Applied Computing and Intelligence

https://www.aimspress.com/journal/aci

Applied Computing and Intelligence
4(2): 107–124.
DOI: 10.3934/aci.2024007
Received: 02 August 2024
Revised: 07 October 2024
Accepted: 09 October 2024
Published: 21 October 2024

Research article

Iterative transfer learning with large unlabeled datasets for no-reference image quality assessment

Sheyda Ghanbaralizadeh Bahnemiri*, Mykola Pnomarenko and Karen Eguiazarian

Faculty of Information Technology and Communication Sciences, Tampere University, Finland

* **Correspondence:** Email: sheyda.bahnemiri@gmail.com.

Academic Editor: Chih-Cheng Hung

Abstract: No-reference image quality assessment is crucial for evaluating perceptual quality across diverse image-processing applications. Given the challenge of accruing mean opinion scores for images, utilizing data augmentation and transfer learning is vital for training predictive networks. This paper presents a new iterative transfer learning technique, which helps to transfer knowledge between heterogeneous network architectures, and overcomes the problem of overlearning when training on small datasets. The proposed method used a large amount of unlabeled data during training, improving its ability to handle different image quality conditions. We also presented a two-branch convolutional neural network architecture, which merges multi-scale and multi-level attributes efficiently. This architecture emphasizes both local detail extraction and high-level comprehension, and the result was fast execution time and minimal memory overhead. Empirical results showed that applying iterative transfer learning to train a two-branch convolutional neural network achieved superior real-time performance and at the same time exhibited good performance in spearman's rank order correlation coefficient. Furthermore, the model manifested robustness for the noisy mean opinion score, which is prevalent in available datasets, and during data augmentation processes.

Keywords: no-reference image visual quality assessment; deep convolutional neural networks; transfer learning

1. Introduction

A growing demand for high-quality images and videos has prompted an increasing need for image quality assessment (IQA) tools [1]. The assessment of image quality is crucial for Internet-based platforms like YouTube and Facebook, as well as for handheld devices such as smartphones, as it significantly influences user perceptions [2]. Given that reference images may not be available in these applications, no-reference image quality assessment (NR-IQA) techniques are widely employed

to estimate image quality. The basics of NR-IQA design is extracting and analyzing image features to detect distortions present in the images.

Images are often subjected to various types of artifacts. For instance, blocking artifacts resulting from JPEG compression and ringing artifacts caused by JPEG2K were explored in [3–5] and [6, 7], respectively. Papers [8–10] investigated image blurriness, while [11, 12] examined artifacts related to super-resolution techniques. Additionally, noise has been the subject of study for years, for instance in [13], a technique based on the discrete cosine transform (DCT) was proposed to reduce speckle noise, and similarly in [14], DCT is used for estimating non-stationary noise. Authors in [15] applied principal component analysis (PCA) to estimate noise levels from image patches. Lately, deep learning methods have been applied for estimating non-stationary noise [16, 17].

However, in real-life applications, one usually encounters images that are corrupted by a combination of multiple distortions. Therefore, there is a need to develop a comprehensive NR-IQA metric capable of detecting a wide range of possible distortions. Distortions normally alter statistical properties of images both in spatial and frequency domains [18]. These statistical properties are known as natural scene statistics (NSS) and there are various methods that can extract such properties in the spatial domain [19–21], and transform-based domain, e.g., [22, 23] which used a DCT transform or [24, 25] which used a wavelet and curvelet. Additionally, techniques presented in [26–29] combine both spatial and frequency characteristics to provide more robust and reliable results. When an image is distorted, the distribution of its NSS values shifts, reflecting the degree of degradation. This shift allows us to measure the deviation of a distorted image from a normal pristine. For NR-IQA, this measurement is performed independently for each type of distortion (e.g., noise, blur, or compression artifacts) to capture the unique impact of each distortion on the image's quality. Finally, these individual measurements are combined in a pipeline using regression techniques, e.g., support vector regression (SVR) [30], and the final score value is obtained. NSS thus serves as a powerful tool to extract meaningful statistical differences between distorted and pristine images in the NR-IQA framework.

The emergence of deep learning and convolutional neural networks (CNN) has brought about a significant transformation in the field of image processing, including NR-IQA. The first CNN to design NR-IQA was a relatively shallow network utilized on synthetic datasets [31]. CNNs have been mostly used for end-to-end optimization [32, 33] and training NR-IQA models requires extensive datasets labeled with mean opinion scores (MOS) as the target value. Some well-known datasets for this purpose are FLIVE [34], KonIQ-10k [35], NRTID [36], HTID [37], SPAQ [38], and LiveInWild [39]. However, many of these datasets are insufficient in size for training deep models, which can lead to overfitting as deep networks tend to memorize key features from tens of thousands of images. One of the approaches to address this issue is employing transfer learning [40, 41].

The idea behind transfer learning is that knowledge acquired from solving one problem can be used for solving a different but related problem. Instead of training a model from scratch on a new task, transfer learning allows utilizing the learned features, representations, or parameters from a pre-existing base model. In the case of NR-IQA design, pre-trained weights can effectively avoid overfitting and compensate for the insufficiency of datasets. In most cases, the desired task is different from the initial task which the base model was trained for. Therefore, the last layers of the base model should be modified according to the new task. It means that the number of output neurons and loss function in the last layer should be changed. Authors in [32] used pre-trained models with 27 layers, added 5 extra layers to modify the architecture according to the desired task and fine-tuned the model

on NR-IQA datasets.

Generally, there are popular architectures that are frequently used for transfer learning. VGG16 [42], Inception-v2 [43], and MobileNet [44] are examples of such architectures, and authors of [45] and [35] used them to develop NR-IQA models. In [45], a new loss function, *Earth mover's distance* (EMD), was used as the new regression loss. Likewise, in [40], authors used Xception [46] as the base model to predict the opinion score distribution instead of MOS. In [47], pre-trained CNNs were used as feature extractors where both the image and its sub-regions were fed to the pre-trained model to extract features. In [48], a Siamese network [49] was trained to rank images based on the image quality, and then, by fine-tuning, the knowledge gained by the Siamese network was transferred to a CNN that estimated a final image quality. However, transfer learning has this limitation that is only feasible between similar network architectures, because it is using the pre-trained weights of a strcuture so it is bound to use the same structure.

