that the image evaluation process is more in line with the HVS. In addition, the method handles with the pre-processing problem by accepting arbitrary image sizes as input since humans assess the whole picture at the same time.

We tested the model on multiple databases. In these experiments the test image remained at the original

size so that the quality score of the test image accurately represents the truth quality. We conducted individual database, cross-database, individual distortion type, and ablation experiments. These test results show that the proposed adaptive model can handle complicated distortions and own high accuracy for predicting image quality due to the superpixel information extraction and sufficient multi-scale features.

In summary, our contributions are summarized as follows:

- To the best of our knowledge, the proposed model is the first to apply superpixel segmentation in BIQA to extract local features and multi-scale semantic features. Experiments demonstrate that these features are highly consistent with HVS.
- We analyzed the influence of image cropping training on complete image evaluation, and adjusted the pooling layers to design a model that can overcome the problems associated with the size of images.
- Our approach is deployed properly and every part fits each other. The results of our experiments indicate that the approach outperforms in predicting quality and handles images with complicated distortions.

The remaining parts of this paper are arranged as follows. In Section 2, we review the development of superpixel segmentation for IQA and CNN-based BIQA models and show what our model builds from. Section 3 details the construction of the proposed DSN-IQA model. Section 4 provides extensive experimental results and a comparative analysis of our proposed model. Section 5 summarizes our work and draws some conclusions.

2 Related Works

2.1 Superpixel segmentation for IQA

A superpixel, as defined by Ren *et al.* [10] in 2003, refers to irregular pixel blocks with certain visual significance composed of adjacent pixels with similar texture, color, brightness and other characteristics. Superpixel segmentation uses a small number of visually meaningful superpixels to represent groups of adjacent similar pixels, reducing the volume of data.

The widely used superpixel segmentation algorithms do not depend on CNNs relying instead on statistical models. These algorithms can be divided into two groups: the graph-based and the gradient-ascent-based algorithms. Graph-based algorithms are based on a data structure that contains vertices and weighted edges, and segments images by minimizing a cost function [11, 12]. Some of the typical graph-based methods are Normalized cuts [13], Graph cuts [14], and Entropy Rate Superpixel Segmentation algorithms [15].

Gradient-ascent-based algorithms are iterative and cluster pixels based on shifts between groups of pixels with similar values [12]. There are a number of such segmentation algorithms currently in use, these include the Watershed [16], Mean Shift [17], Quick Shift [18], Turbopixel [19], and Simple linear iterative clustering (SLIC) [12] algorithms. SLIC in particular, achieves superpixel segmentation based on clustering pixels by color and distance similarities. Most of superpixel segmentation algorithms are unsupervised and produce superpixels with uniform size and regular shapes. These superpixels have clear visual meaning and are widely used in computer vision preprocessing.

IQA, includes two main improvements for superpixel assessment accuracy. First, superpixels reduce the pixel redundancy and help the automated assessment process be perceptive. Many common IQA algorithms use square convolution kernels to ensure that a sufficient number features are extracted from images for quality prediction [6, 20, 21]. The 3×3 square kernel however, concentrates on a tiny zone at a time and thus loses visual meaning [22]. As a result, a square kernel does not exploit the connections between adjacent pixels while promoting information redundancy. In contrast, superpixel segmentation performs much like human vision. When humans observe and assess a picture, adjacent similar pixels are recognized and gathered to one local region [23]. Second, superpixels help IQA models to assess regionally. Superpixel-Based Similarity Index (SPSIM) [22], proposed by Sun *et al.*, illustrates that different types of regions respond to noise distinctively. Textured areas perform more resistance to Gaussian noise than flat areas. The situation is reversed for image blur. If the extracting network ignores these meaningful effects in regions, it will locally lose some common details and the predicting result will definitely deviate from human's subjective scores. All in all, superpixels are a vital tool for improving IQA algorithms.

Taking notice of superpixels, many superpixel-based methods emerged. SPSIM is a full-reference IQA model based on SLIC. In this method, reference images without distortion and distorted target images are segmented into visually meaningful regions, using superpixels. The mean values of the intensity and chrominance components are extracted within each superpixel and compared between a reference and target image to describe local similarity index precisely. Frackiewicz *et al.* improved SPSIM and developed an improved SPSIM index for IQA [24]. Their method revises SPSIM in two ways: new color space replaces the YUV space and exploits the calculation method from Mean Deviation Similarity Index (MDSI) index [25]. Fang *et al.* [26] use SLIC to distinguish between the

fused image and the exposed image. In this way, they compute the quality maps based on the Laplacian pyramid for large-changed and small-changed regions separately and take different regional strategies. All of these works bring regional solution to IQA by deploying superpixels.

