IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 2 November 2023, accepted 26 November 2023, date of publication 7 December 2023,
date of current version 8 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3340266

RESEARCH ARTICLE

VRL-IQA: Visual Representation Learning for Image Quality Assessment

MUHAMMAD AZEEM ASLAM[1], XU WEI[2], NISAR AHMED[3], GULSHAN SALEEM[4], TUBA AMIN[5], AND HUI CAIXUE[1]

[1]School of Information Engineering, Xi'an Eurasia University, Xi'an, Shaanxi 710065, China
[2]Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, Jilin 130033, China
[3]Department of Computer Engineering, University of Engineering and Technology, Lahore, Lahore 54890, Pakistan
[4]Department of Computer Science, Lahore Garrison University, Lahore 54000, Pakistan
[5]Department of Computer Science, Government College University Faisalabad, Faisalabad 38000, Pakistan

Corresponding author: Muhammad Azeem Aslam (azeem@eurasia.edu)

This work was supported in part by the School of Information Engineering, Eurasia University.

ABSTRACT With the increasing prevalence of digital multimedia devices and the growing reliance on compression and wireless data transmission, evaluating image quality remains a persistent challenge. This study addresses the limitations of image quality assessment stemming from the expense of data annotation and the scarcity of labeled training datasets. Leveraging visual representation learning, our approach involves training a deep Convolutional Neural Network on a large image dataset generated by simulating 165 distortion scenarios across 150,000 images, resulting in 24.75 million distorted images. These distortions are labeled using an ensemble of full-reference quality assessment models. The trained model undergoes fine-tuning on diverse datasets, including TID2013, Kadid-10K, KonIQ-10K, and BIQ2021, encompassing both simulated and authentic distortions. The fine-tuning process achieves state-of-the-art image quality assessment performance, yielding Spearman's correlation coefficients of 0.921, 0.893, 0.884, and 0.793, respectively, for the four datasets. Comparative analysis with an ImageNet pre-trained model demonstrates superior performance in terms of Pearson and Spearman's correlations, achieving validation criteria with fewer epochs. These findings contribute to the advancement of IQA, offering a promising approach for robust and accurate quality prediction in various applications.

INDEX TERMS Convolutional neural network, image quality assessment, image quality, IQA, transfer learning, visual representation learning.

I. INTRODUCTION

The Human Visual System (HVS) is a sophisticated sensory system that enables us to perceive the world around us. Vision plays a pivotal role in acquiring and retaining over 70% of the information we learn and experience [1]. Deep within the cerebral cortex lies the visual cortex, which is responsible for all visual processing in the human brain. Among all animal species, humans possess the most intricate and advanced visual system.

This transformation has elevated visuals as the primary mode of communication and information transmission. However, the fidelity of this information is heavily contingent on both the fine details of the image and the observer's visual acuity. This, coupled with the rapid advancements in digital multimedia technologies, has amplified the significance of images in conveying information.

Images captured by digital cameras are essentially electrical impulse representations of an object's visual attributes. Artifacts, or unintended characteristics within a digital image, can inadvertently be introduced during acquisition, processing, storage, or transmission. Hence, it becomes imperative to assess the efficacy of different systems in maintaining a high degree of perceived image quality.

Perceptual image quality pertains to how an image is perceived by a human observer. There are two primary approaches for conducting Image Quality Assessment (IQA): subjective approaches, which rely on quality ratings provided

The associate editor coordinating the review of this manuscript and approving it for publication was Gangyi Jiang.

by human observers and are considered the gold standard in quality assessment, and objective approaches, which leverage algorithms to compute a quality score, offering a convenient and expeditious alternative.

Furthermore, objective quality assessment can be broadly categorized based on the availability of reference information, leading to full-reference and no-reference IQA. Traditional metrics like Mean Squared Error (MSE) and Peak Signal-to-Noise Ratio (PSNR) are employed in full-reference IQA, though they are deemed less reliable.