The primary motivation for employing CNN-based models is their efficient handling of local features and spatial dependencies. This is especially critical in the context of NR-IQA, where the extracted features should closely align with the human visual system (HVS), considering both local and global characteristics. Both of these characteristics which can be also be called low-level and high-level features, are important in estimating the image quality because the HVS analyzes the composition of the scene and colors, which are high-level features, and blur or noisy artifacts that are found in low-level features. One of the important parts of an image is the salient regions. The salient regions generally contain rich semantic information and thus have a great impact on the overall quality. HVS also focuses on salient regions, therefore, utilizing techniques to separate such regions for better analysis have been used extensively because they can model the focusing characteristic of HVS and assign larger weights to the focused parts. Authors of [50] and [51] used CNN architecture in combination with the visual attention modules or saliency maps, which helped them to focus on specific regions of the image. Another instance of employing attention mechanisms can be found in [52], where feature-product networks were crafted by considering principles inspired by biological vision. One of the state-of-the-art methods that utilized saliency for image quality, specifically, on surveillance face images, was presented in [53]. The model called Hyper-IQA proposed in [54] also fits in the category of semantic-based or content-aware techniques. It consists of three main steps: 1) content understanding, 2) perception rule learning, and 3) quality predicting.

One of the recent methods, proposed in [55], presented a hybrid method based on a CNN and transformer. An off-the-shelf pre-trained CNN was used as a feature extractor and its output was used as an input to the transformer. The authors combined the self-attention mechanism of transformers and a locality bias in the extracted features of the CNN, which led to a robust analysis of the image because the correlation between different local features was measured and considered during quality assessment.

In this paper, we addressed the limitation of transfer learning and incorporaing HVS in NR-IQA design by introducing a two-branch CNN-based network that accepts both high- and low-frequency image patches as inputs. In our study, we propose an improved version of the transfer learning proposed in [56] and created a dataset by collecting thousands of images from Google Open Images (GOI) [57]. This extensive dataset covers a broad spectrum of concepts and categories, including common objects, animals, people, and nature.

Many CNN-based NR-IQA approaches frequently face high computational costs and varying

execution times due to variations in input image sizes and, at the same time, pre-trained models often require image resizing, which, when performed non-proportionally, has the potential to jeopardize image quality and the image aspect ratio. Moreover, many of the pre-trained network parameters are large, rendering them unsuitable for hardware implementation. Our models possess a relatively small number of parameters. Additionally, our approach to image patch selection, which involves the random sampling of local patches and global patches that incorporate low-level and high-level features, respectively, both TBCNet parameters and the image patch selection procedure aid in accelerating the model's processing speed, eliminating the need for image resizing.

Our contributions in this paper can be explained in two main parts: first proposing a two-branch convolutional neural network (TBCNet) , that uses both low-level and high-level features of images for estimating image quality. The network has low memory requirements and a small constant computational time for images of different resolution. This feature makes our model particularly well-suited for real-time applications and resource-constrained environments.

Second, proposing iterative transfer learning (ITL), which is an iterative version of the new transfer learning method introduced in [56], we propose an iterative training scheme and ensure the robustness of the method by training on noisy MOS.

The rest of this paper is organized as follows: Section 2 details the proposed iterative transfer learning method. Section 3 describes the TBCNet model, used for verifying the ITL approach. Section 4 explains the application of transfer learning on the KonCept512 NR-IQA model and TBCNet. Section 5 presents numerical analysis, comparisons, efficiency and generalization of our method, and analysis of the denoising capability of our proposed training scheme. Conclusions and future work are explained in Section 6.

2. Iterative transfer learning

In this research, we enhance and expand upon the transfer learning technique that was initially presented in our prior work [56]. The framework of the proposed method is depicted in Figure 1. Additionally, a comprehensive description of the (ITL) process is provided in Algorithm 1. The principle of this method which is proposed in [56] involves transferring knowledge from a well-developed, pre-trained NR-IQA model, referred to as metric A, to another NR-IQA model with a different structure, named metric B.

This approach utilizes two separate datasets: a smaller, MOS-annotated dataset (MD) for training metric A, and a larger, unlabeled dataset, referred to as the unlabeled large dataset (ULD), for training metric B. The ULD is instrumental in preventing the model from overfitting during training. To train model B using the ULD dataset, MOS values (target labels) are necessary for each image.

As such, the transfer learning process presented in [56] begins by using the pre-trained model A to estimate the quality values of images in the ULD . Therefore, metric A evaluates the quality of images and produces a predicted assessment (PA) array for the ULD. Assuming model A is sufficiently accurate, the PA values can serve as substitute MOS, albeit with inherent noise:

$$MOS_{ULD} = PA = MOS_{true} + \epsilon. \tag{2.1}$$

Here, MOS_{ULD} denotes the MOS of the images in the ULD, PA denotes predicted pssessments of image quality by metric A, and MOS_{true} denotes the true, but unknown, MOS of the ULD.

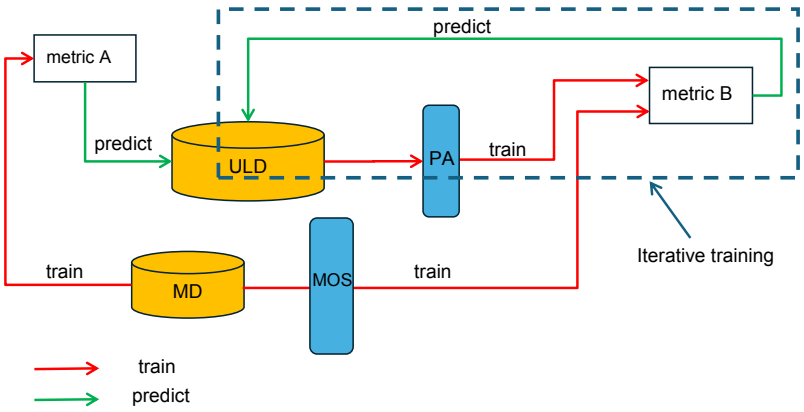


Figure 1. Structure scheme of the proposed iterative transfer learning.

Algorithm 1 Iterative Transfer Learning

1: train metric A on dataset MD

2: collect large ULD dataset ⟹ *Images were randomly selected from GOI [51] dataset*

3 : $A(ULD) \rightarrow PA_1$ ⟹ *Utilize model A to generate quality predictions PA for images in ULD*

4: $P(MD) = N(MD)/(N(MD) + N(ULD))$ ⟹ *Probability of usage of images from MD; N(MD) and N(ULD) are the dataset sizes*

5: $P(ULD) = 1 - P(MD)$ ⟹ *Probability of selecting training images from ULD*

6: **for** (j) ∈ Iteration **do**

7: $Batch(ULD\ and\ MD) = (P(MD) \times N(MD) + P(ULD) \times N(ULD))$ ⟹ *selecting images of the batch according to the probability*

8: Train metric B on dataset $Batch(ULD\ and\ MD)$ ⟹ PA_j *and MOS of MD are used as ground truth data*

9: $B(ULD) \rightarrow PA'_j$ ⟹ *Use pre-trained metric B to re-calculate PA*

10: $PA_{j+1} = PA_j + K(PA'_j - PA_j)$ ⟹ *Updating the PA values, the default K is 1*

11: **end for**

The noise variable ϵ represents the prediction errors of metric A. If metric A performs satisfactorily, it is anticipated that the error will exhibit a distribution that closely approximates a Gaussian distribution. Notably, a noisy MOS situation is a common condition in NR-IQA metric training. MOS are prone to noise due to the sample size of observers' opinions that is used for averaging and obtaining

MOS.