Superpixel implementation improves the consistency between algorithms and HVS, but there is a problem that still must be solved. It is difficult to combine non-CNN superpixel segmentation algorithms with CNN networks and ensure a better performance. CNN creates dimensional feature tensors which have their visual meanings. Thus, these tensors can't combine simply with labeled superpixels created by non-CNN algorithms. At the same time, because CNN needs back propagation to train, these non-differentiable segmentation algorithms will block whole network and prevent training [27]. Nevertheless, directly using CNN to segment superpixel can overcome this contradiction. After comparing many other CNN superpixel segmentation methods, we opted for the Superpixel Segmentation Via CNN (SSVCNN) method proposed by Teppei [28]. This is an unsupervised superpixel segmentation method that optimizes a randomly-initialized CNN. SSVCNN is easier to integrate with existing image quality models, and the output is a clear and meaningful probabilistic map representing the belonging of pixels.

2.2 CNN-based BIQA models

Due to the insufficient computational power, the early CNN-based IQA models cannot extract and predict quality in one network. These methods initially only extract hierarchical features and subsequent operations calculate quality from the feature sets [29], which is different from an end-to-end model [7]. Tang *et al.* [30] proposed a non end-to-end model using a radical basis function to pre-train a deep belief network with unlabeled data and fine-tune it with labeled data. In addition, Bianco *et al.* [20] adopted CNN features pre-trained on the image classification task as input in order to create quality evaluator through support vector regression (SVR) [31], a model that regresses quality scores from feature sets. These researchers quantified the mean opinion scores (MOS) into five categories, fine-tuned the pre-trained features in the multi category classification settings, and fed the fine-tuned features to SVR. This method however, cannot be optimized end-to-end since making many manual parameter adjustments necessary.

Many end-to-end models started to emerge in BIQA however, with the development of computational power and deep CNN. Based on the BIQA model of

CNN [32], Kang *et al.* put forward the CNNIQA [6] algorithm, which takes the image patch as input and employs back propagation and other methods for training. Since the feature extraction and regression are integrated into the CNN, the depth of the neural network is deepened to improve the learning ability. Kang *et al.* proposed another algorithm CNNIQA++ [33], which increased the number of convolutional layers of CNNIQA so that it simultaneously estimates image quality and distortion type. Zhang *et al.* proposed the Deep Bilinear CNN (DBCNN) algorithm [34] based on VGG-16 [35]. VGG-16 initially was designed for image recognition but can be fine-tuned to assess image quality. They designed two branches of deep CNNs, specializing in synthetically and authentically assessing distortion scenarios separately, using bilinear pooling to fuse the network of the two branches. These methods however, still cannot accurately predict authentic databases that contain many images with complicated objects.

In recent years, semantic features based BIQA models have become a research hotspot because of their ability to perceive semantic information and more accurately predict authentic image databases. Kim *et al.* [36] found ResNet [37], a deep semantic CNN trained by classification databases, also assists in improving accuracy in IQA. In [38], researchers tested several deep CNNs and confirmed the advantages of semantic features in dealing with authentic IQA databases. In Semantic Feature Aggregation method (SFA) [39], Li *et al.* use statistics from ResNet-50 multi-patches features for quality prediction. They proposed that the content of images also has an impact on predicting the quality. They exemplified that people will score a clear blue-sky image as a high-quality one while the traditional prediction model mistakenly recognizes it as an image with blur noise. This phenomenon could be explained by the semantic losses during extracting features. Considering the contents varied from image to image, Su *et al.* proposed Hyper-IQA [8]. Hyper-IQA separates the IQA procedure into three stages including content understanding, perception rule learning and quality predicting. It designs a hyper network connection to mimic this mapping from image contents to the manner of perceiving quality. Thus, using semantic features in IQA will make an automated assessment process more like human assessment of the semantic contents in images when evaluating image quality.

The problem remaining is that all of the current semantic models lack attention to local visual meaning of images. As we illustrated in Section 2.1, superpixel fills this gap. Our proposed adaptive model takes a step further not only by using multi-scale semantic features, but also by introducing extraction of superpixel

information to imitate HVS. These two innovations improve the consistency of assessment process between our model and human eyes.

3 Proposed Method

3.1 Model framework

Our method contains two models: multi-scale feature extraction and a superpixel network. In first model, we deploy a backbone network to extract multi-scale features including semantic information. In second model, we implement CNN based superpixel segmentation to make the deep neural network aware of local superpixel regions. In this way, we ensure that the fused features will contain the visual information from adjacent regions as provided by superpixels. By introducing these two models, our prediction process contains three steps:

- (1) extract features with semantic information,
- (2) generate the superpixel adjacency map,
- (3) predict the quality score.