More advanced methods like the Structural Similarity Index Metric (SSIM) [2] and its derivatives such as Multi-Scale SSIM (MS-SSIM) [3], three-component SSIM (3)-SSIM) [4], Complex Wavelet SSIM (CW-SSIM) [5], Information Content Weighted SSIM (IW-SSIM) [6], along with other modern alternatives, prove to be more robust choices for quality assessment in the presence of reference information. In contrast, no-reference IQA approaches have not yet reached the same level of maturity due to the complexity of the problem, stemming from multiple sources of degradation and a lack of access to reference information [7].

In the dynamic world of visual content, the pursuit of optimal image quality has never been more crucial. With the ubiquity of digital media devices and the surge in data transmission through compression and wireless channels, the demand for precise IQA has reached unprecedented heights. IQA plays a pivotal role in fine-tuning bit rates, compression techniques, and processing strategies for these cutting-edge multimedia technologies.

Yet, despite its paramount importance, the domain of IQA has grappled with challenges that have hampered progress. Traditional approaches, relying on handcrafted features or standard regression-based algorithms, have fallen short of achieving the level of predictive accuracy demanded by today's complex visual environments. This is where the power of deep learning and Convolutional Neural Networks (CNNs) steps into the limelight.

The potential of CNNs and other deep learning-based techniques for visual representation learning in IQA is immense. These modern methodologies promise to revolutionize how we perceive and quantify image quality. However, to unlock their full potential, a significant volume of task-specific data and computational resources are required. This is where a critical hurdle arises: the reliance on transfer learning from a generic dataset like ImageNet [8]. While effective, this approach demands prolonged training periods and often leads to less-than-optimal predictive performance due to the fundamental disparity between ImageNet's object recognition focus and the nuanced intricacies of IQA.

In our proposed approach, we embrace the fundamental concept of pre-training on a comprehensive dataset, followed by focused fine-tuning for the specific IQA task at hand. The objective of the study is to streamline the fine-tuning process, requiring fewer data points and computations, all while leveraging the capabilities of the sophisticated NASNet-large CNN model [7]. This model, selected for its remarkable ability to capture complex image-quality features from a large-scale dataset, forms the cornerstone of our innovative strategy.

Through simulated distortions across a staggering 165 unique scenarios applied to 150,000 pristine images from the KADIS-700K dataset [9], we generate a massive 24.75 million distorted images. These, along with 150,000 reference images, become the canvas for our IQA journey. Employing ten full-reference IQA algorithms, we evaluate image quality with meticulous precision. The resulting quality scores, derived through a weighted ensemble of predictions, serve as the bedrock for our pre-training process.

Our journey culminates in a thorough evaluation of benchmark datasets: TID2013 [10], Kadid-10K [9], KonIQ-10K [11], and BIQ2021 [12]. These datasets, ranging from artificially distorted images to authentically distorted ones, provide a comprehensive testing ground for our approach. The results, we believe, not only push the boundaries of IQA performance but also hold tremendous promise for enhancing image quality assessment across a myriad of applications.

In this study, we embark on a transformative quest driven by a profound understanding of the critical role image quality plays in our digital landscape. Through pioneering techniques, we endeavor to not only meet but exceed the demands of modern multimedia technologies, unlocking new possibilities in the realm of image quality assessment. The following are some of the specific contributions of the proposed VRL-IQA model:

- Proposed a novel pre-training method that involves simulating 165 distortion scenarios on a large set of 150,000 pristine images, resulting in 24.75 million distorted images. This innovative approach addresses the challenge of limited annotated data and enables the development of deep learning-based solutions for distortion-agnostic IQA.
- Leveraged the well-established field of full-reference quality assessment by selecting 10 full-reference models to predict quality scores for the 24.75 million distorted images. The ensemble approach, utilizing a weighted average of these models, provides a reliable ground truth for pre-training the IQA model.
- Employed the NASNet-large architecture, a complex and larger CNN model, for pre-training on the upstream data and fine-tuning on the downstream data. This architecture was specifically chosen for its ability to capture intricate image-quality features learned from a large-scale dataset.
- Introduced a quality-aware loss function incorporating an adjusted correlation term alongside error-based terms, with the aim of enhancing the robustness of the trained model.
- Demonstrated the effectiveness of the proposed technique, which integrates quality-aware pre-training and model fine-tuning, in achieving high prediction performance on four benchmark datasets encompassing a diverse range of synthetic and authentic distortions.