Above, we detailed the method described in [56]. In this paper, we have elevated this technique by introducing an iterative training procedure leveraging both the ULD and MD datasets. With this approach, we harness the combined power of ULD and MD for training purposes. Following this, we use metric B to calculate the PA of ULD for another round (updated PA values). These updated PA values are then used to train metric B again. This iterative process can be repeated multiple times, enabling the PA values to evolve and eventually stabilize. Once stable, these values represent the MOS of ULD.

To gain a thorough understanding of the image selection process from both datasets and the updating of PA values, please refer to Algorithm 1.

The role of ULD is essential in the training process and it is created based on three characteristics:

1) Comprehensiveness: ULD must encompass a diverse range of image quality factors typical of existing MOS-annotated image databases.

2) Consistency: ULD images should span the same PA value range as the MOS of MD used in training model A.

3) Challenge: ULD images should pose a sufficient challenge for PA prediction.

For refining PA, the equation in line 10 of Algorithm 1 is used. The K coefficient is introduced within this equation, and in this experiment, it is set to 1. Choosing a smaller value for K provides smoother, and potentially more effective training and increasing the number of iterations that can lead to the better performance.

3. TBCNet architecture for NR-IQA

This section introduces TBCNet, a CNN-based architecture for NR-IQA, which we train using ITL. Our model, illustrated in Figure 2, incorporates two branches for image analysis, as our goal is to encompass both high-level and low-level image features. High-level image characteristics contain spatial scene configurations, object placements, color composition, and image clarity. The majority of NR-IQA models exclusively depend on the examination of patches with dimensions of 224×224 or 512×512. This approach tends to neglect the elements of high-level image characteristics and misinterpret the global attributes of the image. Additionally, our model effectively handles low-level image characteristics, including factors like noise type and intensity, blur type and intensity, sharpness, and the presence of compression artifacts.

This network has a minimal number of parameters, making it lightweight and efficient for implementation. Furthermore, its processing time remains consistent across varying image sizes due to the method of input patch selection. In contrast, many neural networks that follow the full-image analysis method such as KonCept512, do not adhere to this criterion, resulting in prolonged processing times.

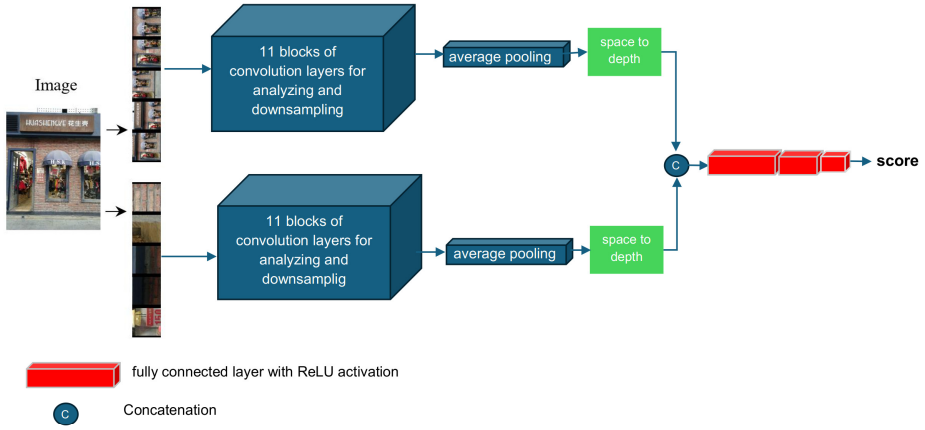


Figure 2. Structure of TBCNet with two branches: one for high-frequency and the other one for low-frequency features.

3.1. Structural scheme of the proposed metric

The structural scheme of TBCNet is presented in Figure 2. TBCNet accepts two inputs, each receiving RGB images standardized to a size of 720×128, regardless of the original image size. This design guarantees consistent analysis time for all images. The first branch is designed to input the low-level image features and the second branch is set up to analyze high-level image features. Each input comprises five segments of the input image, separated by black stripes, which simplifies the network's edge analysis.

Each branch of TBCNet consists of 11 blocks that perform analysis and downscaling. The analysis and feature extraction are performed by eight blocks. Regularly, after each two/three blocks of analysis, one block of downscaling is applied to reduce the size of the feature maps. In total, it has nine blocks of analysis and three blocks of downscaling. The architecture of the blocks are similar and consist of three parallel connections. An Example of one analysis block is illustrated in Figure 3.

In the first connection, there are three convolution kernels followed by ReLU activations. The second and the third connections have two and one kernel, respectively. For downscaling, instead of pure pooling layers, the last convolution kernel of each connection uses strides and no padding.

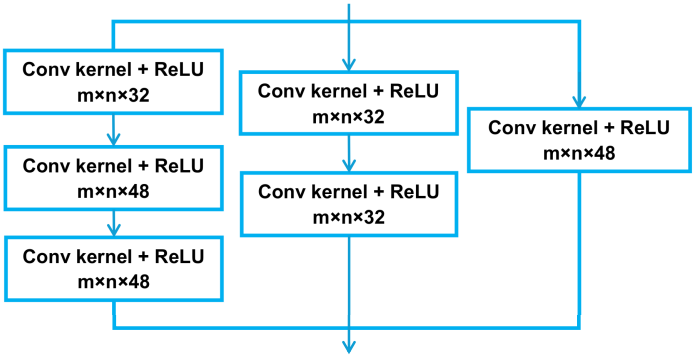


Figure 3. Structure of an analysis block of TBCNet.

Three parallel connections are inspired by skip connections in ResNet [58]. Skip connections link features before and after processing [49, 58]. In our model, as a skip connection, we incorporated information from three distinct parallel processes. By concatenating features from different depths,

the network can capture a more comprehensive representation of the data. Each depth level, in the parallel connections within each block, has the ability to capture different features of the image patch. Concatenating them enables us to integrate the maximum amount of image feature information.

Previously, it was mentioned that a black stripe is used to separate image patches. This separation of patches allows the network to distinguish between different scenes and their compositions. Since each input contains five image patches, at the end of the process for each branch, five tensors with the size of 1×128 are created and result in a $5 \times 1 \times 128$ feature space, which is delivered to fully connected layers. Finally, two branches will unite into one fully connected layer by concatenating the outputs. The final unified layer is responsible for integrating the feature maps and the final score. As the loss function, L2 is adapted for this model.