The structure of the network is shown in Figure 2. In detail, the input image is fed into one backbone network to extract semantic features and one CNN-based superpixel segmentation network to gain a superpixel probability map, respectively. For further consideration of features aggregation and refining the crucial part, we design a map generation network to gain a superpixel adjacency map after the segmentation. After generated, the superpixel adjacency map and the semantic features are combined together through features fusion. The mixed features are put into the prediction network to predict a final quality score. These mixed features, which are composed of semantic and local adjacent information, are highly comprehensive features and represent the image quality exactly. Consequently, our approach is consistent with the procedure of assessment, in which people concentrate on the semantic meaning and the local details of the image at the same time.

3.2 Semantic Features Extraction

We deploy pre-trained ResNet-50 as the backbone network to extract the semantic features and make the features more comprehensive. This network is aware of both the semantic contents and the quality of images. Inspired by earlier works [8, 9], we apply a multi-scale features extraction model in our backbone network. In this way, the local contents and distortions are extracted more completely. Also, the multi-scale feature extraction strengthens the effect of superpixel adjacency map by conducting wider information fusion. Figure 2 shows the details of our multi-scale extraction model. We acquire the local features from the key points of the backbone network. In order to keep

the principal component and fast speed of calculation, we apply a 1×1 convolution layer, an average pooling layer, and a full connection layer to refine the multi-scale features. With the introduction of multi-scale features, the network now can be defined as follows:

$$V_{ms} = [L_1, L_2, L_3, F] = \varphi(\mathbf{x}) \quad (1)$$

where \mathbf{x} represents the input images, φ means the extraction model, $L_i (i = 1, 2, 3)$ means the i -th local feature and F means the holistic feature, and V_{ms} means the multi-scale features.

In addition, because our model is based on the consideration of semantic features and HVS, the input images should incorporate all of the original content. This requires that the pre-processing of input images should not modify the main contents and quality of image themselves. However, some researchers apply various augmented methods, including cropping the images to small patches unduly, resizing the images or padding the edges of images. All of these processes will change the content and quality of images, reducing the adjacent information thus making the MOS/DMOS labels unsuitable for the processed images. As a result, under the condition that the size varies from picture to picture, our model must be available for arbitrary size without affecting the quality. Although the convolution layers can accept arbitrary size images as input, FC layers only allow the fixed vectors, making the whole network accessible merely for fixed-size images. One common tactic is to use global average pooling (GAP) [40, 41] or global maximum pooling (GMP) [42] to regularize the features. Although these approaches aim to establish the relation between the scalar amounts in features and channel amounts of the features, they lose too much information. For this reason, we displace the average pooling and maximum pooling with adaptive pooling, which GAP and GMP can be viewed as a special case of adaptive pooling. In this way, we adjust the size of features and save most of the useful information.