II. RELATED WORK

Visual representation learning is concerned with the automatic learning of suitable representations from visual data such as photos or movies. The objective of visual representation learning is to use raw data to learn meaningful representations that capture the underlying semantics and structures in the images. These learned representations can be used for a wide range of tasks, including visual recognition, semantic segmentation, image captioning, and image regression. The development of deep learning methods and the availability of large image datasets have led to tremendous progress in the field of visual representation learning. These advancements paved the way for further research into computer vision and artificial intelligence, resulting in significant improvements in the performance of visual identification tasks. Learning visual representations may be accomplished in several ways, which are explained below.

A. SUPERVISED LEARNING

Visual representation learning using supervised learning refers to training a neural network to learn a mapping from unlabeled visual input (such as images or videos) to a set of desired labels or output. The ImageNet-1K and 21K datasets [8], [13], which contain more than a million and 14 million annotated images, respectively, are a good example of a large dataset that may be used to train a supervised visual representation learning model [14], [15], [16], [17], [18]. High-level image features that are associated with the target labels are learned during training [19]. This trained network may then be applied to another related domain with images of relatively similar types.

CNNs are the most widely used method for visual representation learning, which predominantly performs supervised learning, producing a series of backbone architectures [7], [11], [20], [21], [22], [23], [24], [25]. Most of these architectures are constructed by stacking high-resolution to low-resolution convolutional layers by going deeper to learn high-level representations. These architectures are mainly focused on learning visual recognition using the ImageNet dataset and are fine-tuned to various tasks via transfer learning.

An alternative to these architectures is multi-scale CNN backbones such as Res2Net [26], which performs granular-level multi-scale feature extraction. The architecture employs hierarchical residual connections to replace the bottleneck layer, consequently expanding the receptive field range for each layer in the network. Another noteworthy approach to multi-scale learning is HRNet by Wang et al. [27], wherein the authors adopt a visual representation learning approach. They have explored the use of a multi-stream architecture with high-resolution streams and low-resolution streams, which perform learning in parallel. The information is shared from the low-resolution stream to the high-resolution stream, which provides semantically richer and spatially precise features. Similarly, Ahmed et al. [7] proposed DeepEns, which is a two-stream architecture that performs learning in two parallel streams. Both streams learn features with different CNN backbones and combine the weights via global average pooling. The outcome of their architecture results in improved predictive performance.

Liu et al. [28] proposed Mix-MAE, which performs mix embeddings and masked attention for pre-training. The authors have used the ImageNet-1K dataset to train a hierarchical vision transformer and performed training for 600 epochs on input images of 224×224 pixels. Mix embeddings are used in conjunction with positional embeddings to perform efficient learning of the masking operation. The downstream fine-tuning is performed on ADE20K and COCO datasets, and state-of-the-art performance is claimed.

Yao et al. [29] proposed Wave-ViT, which is a wavelet decomposition-based vision transformer. They claimed that the downsampling operation performed through average pooling is an invertible process and results in information loss. They have used wavelet decomposition to perform downsampling, which is invertible and believed to cause less information loss during downsampling. They have performed pre-training on the ImageNet dataset with an input spatial resolution of 224×224 pixels. It is claimed that state-of-the-art performance is achieved by doing downstream fine-tuning on the ADE20K and COCC datasets.

B. SELF-SUPERVISED LEARNING

Yang et al. [30] proposed VISTA-Net, which used spatial and channel attention to perform visual representation learning using variational structures. Their method combines a probabilistic framework with a structured attention model to learn deep feature representation, providing rich channel and spatial interdependencies with effective performance on a variety of tasks.

C. UNSUPERVISED LEARNING

Unsupervised visual representation learning is a machine learning technique that entails learning representations of visual data without explicit supervision, i.e., without using labeled examples. Unsupervised visual representation learning seeks to find useful patterns and structures in data to be used for another task of similar or related nature [17]. Unsupervised visual representation learning uses generative models, such as variational autoencoders [31], [32], [33] and generative adversarial networks [34], [35], [36]. These models build synthetic examples of the input data by first creating a latent representation, which captures the underlying structure of the data. The latent representation can then be used as a representation of features for subsequent tasks.