3.2. Formation of network inputs

The proposed TBCNet model is a deep network with two branches and two specially prepared inputs (see Figure 2). The input for analyzing low-level features consists of five image fragments selected randomly with a size of 256×256 pixels, which are downsampled to a size of 128×128. An example of the patch selection process is shown in Figure 4 (left). These selected patches are desirable for analyzing noise level, blur level, and other low-level features and were selected randomly, which guaranteed a wide variety of patches, from the simplest to those with high-frequency features. This element of randomness ensures the diversity of the patches chosen.

For high-level features, however, a larger square window is used to crop parts of the image. To select the square size, we used the concept of the golden ratio (demonstrated in Figure 4 (right)), the ratio of the desired square edge to the rest of the image. This proportion has been appreciated for its aesthetic qualities and is believed to be visually pleasing to the human eye. It has been used in various fields, such as art and architecture. The window is moved to all directions (center, top, bottom, left, and right), five image fragments containing high-level features are cropped, and afterward, they are downscaled to 128×128 size. Upon extracting and preparing high-level and low-level patches, they undergo separation by inserting dark pixels between them.

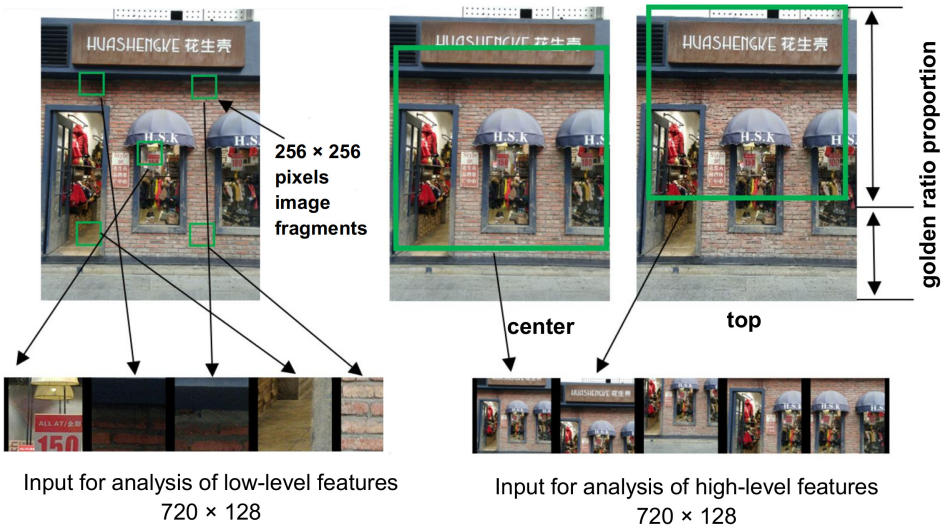


Figure 4. Low-level and high-level patch selection.

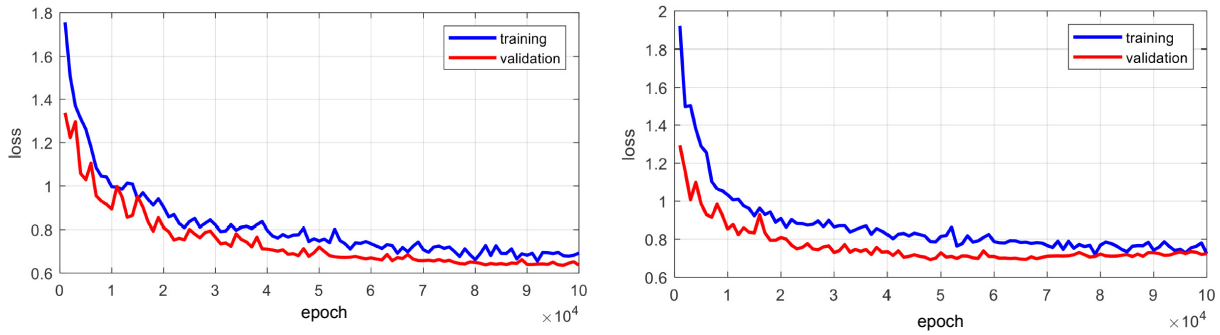
4. Transfer learning from KonCept512 to TBCNet

In this paper, we select KonCept512 as metric A due to its good performance. Koncept512 is a conventional method of transfer learning that employs the pre-trained Inception-v2 model [43] as its base model. To train this metric, we need to construct our dataset MD. For this purpose, we selected six datasets: FLIVE [34], Konik-10k [35], NRTID [36], HTID [37], SPAQ [38], and WILD [39]. These datasets are among the most commonly used no-reference (NR) datasets for training and validating the performance of NR-IQA methods. They encompass a vast number of images, collectively around 60,000 images, and contain authentic distortions (except for HTID, which has synthesized distortion), ensuring the robustness and universality of the metric A. FLIVE is one of the largest datasets, consisting of real-world images from user-generated content (images captured by digital cameras and smartphones). The distortions it covers stem from poor lighting, camera quality, and compression artifacts. Konik-10k, with nearly 10,000 images, is also considered one of the large datasets. Its content is gathered from online photo and video hosting platforms, making it diverse in content. On the other hand, SPAQ is a dataset with images captured by smartphones that reflect the limitations of smartphone cameras in its distortions. Finally, the WILD and NRTID datasets are smaller but worth including among the datasets because they contain images from different sources with naturally occurring distortions, including noise, blur, and artifacts from hardware limitations. To create a large and diverse dataset with MOS values, we followed the method outlined in [59] to construct the MD and combine the MOS values.

As previously stated, our approach also requires the ULD dataset. The ULD dataset was formed by randomly choosing 360,000 images from the Google Open Images (GOI) datasets [57], which contain a vast range of images with varied content. The reason behind selecting images from GOI is that it has a reach image selection that helps the model to adapt to different real-world scenarios, and it includes images from 60,000 different categories. Subsequently, the KonCept512 metric trained on MD was utilized to compute quality scores for ULD. In our setup, the suggested TBCNet functions as the metric B, which means that knowledge is to be transferred from Koncept512 to TBCNet. As we are using ITL, TBCNet is trained on both the ULD and MD datasets for two iterations.

TBCNet is trained in three states. The first state is the transfer learning proposed in [56], and then we use the proposed iterative training for one and two iterations. TBCNet is designed and trained in MATLAB 2022 with the custom training loop technique. The model was trained on 100,000 epochs. The learning rate starts from 0.0001 and is decreased by half at every 20,000 epochs.

According to the curves of training and validation in Figure 5 (left), which shows the direct training on the MD dataset, the decrease of loss stops near epoch 50,000 (evidence of overfitting) though in Figure 5 (right), the reduction continues near to epoch 100,000 (no visible overfitting). Hence, we can assert that our approach also addresses the issue of overfitting.