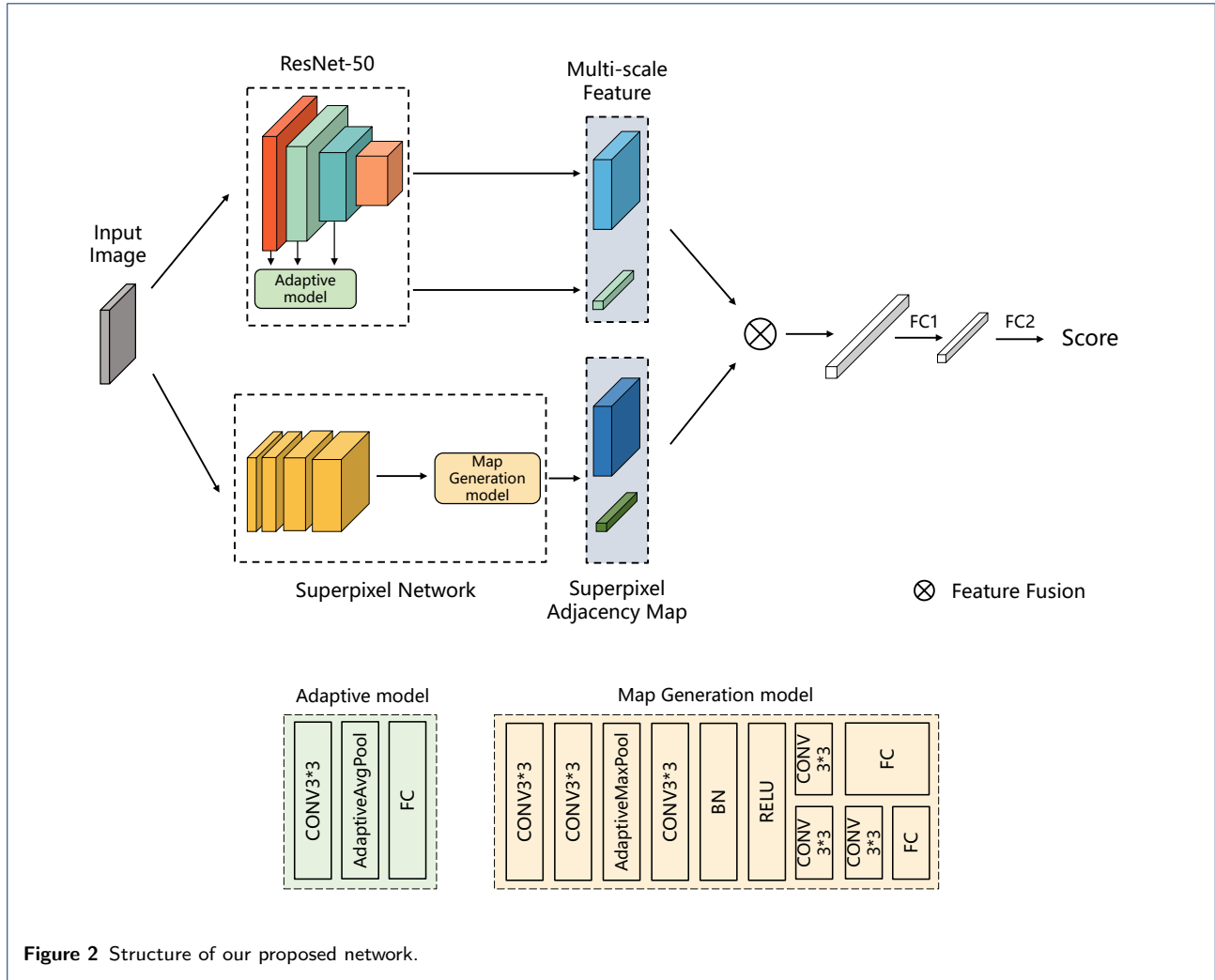
3.3 Superpixel Map Generation Net

3.3.1 Superpixel Segmentation

SSVCNN defines the superpixel segmentation as N class classification problem. Network is devised by a five-layer CNN and the procedure of segmentation can be simply defined as:

$$P = S(\mathbf{x}) \quad (2)$$

where $P \in \mathbb{R}_+^{H \times W \times N}$; $\sum_n P_{h,w,n} = 1$ is the probabilistic representation map of superpixels, $S(\mathbf{x})$ represents



the whole network, \mathbf{x} means the input image with size $H \times W$. We delete the operation $\arg \max_n P_{h,w,n}$, which is used to transform the supersixel probability map to visible supersixel for visual appreciation. In practical process, the visible supersixels are not necessary and thus we directly use the supersixel probability map. For implemented details, we mostly use the default parameters set by author. Particularly, we set the maximum number $N(n)$ of supersixel to 100 to accelerate the processing speed.

Figure 3 are two examples processed by our supersixel segmentation algorithm. The pictures in Figure 3 are divided into 100 supersixels and processed using the $\arg \max$ function to present a crisper, clearer visual result. Adjacent pixels with the same type of features are grouped together in the image. These pixels are similar to each other in color and intensity and always belong to one semantic object. For instance, through the Figure 3(a) we can clearly see that the area marked by red frame have semantic information implying a

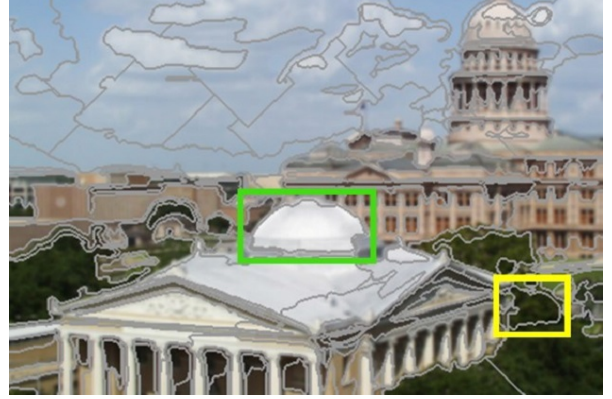
window. Also, the black shadow caused by roof and the body of window are separated clearly. Through Figure 3(b) we can see that the area marked by green frame represents the roof and the area marked by yellow frame is part of a tree. The textured tree and the flat roof are successfully segmented and they have different resistance to blur noise [39]. Thus, the information carried by those supersixels enables further IQA processing for enhanced visual results. This processing makes multi-scale semantic feature extraction conform more closely to quality of the picture for improved the adaptability and effectiveness of the algorithm. In this way, combining the result of supersixel segmentation with CNN can also make up for inability of existing algorithms to exploit the semantic content of images.