Unsupervised representation learning has the potential to overcome data scarcity and lack of availability by revealing previously hidden patterns and clusters. Clustering-based methods for this purpose may use either k-means [37], [38],

[39] or Gaussian mixture models [40] to classify images with shared visual characteristics. The clusters may be used to create a group of visual prototypes that can serve as a feature representation for the subsequent tasks. However, evaluating the quality of learned representations is difficult because there is no objective metric that can be optimized for these types of tasks.

III. PROPOSED APPROACH

The proposed approach uses upstream training to create a pre-trained model using a sizable collection of artificially generated images. On one of the several IQA datasets containing authentic or synthetic distortions, downstream training can be performed with fewer hyperparameter adjustments, fewer training samples, faster convergence, and greater correlation with MOS. Figure 1 depicts the overall structure of the proposed technique, and this section provides specifics on each of its component parts.

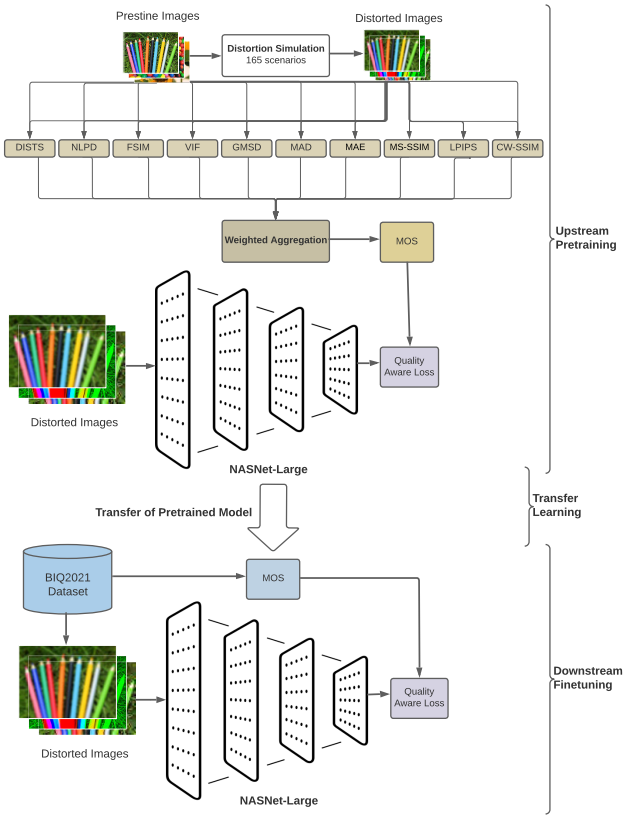


FIGURE 1. Overall framework of the proposed approach.

A. UPSTREAM PRE-TRAINING

It has been demonstrated [41] that larger and more complex CNN models are usually better at visual recognition tasks, and the same is the trend in the task of IQA [42]. It is apparent from deep learning research [13] that larger networks require massive amounts of annotated data to perform learning and obtain the benefits of deeper and more complex architectures. BiT [14] performed domain transfer and demonstrated that deeper architectures could perform better when trained on larger datasets than less complex architectures and vice versa. Therefore, to perform visual representation learning for image quality, we are required to have a deeper and more complex architecture along with a massive amount of annotated data that can be used for upstream training. Upstream training refers to the process of training a machine learning model for a similar but not identical task for which the model will be used. The concept is that the model may improve its performance and reduce the quantity of data required for training if it is first trained on a similar task and then fine-tuned to the target task.