Direct training on MD. Iterative transfer learning.

Figure 5. Training and validation curves for 100,000 epochs.

5. Experimental results

In this section, we present the results of our proposed NR-IQA method and compare it to state-of-the-art metrics. spearman's rank order correlation coefficient (SROCC) is adopted as the primary quality metric for evaluating the performance. SROCC is a non-parametric measure of statistical dependence between two ranked variables, and it is commonly used to evaluate the correlation between predicted and ground truth quality scores in IQA tasks.

For testing purposes, a selection of images was gathered from each of the six NR datasets: FLIVE, Konik-10k, NRTID, HTID, SPAQ, and WILD. These images were not part of the training set *MD*. The number of images selected from each dataset is as follows: FLIVE = 250 images, HTID = 200 images, KonIQ-10k = 200 images, NRTID = 50 images, SPAQ = 200 images, and WildLive = 100 images. Table 1 illustrates the training of TBCNet in three modes: 1) trained with the *ULD* dataset (based on transfer learning in [56]), 2) trained with the *ULD* and *MD* datasets (ITL), and 3) trained with *ULD* and *MD* in two iterations (ITL 2 iterations).

Table 1. SROCC of metrics on 6 datasets and the average SROCC. TBCNnet (*ULD*, *MD*) indicates that TBCNet was trained on the *ULD* and *MD* datasets.

Metric	KonIQ-10k	FLIVE	WILD	NRTID	SPAQ	HTID	ALL
TBCNet (*ULD*)	0.813	**0.425**	0.657	**0.812**	0.884	0.730	0.791
TBCNet (*ULD*, *MD*) (Ours)	0.810	0.416	0.692	0.775	0.893	0.837	0.850
TBCNet (*ULD*, *MD* two iterations) (Ours)	0.816	0.382	0.719	0.785	**0.903**	**0.849**	**0.857**
KonCept512 [35]	**0.925**	0.411	**0.768**	0.710	0.855	0.681	0.790
IMQNet [56]	0.844	0.407	0.699	0.691	0.732	0.826	0.700
Otroshi [40]	0.827	0.274	0.636	0.633	0.759	0.528	0.673
UIQA [33]	0.635	0.226	0.515	0.540	0.788	0.613	0.625
Paq2Pic [34]	0.618	0.461	0.638	0.741	0.851	0.182	0.624
DBCNN [60]	0.565	0.182	0.450	0.494	0.615	0.579	0.605
CDIVINE [24]	0.492	0.215	0.525	0.723	0.669	0.415	0.572
Desique [26]	0.388	0.004	0.435	0.655	0.605	0.228	0.452

We included several well-known metrics in the analysis and the results of the comparison are demonstrated in Table 1. It can be seen that using both ULD and MD together in training TBCNet results in a better SROCC compared to using ULD alone. Furthermore, employing two iterations marginally enhances this SROCC value. It is important to note that the KonCept512 model, trained exclusively on the KonIQ-10k dataset, delivers the highest SROCC value of 0.925 for the same test set.

This can be explained by the model's specific training on the KonIQ-10k dataset, leading to a bias and optimal performance on this particular dataset. Given these considerations, allow us to discuss the results for the three modes of TBCNet and compare them with KonCept512. TBCNet, trained on a vast dataset with artificial MOS, demonstrates increased generality as the dataset size expands.

Our TBCNet, trained on ULD and MD, benefits from extensive data and does not display bias toward specific test sets. The rationale behind employing ITL is to show that even without MOS data, our model can approximate the performance of traditionally trained CNNs. This shows the robustness of our model and its ability to achieve high accuracy without being influenced by the dataset on which it was trained.

5.1. Generality of ITL

In the next experiment, we will perform a cross-dataset evaluation of the efficiency of ITL in comparison with KonCept512 and the methods that use a single-stage transfer learning. We made a slight change in data selection so the training data and the test data are taken from different databases. In our experimental setup, FLIVE and SPAQ datasets serve as the ULD_{new}, and Wild and Koniq as the dataset MD_{new}, and NRTID as the dataset for testing.

In Table 2, we have presented the results of SROCC for the different versions of TBCNet and the KonCept512 method trained on MD_{new}.

It should be noted that in this experimental setup, we made modifications to the patch selection process. Specifically, we did not downsample the low-level patch dataset. For the high-level image patch, we selected a single patch of size 348×512, extracted from the center of the image.

Table 2. SROCC of the metrics for the cross-dataset experiment.

Metric	NRTID
TBCNet (ULD_{new})	0.714
TBCNet (ULD_{new} +MD_{new})	0.735
TBCNet(ULD_{new} +MD_{new} 2 iterations	0.729
TBCNet (ULD_{new} +MD_{new}) 3 iterations	**0.740**
KonCept512(D_{new})	0.708

5.2. Efficiency of TBCNet implementation

Another critical characteristic of TBCNet is its light structure and suitability for implementation, which is analyzed and compared with other state-of-the-art metrics in Table 3. This table compares model sizes (number of trained parameters) and the time (seconds) it takes to predict the quality of one image in different resolutions on a computer with CPU i5-9300H 2.4 GH and GPU NVIDIA GeForce

GTX 1660 Ti 6 GB. The results displayed in Table 3 illustrate TBCNet's remarkably low computation time, indicating the efficiency of our method in terms of parameter count. This provides evidence for its suitability in practical implementation. Also, the execution time for TBCNet for images with varying resolution is constant due to the efficient strategy of selecting input patches.

Table 3. Computational cost and number of parameters.

Model	Parameters (millions)	Image quality prediction time (seconds)		
		1024×768	3840×2160	7680×4320
TBCNet (Ours)	**4**	**0.09**	**0.09**	**0.09**
IMQNet [56]	21	0.25	2	out of memory
HyperIQA [54]	27	0.77	0.77	0.77
Contrique [61]	28	5.4	5.7	6.0
KonCept512 [35]	60	**0.09**	0.3	1.1
NasNet [62]	89	3.1	3.1	3.1
Tres [55]	152	3.2	3.2	3.2

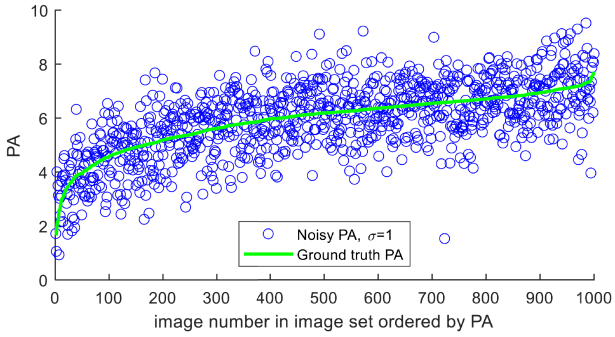
5.3. Analysis of the denoising ability of iterative transfer learning

The phenomenon of noisy MOS frequently occurs due to the large number of images and the limited number of evaluators available to rate them. For example, FLIVE is among the largest datasets available. However, the ratings for each image are only averaged across 20 subjects, leading to a noisy MOS.