3.3.2 Supersixel Map Generation

After the supersixel segmentation, we have the probabilistic representation map P with the size $N \times H \times W$. For future aggregation and reducing the redundancy,



(a) Buildings



(b) Church and Capitol

Figure 3 Exmaple of superpixel segmentation. (a) Buildings after superpixel segmentation. In red frame the window and the shadow of roof are separated, which demonstrated that superpixel always belongs to one semantic object. (b) Church and capito after superpixel segmentation. The gathered pixels of cupola share same smooth and flat characteristics. The clumped pixels of tree share similar texture. Both of them have different perception for human eyes collectively.

we designed a map generation model to generate a superpixel adjacency map. The structure of the network is shown in Figure 2. We use several 3×3 convolution layers to gather meaningful features and then apply adaptive maximum pooling to make the network suitable for arbitrary sizes. According to the representation of features, two different branches are devised to fit the local features and holistic features, respectively. Owing the multi-scale features and superpixel adjacency map, we integrate them with direct multiplication and the mixed features are fed into FC layers to predict the final score.

3.4 Model Training

3.4.1 Implementation Detail

We deployed our model in PyTorch 1.7.1 [43] and conducted training and testing in NVIDIA Tesla V100 GPUs 16G. For the consideration we stated in Section 3.2, we tested different size images according to their original sizes in database, but all size of training images is the same in one database. This is the requirement of mini-batch based training, which stabilizes the loss and increases the generalization ability. We used the Adam [44] optimizer with the weight decay rate $\lambda = 0.0005$, contributing to avoiding over-fitting. It can be defined as following:

$$\mathcal{L} = \mathcal{L}_0 + \lambda \sum_{\omega} \omega^2 \quad (3)$$

where \mathcal{L} and \mathcal{L}_0 mean the patch loss and the original patch loss, and ω means all training patches. Also, we set initial learning rate as 10^{-3} and applied dynamic adjustment.

3.4.2 Loss function

For convolution neural networks, stochastic gradient descent and backward propagation are widely used to calculate the gradient and update the parameters. Loss function acts as an index of the entire network. In our work, we minimize L1 loss, which describes the absolute error between predicted score and the ground-truth score over training set. It can be defined as follows:

$$\ell = \frac{1}{M} \sum_i^M |\Phi(\varphi(\mathbf{x}_i), S(\mathbf{x}_i)) - Q_i| \quad (4)$$

where \mathbf{x}_i and Q_i represent the i -th training patch and its ground truth score, Φ represents the prediction model, and M represents the number of input samples.

4 Results and discussion

In experiment stage, we selected 5 databases for training and testing. They are KonIQ-10K [38], LIVE In the Wild Image Quality Challenge (LIVEC) [45], LIVE Facebook (FLIVE) [46], LIVE [47], and CSIQ [48] database. All details are described below and shown in Table 1.

The first three lines are the authentic distorted image databases, which aim to achieve the distribution of real distorted images. Compared to the rest two synthetic databases, their distortion types are more complex and composite, which always become a harder task for image quality prediction model. KonIQ-10K database is composed of 10,073 distorted images in size 512×384 .

Table 1 Comparison of basic database conditions

Database	Unique contents	Resolution	Distortions	No. of distorted images	Distortion type	Quality Index
LIVEC	1,162	$500 \times 500 \sim 640 \times 960$	-	1,162	in-the-wild	MOS
KonIQ-10k	10,073	512×384 and $1,024 \times 768$	-	10,073	in-the-wild	MOS
FLIVE	39,810	$160 \times 186 \sim 1,200 \times 660$	-	39,810	in-the-wild	MOS
LIVE	29	$480 \times 720 \sim 768 \times 512$	5	779	synthetic	DMOS
CSIQ	30	512×512	6	866	synthetic	DMOS

and 1024×768 . They are chosen from a huge public database YFCC100m [49] and their MOS ranged in [0, 100] is provided as the ground truth. LIVEC database contains 1,162 images in size 500×500 created by cameras in the real world. MOS ranged in [0,100] is used and collected through online platform. The LIVE Facebook database containing 39,810 images selected from several databases, was released and implemented on the Amazon Mechanical Turk (AMT) crowd sourcing system to gather each image MOS, ranged in the [0, 100].