1) ARCHITECTURE SELECTION

To perform upstream training, we have selected NasNet-Large [43] as it is sufficiently deeper and complex and has been shown to perform remarkably on IQA [42]. The architecture belongs to the NASNet family and is the largest variant with over 88 million parameters. The architecture is created via neural architecture search [44] and contains a series of blocks connected in a feedforward manner and is optimized for visual classification tasks. In order to reduce the number of parameters while preserving predictive performance, the architecture includes normal and reduction cells. The architecture contains repetitions of these cells along with skip connections to improve the information flow across layers and minimize the vanishing gradient problem. Figure 2 provides the architectural arrangements of normal (a) and reduction cells (b) used in the construction of NasNet-Large, which is used for upstream training. Let X be the input image. The NASNet applies a set of initial convolutional operations to the input image X, which can be represented as 1:

$$C_1 = \text{Conv1}(X) \tag{1}$$

where C_1 represents the output feature map after the initial convolutions. NASNet utilizes a series of normal and reduction cells to extract hierarchical features. Each cell consists of multiple operations that are selectively chosen during the architecture search process. Let's denote the output of the normal cell as $N = NormalCell(C_1)$ and the output of the reduction cell as $R = ReductionCell(N)$. The NormalCell takes the output feature map C_1 as input and performs a sequence of operations, which can be represented as 2:

$$N = f(N-1, N-2, \ldots, N-k) \tag{2}$$

where f represents the sequence of operations in the $NormalCell$, and $N-k$ represents the k_{th} intermediate feature map. Similarly, the $ReductionCell$ takes the normal cell output N as input and performs a series of operations to reduce spatial dimensions, which can be represented as 3:

$$R = g(R-1, R-2, \ldots, R-k) \tag{3}$$

where g represents the sequence of operations in the reduction cell, and $R-k$ represents the $k-th$ intermediate feature map. In a normal or reduction cell, the input of the cell, h_i,

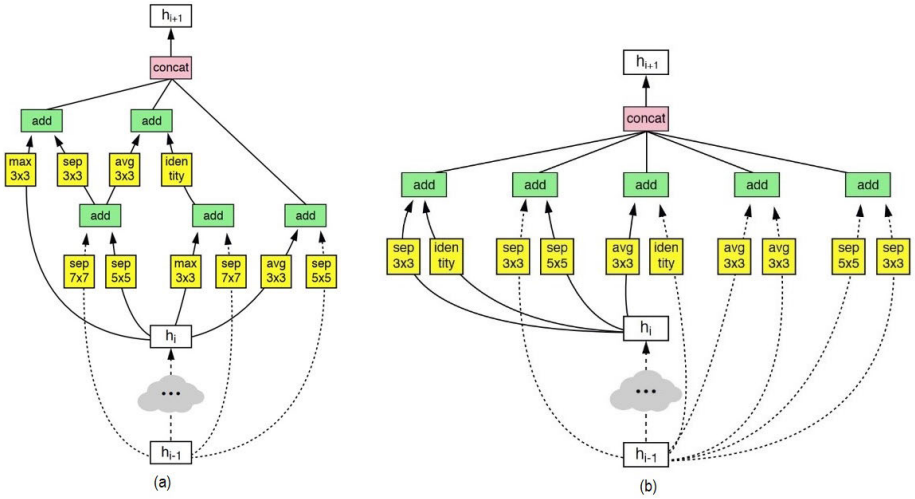


FIGURE 2. Architectures of the normal cell (a) and reduction cell (b), used in NasNet-Large.

is fed into the cell, and the results, h_i+1, are obtained through concatenation operations from all branches represented as Br. The cell operations can be represented as 4, which represents the different branches within the cell.

$$h_{i+1} = \text{Concatenate}(\text{Br}_1(h_i), \text{Br}_2(h_i), \ldots, \text{Br}_N(h_i)) \quad (4)$$

Within each branch, various operations are performed, including separable convolutions ''sep'', identity operations ''identity'', average-pooling ''avg'', and max-pooling operations ''max'' which are discussed further.

a: SEPARABLE CONVOLUTIONS

Let X be the input feature map with the dimensions $H \times W \times C$, where H stands for height, W for width, and C for input channel count. The depthwise convolution and the pointwise convolution processes make up the separable convolution. Each input channel is convolved with a different set of filters during the depthwise convolution step. Let $F_{depthwise}$ be the set of filters for depthwise convolution, which can be expressed as 5.