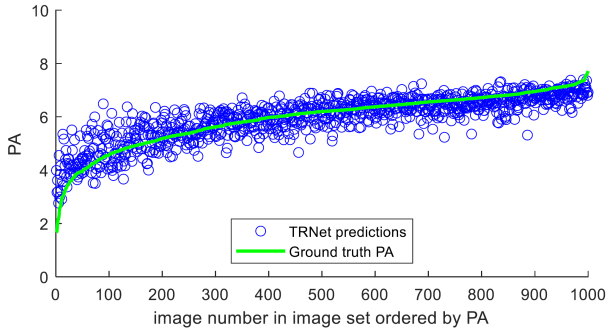
In response to this challenge, ITL can be an effective solution. We demonstrate that TBCNet, when trained via ITL for two iterations, can effectively reduce noise in *PA* values, leading to more consistent image scores. This results in a denoising effect, suggesting the potential application of our model in enhancing the quality of MOS in recently created image databases.

To validate the effectiveness of ITL in this scenario, we generated noisy *PA* values by introducing additive white Gaussian noise with a variance of 1 to the ground truth *PA* values. The visual demonstration of noisy values can be seen in Figure 6 (left). The results show that the noisy *PA*s are converging toward the ground truth *PA* after being updated by ITL in two iterations (Figure 6 (right)), and the noisy values of *PA* are scattered with more deviation from the ground truth *PA* compared to the reconstructed *PA*s (Figure 6 (left)). It should be mentioned that the model used for this experiment is TBCNet, trained by ITL in two iterations. *PA* values predicted by TBCNet after two iterations of the ITL are visualized in Figure 6 (right).

Figure 7 shows the curve of the quality of TBCNet in terms of SROCC versus the level of noise presented in the *PA*. It is clearly seen that as the noise level in the *PA* increases, the efficiency of the proposed transfer learning decreases. However, it still remains operational over a vast range of noise levels.



Comparison of ground truth *PA* and noisy *PA*.



Comparison of ground truth *PA* and reconstructed PB by TBCNet.

Figure 6. Comparing *PA*, noisy *PA*, and reconstructed *PA* by TBCNet trained by ITL in two iterations.

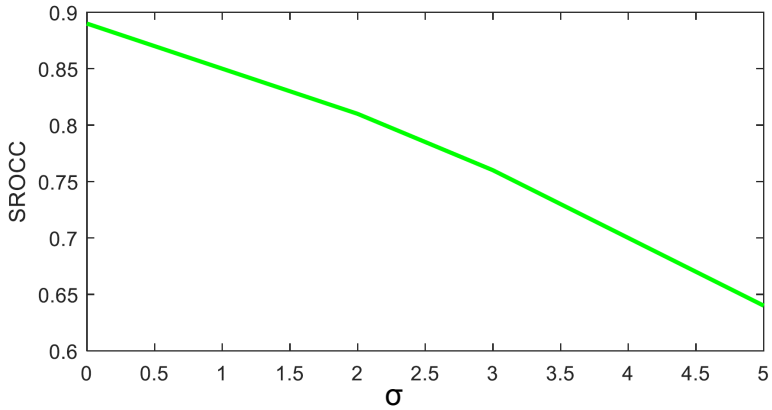


Figure 7. The efficiency of the TBCNet model (trained by ITL) according to the noise standard deviation 0...5 in *PA*.

6. Conclusions and future work

This paper introduces a robust and efficient solution to address key challenges in no-reference image quality assessment. We proposed an iterative transfer learning methodology that overcomes the limitations of traditional transfer learning and addresses the issue of over-learning. By utilizing large and diverse datasets, our model ensures a better generalization compared to the conventional transfer learning for a wide range of image quality conditions. Additionally, we introduced TBCNet, a two-branch CNN architecture designed to achieve fast and efficient implementation while maintaining high accuracy. TBCNet's multi-scale and multi-level feature extraction enables it to capture both local details and global semantics, making it a powerful tool for image quality assessment. The experimental results showcase the notable performance of TBCNet, as evidenced by the short execution time and the good SROCC values simultaneously. Moreover, the TBCNet model, along with iterative transfer learning, demonstrates exceptional resilience to noisy MOS data, a common occurrence in real-world scenarios. This robustness allows the model to be used effectively for denoising MOS, enhancing the

accuracy and credibility of NR-IQA predictions. In our current work, we exclusively choose random segments for low-level features and utilize large patches of the image for high-level features. However, for future studies, there is potential to incorporate attention modules into the TBCNet architecture for improved segmentation of the image's semantic regions. Additionally, the use of various types of loss functions could yield enhanced results.

Conflict of interest

The authors declare no conflict of interest.

References

1. P. Mohammadi, A. Ebrahimi-Moghadam, S. Shirani, Subjective and objective quality assessment of image: a survey, arXiv: 1406.7799. http://dx.doi.org/10.48550/arXiv.1406.7799
2. S. Möller, A. Raake, *Quality of experience: advanced concepts, applications and methods*, Cham: Springer, 2014. http://dx.doi.org/10.1007/978-3-319-02681-7
3. H. Liu, I. Heynderickx, A perceptually relevant no-reference blockiness metric based on local image characteristics, *EURASIP J. Adv. Sig. Pr.*, **2009** (2009), 263540. http://dx.doi.org/10.1155/2009/263540
4. S. Golestaneh, D. Chandler, No-reference quality assessment of JPEG images via a quality relevance map, *IEEE Signal Proc. Lett.*, **21** (2014), 155–158. http://dx.doi.org/10.1109/LSP.2013.2296038
5. Y. Zhan, R. Zhang, No-reference JPEG image quality assessment based on blockiness and luminance change, *IEEE Signal Proc. Lett.*, **24** (2017), 760–764. http://dx.doi.org/10.1109/LSP.2017.2688371
6. H. Liu, N. Klomp, I. Heynderickx, A no-reference metric for perceived ringing artifacts in images, *IEEE T. Circ. Syst. Vid.*, **20** (2010), 529–539. http://dx.doi.org/10.1109/TCSVT.2009.2035848
7. L. Liang, S. Wang, J. Chen, S. Ma, D. Zhao, W. Gao, No-reference perceptual image quality metric using gradient profiles for JPEG2000, *Signal Process.-Image*, **25** (2010), 502–516. http://dx.doi.org/10.1016/j.image.2010.01.007
8. N. Ponomarenko, V. Lukin, O. Eremeev, K. Egiazarian, J. Astola, Sharpness metric for no-reference image visual quality assessment, *Proceedings of Image Processing: Algorithms and Systems X; and Parallel Processing for Imaging Applications II*, 2012, 829519. http://dx.doi.org/10.1117/12.906393
9. N, Narvekar, L. Karam, A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection, *Proceedings of International Workshop on Quality of Multimedia Experience*, 2009, 87–91. http://dx.doi.org/10.1109/QOMEX.2009.5246972
10. F. Crete, T. Dolmiere, P. Ladret, M. Nicolas, The blur effect: perception and estimation with a new no-reference perceptual blur metric, *Proceedings of Human vision and electronic imaging XII*, 2007, 649201. http://dx.doi.org/10.1117/12.702790