In addition, we also used synthetic databases to test our work. The LIVE database has 779 synthetically distorted images and provides DMOS. CSIQ database provides 866 images and takes DMOS ranged in [0,1] as ground truth score

Two commonly used criteria, namely Spearman's rank order correlation coefficient (SRCC) and Pearson's linear correlation coefficient (PLCC) were selected to evaluate the model. SRCC represents the monotonicity of the algorithm and PLCC represents the accuracy of prediction. Both range from -1 to 1 and a higher value indicates a model more consistent with human eyes.

We conducted the following splitting approaches in several experiments, including individual database, individual distortion type, ablation, and sub-image size experiment. From the authentic distorted image databases, KonIQ-10k, and LIVEC we randomly selected 80% images as training set and 20% images as testing set. In synthetic image databases, we also used the 8:2 train-test ratio to split the reference images, contributing to making the contents of image independent in the training and testing set.

4.1 Performance on individual database

The proposed model was trained and tested on the same database split by method detailed in Section 4. The splitting and train-test procedures were repeated 10 times and the median results are provided. This decreases chance results and avoid dependence of the model on the specific training set. We compared our model with 10 other approaches, including:

- Full reference based approaches: PSNR, SSIM [50].
- Hand-crafted features based BIQA approaches: BRISQUE [51], BMPRI [52].

- Deep learning based BIQA approaches: CNN-IQA [6], WaDIQaM-NR [53], SFA [39], DIQA [21], HFANet [41], DBCNN [34].

Particularly, for some deep learning-based methods, it is difficult to reproduce the result due to the unavailability of code and related parameters so their results were extracted from the corresponding papers [34, 41, 54, 55] for comparative analysis. The comparative results are shown in Table 2 and the best performances in each database are shown in bold.

Table 2 Overall Performance on individual database

SRCC	LIVEC	KonIQ10k	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	KonIQ10k	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

On authentic image databases, our model outperformed the other tested algorithms. These results meet our expectation that our visual system-based model can handle the complex situation in authentic images. Moreover, compared to synthetic databases, authentic databases contain more different semantic contents and thus have less repetition. This suggests that with more data driving, the model is effective and avoids over-fitting. Among all of the other approaches, Deep Bilinear CNN (DBCNN) also outperformed in FLIVE database. Because it contains two branches for dealing with authentic and synthetic images. One of

the branches is pre-trained by PASCAL VOC 2012 database [56] which is also used by FLIVE database partially. As a result, it brings some advantages to DBCNN when it trains and tests in FLIVE.

For synthetic databases, although our model is not intended for them originally, it still gained the top three SRCC values on CSIQ database and above the average on LIVE. This illustrates that our model gives the high score to high quality images and gives low score to bad quality images correctly. But the predicted scores are not as accurate as the ground-truth scores in label.

4.2 Performance on cross-database

Cross-database experiment was designed to test the generalization ability of our approach. Cross-database experiment trains the model on one database and tests it on another independent database. A robust model can not only perform effectively on training database, but also on other databases. The whole cross-database test is separated into two parts, an authentic part and a synthetic part. We should mention that because of the distorted types, processes of taking the image and the distribution of ground truth are different in each database, the evaluation metrics will certainly degrade, especially when the type of database is different.

In the authentic distorted databases, we selected the most competing approach DBCNN to compare. We test the models with other two authentic databases, which can bring a sufficient view of the generalization ability. For FLIVE database, we follow the process in [46] to collect the testing set. It is excluded from the training set because of its specific requirement to the pre-processing. Table 3 shows the SRCC result in authentic part. The left values are provided

Table 3 Result of SRCC compared with DBCNN on authentic cross-database

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

by our method and the right values are calculated by DBCNN. The better method's index between them is highlighted in bold text. Our model outperformed 3 out of 4 experiments, which illustrates that our trained model is applicable for more different images. For our model trained by two authentic databases, comparing the test result in FLIVE suggests that the model assesses more perceptively and widely if there are more data for training. Thus, we can ensure the sufficient images for training to improve the generalization ability.

One synthetic database, became a training set and the other was used for testing. We also added LIVEC as the extended testing set. For clearer comparison, we pick 6 outstanding approaches, including BLINDS-II [57], CNN-IQA, BIECON [58], PQR [59], WaDIQaM-NR, and TTL-IQA [55]. The SRCC results are shown in Table 4. The top two indicators of generalization ability are shown in bold.