$$Z = DepthwiseConv(X, F_depthwise) \quad (5)$$

The output feature map is represented by Z, which is obtained after depthwise convolution operation which has the same spatial dimensions $(H \times W)$ but with C channels, as each input channel has been convolved independently. The resulting feature maps from the depthwise convolution are linearly combined using a $1times1$ convolution in the pointwise convolution stage. Let $F_{pointwise}$ be the set of filters for pointwise convolution, which can be represented as 6:

$$Y = PointwiseConv(Z, F_pointwise) \quad (6)$$

where Y represents the final output feature map of the pointwise convolution operation, which has a varied number of channels based on the number of filters in $F_pointwise$ but the same spatial dimensions $(H \times W)$. A separable convolution procedure involves employing pointwise convolution to combine the results after independently applying depthwise convolution to each input channel. This two-step process reduces the computational cost while capturing spatial and channel-wise information effectively.

b: POOLING OPERATIONS

The average pooling takes the average value inside each pooling window to minimize the spatial dimensions of the input feature map. Average pooling can be described mathematically as 7:

$$\text{avg\_pool}[i, j, c] = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} x[i, h, w, c] \quad (7)$$

In 7, avg_pool[i,j,k] represents the value of the output feature map at spatial location i, j and channel index c after average pooling. The double summation $\sum_{h=1}^{H}$ and $\sum_{w=1}^{W}$ represents the summation of the input feature map values within the pooling window, and dividing by $H \times W$ indicates that the total number of elements provides the average value within the pooling window.

Similarly, max pooling reduces the spatial dimensions by selecting the maximum value within each pooling window and can be expressed as in 8:

$$\text{max\_pool}[i, j, k] = \max_{h=1}^{H} \max_{w=1}^{W} x[i, h, w, c] \quad (8)$$

c: GLOBAL AVERAGE POOLING

In order to combine the spatial data and create a fixed-length feature vector, the cell's outputs, $h_i + 1$, can either be used as inputs for succeeding cells or fed to a global average pooling layer. The following is how the global average pooling

procedure is shown in 9:

$$\mathrm{GAP(c)} = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} x[h, w, c] \tag{9}$$

In 9, $GAP(c)$ represents the output value of the global average pooling operation for channel index c. The double summation $\sum_{h=1}^{H}$ and $\sum_{w=1}^{W}$ represents the summation of input feature map values over the spatial dimensions h and w, whereas the division by the entire number of feature map elements, given by $H \times W$, provides the average value over the entire spatial extent.

d: FULLY CONNECTED LAYERS

For each output index i and channel index k in the fully connected layer, the output value $dense[i, 1, 1, c]$ is computed using the ReLU activation function:

$$dense[i, 1, 1, c] = ReLU\left(\sum_{j=1}^{channels} W_{dense}[j, c] \cdot concat[i, 1, 1, j] + b_{dense}[c]\right) \tag{10}$$

For an input index of i and a channel index of k, $dense[i, 1, 1, c]$ represents the output of the fully connected layer in 10. The ReLU function is applied to the sum of the weighted inputs, where the weights $W_{dense}[j, c]$ connect input index j to output channel c. The input values are concatenated using $concat[i, 1, 1, j]$. Additionally, the bias term $b_{dense}[c]$ is added to the weighted sum before applying the ReLU activation function.

We adapted NasNet-Large's original architecture for regression problems even though it was created for image classification. The final fully connected layer is replaced by a regression layer in image quality evaluation since the model has to predict a continuous quality score rather than a class label. Additionally, the model must forecast a single quality score per image rather than class probabilities, so the dimension of the last fully connected layer is set to 1. Additionally, the loss function is discussed in the following part because it is crucial to model training and quality prediction.