11. C. Ma, C. Yang, X. Yang, M. Yang, Learning a no-reference quality metric for single-image super-resolution, *Comput. Vis. Image Und.*, **158** (2017), 1–16. http://dx.doi.org/10.1016/j.cviu.2016.12.009
12. K. Zhang, D. Zhu, J. Li, X. Gao, F. Gao, J. Lu, Learning stacking regression for no-reference super-resolution image quality assessment, *Signal Process.*, **178** (2021), 107771. http://dx.doi.org/10.1016/j.sigpro.2020.107771
13. A. Shulev, A. Gotchev, A. Foi, I. Roussev, Threshold selection in transform-domain denoising of speckle pattern fringes, *Proceedings of Holography 2005: International Conference on Holography, Optical Recording, and Processing of Information*, 2006, 625220. http://dx.doi.org/10.1117/12.677284
14. V. Lukin, D. Fevralev, N. Ponomarenko, S. Abramov, O. Pogrebnyak, K. Egiazarian, et al., Discrete cosine transform—based local adaptive filtering of images corrupted by nonstationary noise, *J. Electron. Imag.*, **19** (2010), 023007. http://dx.doi.org/10.1117/1.3421973
15. X. Liu, M. Tanaka, M. Okutomi, Noise level estimation using weak textured patches of a single noisy image, *Proceedings of 19th IEEE International Conference on Image Processing*, 2012, 665–668. http://dx.doi.org/10.1109/ICIP.2012.6466947
16. Z. Yue, H. Yong, Q. Zhao, D. Meng, L. Zhang, Variational denoising network: toward blind noise modeling and removal, *Proceedings of 33rd Conference on Neural Information Processing Systems*, 2019, 1–12.
17. S. Bahnemiri, M. Ponomarenko, K. Egiazarian, Learning-based noise component map estimation for image denoising, *IEEE Signal Proc. Lett.*, **29** (2022), 1407–1411. http://dx.doi.org/10.1109/LSP.2022.3169706
18. A. Moorthy, A. Bovik, Blind image quality assessment: from natural scene statistics to perceptual quality, *IEEE T. Image Process.*, **20** (2011), 3350–3364. http://dx.doi.org/10.1109/TIP.2011.2147325
19. A. Mittal, A. Moorthy, A. Bovik, No-reference image quality assessment in the spatial domain, *IEEE T. Image Process.*, **21** (2012), 4695–4708. http://dx.doi.org/10.1109/TIP.2012.2214050
20. J. Yan, X. Bai, Y. Xiao, Y. Zhang, X. Lv, No-reference remote sensing image quality assessment based on gradient-weighted natural scene statistics in spatial domain, *J. Electron. Imaging*, **28** (2019), 013033. http://dx.doi.org/10.1117/1.JEI.28.1.013033
21. A. Mittal, R. Soundararajan, A. Bovik, Making a "completely blind" image quality analyzer, *IEEE Signal Proc. Lett.*, **20** (2013), 209–212. http://dx.doi.org/10.1109/LSP.2012.2227726
22. M. Saad, A. Bovik, C. Charrier, Blind image quality assessment: a natural scene statistics approach in the DCT domain, *IEEE T. Image Process.*, **21** (2012), 3339–3352. http://dx.doi.org/10.1109/TIP.2012.2191563
23. A. Moorthy, A. Bovik, A two-step framework for constructing blind image quality indices, *IEEE Signal Proc. Lett.*, **17** (2010), 513–516. http://dx.doi.org/10.1109/LSP.2010.2043888
24. Y. Zhang, A. Moorthy, D. Chandler, A. Bovik, C-DIIVINE: no-reference image quality assessment based on local magnitude and phase statistics of natural scenes, *Signal Process.-Image*, **29** (2014), 725–747. http://dx.doi.org/10.1016/j.image.2014.05.004

25. R. Campos, E. Salles, Robust statistics and no-reference image quality assessment in curvelet domain, arXiv: 1902.03842. http://dx.doi.org/10.48550/arXiv.1902.03842
26. Y. Zhang, D. Chandler, No-reference image quality assessment based on log-derivative statistics of natural scenes, *J. Electron. Imaging*, **22** (2013), 043025. http://dx.doi.org/10.1117/1.JEI.22.4.043025
27. X. Chen, Q. Zhang, M. Lin, G. Yang, C. He, No-reference color image quality assessment: from entropy to perceptual quality, *J. Image Video Proc.*, **2019** (2019), 77. http://dx.doi.org/10.1186/s13640-019-0479-7
28. F. Ou, Y. Wang, G. Zhu, A novel blind image quality assessment method based on refined natural scene statistics, *Proceedings of IEEE International Conference on Image Processing (ICIP)*, 2019, 1004–1008. http://dx.doi.org/10.1109/ICIP.2019.8803047
29. L. Liu, H. Dong, H. Huang, A. Bovik, No-reference image quality assessment in curvelet domain, *Signal Process.-Image*, **29** (2014), 494–505. http://dx.doi.org/10.1016/j.image.2014.02.004
30. A. Smola, B. Schölkopf, A tutorial on support vector regression, *Stat. Comput.*, **14** (2004), 199–222. http://dx.doi.org/10.1023/B:STCO.0000035301.49549.88
31. L. Kang, P. Ye, Y. Li, D. Doermann, Convolutional neural networks for no-reference image quality assessment, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, 1733–1740. http://dx.doi.org/10.1109/CVPR.2014.224
32. Y. Li, L. Po, L. Feng, F. Yuan, No-reference image quality assessment with deep convolutional neural networks, *Proceedings of IEEE International Conference on Digital Signal Processing (DSP)*, 2016, 685–689. http://dx.doi.org/10.1109/ICDSP.2016.7868646
33. T. Lu, A. Dooms, Towards content independent no-reference image quality assessment using deep learning, *Proceedings of IEEE 4th International Conference on Image, Vision and Computing (ICIVC)*, 2019, 276–280. http://dx.doi.org/10.1109/ICIVC47709.2019.8981378
34. Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, A. Bovik, From patches to pictures (PaQ-2-PiQ): mapping the perceptual space of picture quality, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, 3575–3585. http://dx.doi.org/10.1109/CVPR42600.2020.00363
35. V. Hosu, H. Lin, T. Sziranyi, D. Saupe, KonIQ-10k: an ecologically valid database for deep learning of blind image quality assessment, *IEEE T. Image Process.*, **29** (2020), 4041–4056. http://dx.doi.org/10.1109/TIP.2020.2967829
36. N. Ponomarenko, O. Eremeev, V. Lukin, K. Egiazarian, Statistical evaluation of no-reference image visual quality metrics, *Proceedings of 2nd European Workshop on Visual Information Processing (EUVIP)*, 2010, 50–54. http://dx.doi.org/10.1109/EUVIP.2010.5699121
37. M. Ponomarenko, S. Bahnemiri, K. Egiazarian, O. Ieremeiev, V. Lukin, V. Peltoketo, et al., Color image database HTID for verification of no-reference metrics: peculiarities and preliminary results, *Proceedings of 9th European Workshop on Visual Information Processing (EUVIP)*, 2021, 1–6. http://dx.doi.org/10.1109/EUVIP50544.2021.9484005