Our model achieved the top performance of generalization ability. When we focus on the are comparatively harder experiments which contain LIVEC database as training set or test set, our model stands out every time. This verifies that our model owns high generalization ability since the type of LIVEC database is different from two synthetic databases. Also, the results reflect that the models trained and tested in same type databases gain more precise performance. As the data distributions are similar, the models can get considerable ability for generalization.

4.3 Performance on each distortion type

After testing on holistic database, we analyze the results of each distortion type separately. This experiment measures the ability for assessing the quality of specific distortion. We only conducted the experiments on the synthetic databases because the distortion types are extremely complex in authentic databases and also not tagged. The models were trained in database with all types of distorted images and tested by specific distorted type. Table 5 shows the result from LIVE and CSIQ database and the top three performance values are in bold. From these two tables, we can observe that our model is among the top two performing models, showing a significant advantage in dealing with usual distortions. Especially, our model outperforms in Gaussian Blur and Fast Fading. Due to the superpixel segmentation, we can extract their adjacency information more adequately although their contents have suffered from a severe distortion. Thus, our model has ability to gain semantic features and evaluate more accurately for these two distortions. However, our model did not perform as well as others on JPEG and JP2K images. Both contaminate the surrounding pixels severely thus have negative impacts on superpixel segmentation. Under this situation, thanks to the semantic and multi-scales features, our approach still has appreciable results.

4.4 Ablation experiment

In order to verify the improvement of our approach, we set 4 ablation experiments. First, we select the ResNet50 as basic experiment to evaluate the ability of baseline. Resnet50-ft accepts fixed size 224×224 as input and is fine-tuned to be tested on same size images.

Table 4 SRCC Results on synthetic cross-database

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 of specific distortion types on LIVE database and CSIQ database

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

For the Resnet50-arbitrary, the simple adaptive model is added and thus tested on arbitrary size images. Secondly, we only deploy the multi-scale extraction model to the backbone. The result can suggest whether it can improve the ability of baseline network. At last, our model is tested to demonstrate that a combination of two sub models is appropriate for the whole network. All SRCC and PLCC experiments are shown in Table 6. The best results are emphasized by bold and the letters behind the database name indicate the SRCC (S) and PLCC (P) metric.

Table 6 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

Table 6 indicates the function of each specific model. The first-row result, taken from [60], shows that the mere usage of a semantic network can outperform many hand-crafted and CNN based approaches. Compared to the first row, the method on the second row accepts the intact images and thus saves more semantic information for the assessment process. The result on second method shows 2-3% improvement on LIVEC after only allowing arbitrary image sizes. Due to the special component of image size in LIVE and the requirement of same image size in one mini-batch, the arbitrary model cannot perform fully in LIVE. The results found in row 3 and 4, definitely indicate that

both of the sub models fit each other and help the network achieve better performance. Multi-scale features extraction, with its awareness of local and global features, and fits human assessment precisely. The super-pixel segment model permits the extraction of regional contents and simulates the human visual system. The result in this part illustrates that the consolidation of multi-scale and superpixel segmentation is a feasible and effective way to extract the accurate features of image quality.

4.5 Effect of sub-image size

Considering the limited image amounts and various image sizes in some databases, sometimes cropping the image to specific size for data augmented and training is necessary. However, the quality distribution within an image is regional and uneven. Every random cropping for the same image will create sub-images that vary in quality. As a result, there must be a sub-image size that minimizes distortion.

We designed the following experiment to determine the optimal sub-image size. For this experiment we chose the LIVEC and KonIQ-10k databases because all image sizes are same. Thus, we can control the variate. For training set, we randomly cropped [43] the images into various sizes, ranging from 32×32 to full size. The total epoch was adjusted by the size to ensure the sufficient training. For the testing set, all the experiments were evaluated using original size images. The results are shown in Figure 4 displaying the results from LIVEC and Figure 5, depicting those