2) LOSS FUNCTION

The Mean Squared Error (MSE) loss function has become the de facto standard for training quality evaluation models. The MSE outperforms other traditional loss functions for training IQA models, according to the empirical data provided by Ahmed et al. [7]. On the other hand, Hosu et al. [11] suggested that the Mean Absolute Error (MAE) is the best option for assessing image quality. Our research, however, favors the use of a quality-aware loss function in an effort to go beyond the traditional use of MSE or MAE alone. In order to do this, we developed a multi-objective loss function that incorporates MSE, MAE, and Spearman's Rank-Order Correlation Coefficient (SROCC).

$$\mathrm{Loss} = \mathrm{MSE} + \mathrm{MAE} + (1 - \mathrm{SROCC}) \tag{11}$$

The objective of 11 is to minimize this value, which represents the desired behavior, by framing the SROCC component as a loss term, particularly as $1 - SROCC$. The MSE component calculates the average squared difference between the quality scores predicted and those obtained from the ground truth. In order to capture subtle differences in quality, it penalizes larger errors more severely. The MAE component calculates the standard deviation of the absolute difference between the expected and actual quality ratings. It offers reliable error measurement and is less susceptible to outliers. The loss function captures the disparities between the predicted and ground truth scores in terms of squared and absolute differences, respectively. This is in accordance with recognized error metrics, which are supported by the inclusion of both MSE and MAE components. Specifically employing the Spearman Rank Correlation Coefficient, the term SROCC captures the relationship between the projected quality scores and the actual quality scores. When evaluating the monotonic relationship between the projected and actual rankings, this component takes the relative ranking of the quality scores into account. The loss function seeks to optimize the model by integrating these elements to reduce the squared and absolute disparities between the expected and ground truth quality scores and to promote a high correlation between the predicted ranking and the ground truth ranking.

3) PATH DROPPING

Path dropping is introduced as a regularization method that improves the predictive performance of the model and is implemented using ScheduledDropPath [43] that extends the concept of DropPath [45]. The DropPath algorithm stochastically drops a path in the cell with a fixed probability, whereas the ScheduledDropPath performs this dropping with a linearly increasing probability throughout training. For each training iteration I and for each layer L, the ScheduledDropPath operation can be defined as 12:

$$\mathrm{DroppedPath}[i, L] \sim \mathrm{Bernoulli}(P_{\mathrm{scheduled\_drop}}) \tag{12}$$

where $DroppedPath[i, l]$ is a binary random variable indicating whether the paths are dropped for iteration i and layer l. It follows a Bernoulli distribution with a drop probability of $P_{scheduled\_drop}$. During forward pass calculations, the paths that are not dropped $DroppedPath[i, L] = 0$ are scaled by the inverted drop probability, while the paths that are dropped $DroppedPath[i, L] = 1$ are set to zero. This scaling is necessary to maintain the expected value of the output during training. The output of the ScheduledDropPath operation can be defined as 13:

$$\mathrm{Output}[i, L] = \frac{\mathrm{DroppedPath}[i, L] \times \mathrm{Input}[i, L]}{1 - P_{\mathrm{scheduled\_drop}}} \tag{13}$$

where $Output[i, L]$ represents the output of the ScheduledDropPath operation for iteration i and layer L, and $Input[i, L]$ is the input to that layer. By incorporating ScheduledDropPath into the training process, different paths are stochastically dropped during each iteration, encouraging

the network to learn robust representations that are not overly dependent on specific paths. The gradual increase in the drop probability over iterations enables a controlled regularization effect. The use of path-dropping resulted in improved overall predictive performance.

B. TRAINING DATA DESCRIPTION

This section addresses the topic of the dataset that will be utilized for both upstream pre-training and downstream fine-tuning. As discussed, a large-scale dataset of annotated images is required to perform pre-training on the model. The details of such dataset acquisition and preparation are discussed further in this section. Two synthetic distortion datasets and two authentic distortion datasets are utilized to perform downstream fine-tuning, and the results are provided to allow for a comparison with existing approaches.