38. Y. Fang, H. Zhu, Y. Zeng, K. Ma, Z. Wang, Perceptual quality assessment of smartphone photography, *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, 3677–3686. http://dx.doi.org/10.1109/CVPR42600.2020.00373
39. D. Ghadiyaram, A. Bovik, Massive online crowdsourced study of subjective and objective picture quality, *IEEE T. Image Process.*, **25** (2016), 372–387. http://dx.doi.org/10.1109/TIP.2015.2500021
40. H. Otroshi-Shahreza, A. Amini, H. Behroozi, No-reference image quality assessment using transfer learning, *Proceedings of 9th International Symposium on Telecommunications (IST)*, 2018, 637–640. http://dx.doi.org/10.1109/ISTEL.2018.8661024
41. Y. Feng, Y. Cai, No-reference image quality assessment through transfer learning, *Proceedings of IEEE 2nd International Conference on Signal and Image Processing (ICSIP)*, 2017, 90–94. http://dx.doi.org/10.1109/SIPROCESS.2017.8124512
42. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv: 1409.1556. http://dx.doi.org/10.48550/arXiv.1409.1556
43. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, 2818–2826. http://dx.doi.org/10.1109/CVPR.2016.308
44. A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, et al., Mobilenets: efficient convolutional neural networks for mobile vision applications, arXiv: 1704.04861. http://dx.doi.org/10.48550/arXiv.1704.04861
45. H. Talebi, P. Milanfar, NIMA: neural image assessment, *IEEE T. Image Process.*, **27** (2018), 3998–4011. http://dx.doi.org/10.1109/TIP.2018.2831899
46. F. Chollet, Xception: deep learning with depthwise separable convolutions, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, 1251–1258. http://dx.doi.org/10.1109/CVPR.2017.195
47. S. Bianco, L. Celona, P. Napoletano, R. Schettini, On the use of deep learning for blind image quality assessment, *SIViP*, **12** (2018), 355–362. http://dx.doi.org/10.1007/s11760-017-1166-8
48. X. Liu, J. Van De Weijer, A. Bagdanov, Rankiqa: learning from rankings for no-reference image quality assessment, *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2017, 1040–1049. http://dx.doi.org/10.1109/ICCV.2017.118
49. O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, In: *Medical image computing and computer-assisted intervention*, Cham: Springer, 2015, 234–241. http://dx.doi.org/10.1007/978-3-319-24574-4_28
50. J. Gu, G. Meng, S. Xiang, C. Pan, Blind image quality assessment via learnable attention-based pooling, *Pattern Recogn.*, **91** (2019), 332–344. http://dx.doi.org/10.1016/j.patcog.2019.02.021
51. S. Yang, Q. Jiang, W. Lin, Y. Wang, SGDNet: an end-to-end saliency-guided deep neural network for no-reference image quality assessment, *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, 1383–1391. http://dx.doi.org/10.1145/3343031.3350990
52. P. Grüning, E. Barth, Fp-nets for blind image quality assessment, *Journal of Perceptual Imaging*, **4** (2021), jpi0143. http://dx.doi.org/10.2352/J.Percept.Imaging.2021.4.1.010402

53. W. Lu, W. Sun, X. Min, Z. Zhang, T. Wang, W. Zhu, et al., Blind surveillance image quality assessment via deep neural network combined with the visual saliency, In: *Artificial intelligence*, Cham: Springer, 2022, 136–146. http://dx.doi.org/10.1007/978-3-031-20500-2_11

54. S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, Blindly assess image quality in the wild guided by a self-adaptive hyper network, *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, 3667–3676. http://dx.doi.org/10.1109/CVPR42600.2020.00372

55. S. Golestaneh, S. Dadsetan, K. Kitani, No-reference image quality assessment via transformers, relative ranking, and self-consistency, *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, 1220–1230. http://dx.doi.org/10.1109/WACV51458.2022.00404

56. M. Ponomarenko, S. Bahnemiri, K. Egiazarian, Transfer learning for no-reference image quality metrics using large temporary image sets, *Electronic Imaging*, **34** (2022), 219-1–219-5. http://dx.doi.org/10.2352/EI.2022.34.14.COIMG-219

57. A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, et al., The open images dataset v4: unified image classification, object detection, and visual relationship detection at scale, *Int. J. Comput. Vis.*, **128** (2020), 1956–1981. http://dx.doi.org/10.1007/s11263-020-01316-z

58. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, 770–778. http://dx.doi.org/10.1109/CVPR.2016.90

59. A. Kaipio, M. Ponomarenko, K. Egiazarian, Merging of MOS of large image databases for no-reference image visual quality assessment, *Proceedings of IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, 2020, 1–6. http://dx.doi.org/10.1109/MMSP48831.2020.9287141

60. W. Zhang, K. Ma, J. Yan, D. Deng, Z. Wang, Blind image quality assessment using a deep bilinear convolutional neural network, *IEEE T. Circ. Syst. Vid.*, **30** (2020), 36–47. http://dx.doi.org/10.1109/TCSVT.2018.2886771

61. P. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, A. Bovik, Image quality assessment using contrastive learning, *IEEE T. Image Process.*, **31** (2022), 4149–4161. http://dx.doi.org/10.1109/TIP.2022.3181496

62. N. Ahmed, S. Asif, BIQ2021: a large-scale blind image quality assessment database, *J. Electron. Imaging*, **31** (2022), 053010. http://dx.doi.org/10.1117/1.JEI.31.5.053010

AIMS Press