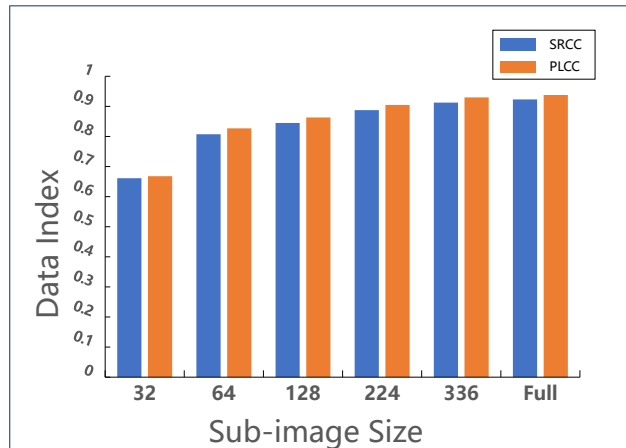


Figure 4 Comparison histogram of sub-image size results in LIVEC

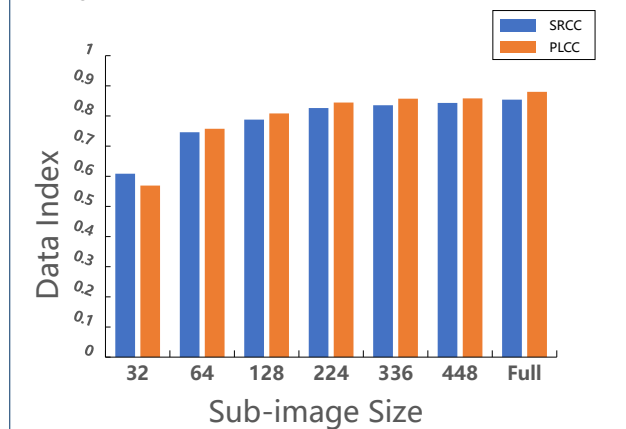


Figure 5 Comparison histogram of sub-image size results in KonIQ-10k

from the KonIQ-10k database. It shows that when the size grows larger, the monotonicity and accuracy of models simultaneously increased. This phenomenon demonstrates that the model can perceive the quality correctly and evaluate arbitrary size images precisely with more training contents being saved. Thus, ensuring the consistency and conformity between training set and testing set is necessary for the network to be fully trained and deal with complicated images.

5 Conclusion

In this paper, we propose an BIQA model based on multi-scale semantic features and superpixels. For the input images, our improved pooling approach avoids changes in the quality of input images caused by pre-processing methods and thus arbitrary scale images can be accepted by the proposed model, and the original information and quality are preserved. In our proposed model, multi-scale features containing semantic

and quality information are gathered by the backbone network. These multi-scale features mimic the information produced by human eyes when people assess one image and thus can predict a credible result. Furthermore, since the adjacent pixels share many similar attributes and have a certain impact on perception, we implement a superpixel model to extract the neighboring information. This information also contains semantic information and complements multi-scale features. As a result, with the fusion of these two elements, the prediction model is highly consistent with human perception and solves complicated images.

The proposed model tackles complicated authentic images and accepts arbitrary size images as input during testing period. How to make the training period accept any size images and explore a more efficient way to exploit superpixels could be the starting point for further research.

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: Full 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; SSVCCNN: Superpixel Segmentation Via CNN; SVR: Support vector regression; WN: White noise;

Acknowledgements

The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

Authors' contributions

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

Funding

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

Authors' information

The authors are with the School of Electronic Information, Wuhan University, Wuhan 430072, China.

Availability of data and materials

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

Competing interests

The authors declare that they have no competing interests.

Author details

School of Electronic Information, Wuhan University, 430072 Wuhan, China.

References

- 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)
- 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)
- Wang, Z., Bovik, A.: Modern image quality assessment. 2006. Morgan & Claypool Publishers
- Zhang, F., Xu, Y.: Image quality evaluation based on human visual perception. In: 2009 Chinese Control and Decision Conference, pp. 1487–1490 (2009). IEEE
- Wang, Z., Bovik, A.C.: Reduced-and no-reference image quality assessment. *IEEE Signal Processing Magazine* **28**(6), 29–40 (2011)
- 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)
- 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)
- 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)
- 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)
- Ren, Malik: Learning a classification model for segmentation. In: Proceedings Ninth IEEE International Conference on Computer Vision, pp. 10–171 (2003)
- Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. *International journal of computer vision* **59**(2), 167–181 (2004)
- 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)
- Shi, J., Malik, J.: Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**, 888–905 (2000)
- 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)
- Liu, M.-Y., Tuzel, O., Ramalingam, S., Chellappa, R.: Entropy rate superpixel segmentation. In: CVPR 2011, pp. 2097–2104 (2011). IEEE
- 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)
- 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)
- Vedaldi, A., Soatto, S.: Quick shift and kernel methods for mode seeking. In: Computer Vision – ECCV 2008, pp. 705–718 (2008)
- Levinstein, 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)
- 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)
- 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)
- 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)
- 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)
- Frackiewicz, M., Szolc, G., Palus, H.: An improved spsim index for image quality assessment. *Symmetry* **13**(3) (2021)
- 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)
- 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)
- 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)
- 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
- 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)
- 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)
- 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)
- 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)
- 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)
- 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)
- Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014)
- 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)
- 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)
- 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)
- 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)
- 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)
- 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)
- 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)
- 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)
- 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)