1) DATASET FOR UPSTREAM PRE-TRAINING

Acquisition of a large-scale annotated dataset for supervised classification is a major challenge that is faced by the research community while training complex deep-learning models. In the adopted approach, we have performed distorted image generation by simulating distortion models on a set of pristine-quality images.

a: GENERATION OF DISTORTED IMAGES

To create a large-scale annotated image dataset, the study used the Kadis-700K [9] dataset's pristine images (denoted by the letter P). For generating distorted images, we employed a distortion generation function defined as follows:

$$\mathrm{Img}_D = f(\mathrm{Img}_P, D_d, L_i) \tag{14}$$

Here in 14, Img_P represents a unique image selected from the pristine dataset P, D_d represents a specific distortion chosen from the set of distortions discussed in III-B2. The function f takes these inputs and produces a distorted image, Img_D. Considering that the number of pristine images P is 0.15 million, the number of distorted images can be calculated using the 15:

$$|\mathrm{Img}_D| = |\mathrm{P}| \cdot |\mathrm{D}| \cdot |\mathrm{L}| \tag{15}$$

In this case, $|P| = 0.15$ million, $|D| = 33$, and $|L| = 5$, hence, $|Img_D| = 0.15 \times 33 \times 5 = 24.75$ million. The type of distortions is synthetically generated, and the code for these distortions is provided at our GitHub repository, which takes an image, distortion number, and the level of distortion intensity as input arguments, and the function returns a distorted image.

The pristine images of the dataset are subjected to the distortion simulation models in order to produce distorted images. A granular degree of control over the intensity of the distortion is possible due to the simulation of each distortion at five different levels of severity. For instance, the compression ratio for JPEG compression can be changed between five distinct classes.

The final collection consists of 24.75 million distorted images and 0.15 million pristine images as a result of this distortion simulation procedure. The technique of distortion simulation is shown in Figure 3, which comprises creating the relevant distorted images from the original, unaltered images taken from the Kadis-700K database. This distorted image collection is known as the VRL-IQA database, and it is used to perform pre-training.

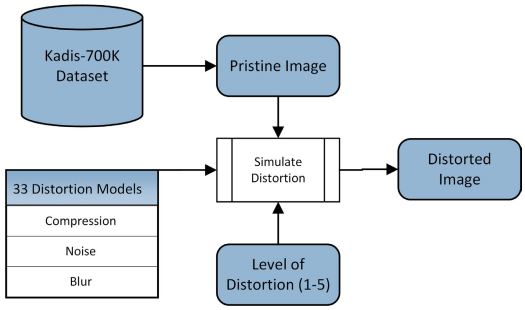


FIGURE 3. Process of distortion simulation.

2) DISTORTION SIMULATION MODELS

The distortion simulation models play a pivotal role in generating distorted images for quality assessment. Each of these models is designed to replicate specific types of distortions commonly encountered in real-world scenarios. The selection of these distortions is guided by their prevalence in digital multimedia content and their impact on perceived image quality. Table 1 provides a comprehensive description of each distortion, categorized by its nature and characteristics. These simulations are integral to creating a diverse and extensive dataset for training and evaluating our IQA model. Each distortion is meticulously designed to represent a distinct facet of image degradation, ensuring a thorough assessment of the model's performance across various quality-altering factors. These simulations collectively contribute to the robustness and effectiveness of our proposed approach to tackling the challenges of IQA in real-world multimedia applications. The distortion models used for the generation of distortions are listed in Table 1.

3) SCORING THE DISTORTED IMAGES

Subjective evaluation of the perceived quality of distorted images by laboratory tests can be an expensive process. The ITU-R BT.500-11 [46] standard recommends using an absolute category rating with a set number of distinct scales for subjective evaluation. It is advised to include at least ten volunteers in the subjective assessment, while more than thirty participants from a variety of backgrounds can yield a more trustworthy result. Given the dataset's 24.75 million images, 742.5 million ratings would be necessary to reach a minimum of 30 volunteer ratings for each image. It is neither practical nor possible to accommodate such a high number of ratings within realistic constraints.

TABLE 1. Description of distortions used to produce distorted images.

Sr.	Distortion	Category	Description
1	JPEG	Compression	Standard JPEG compression
2	JPEG2000	Compression	Standard JPEG2000 compression
3	White noise	Noise	RGB image with Gaussian white noise
4	Chrominance Noise	Noise	YCbCr-converted image with Gaussian white noise
5	Monochromatic Noise	Noise	Introduce Gaussian white noise to (Y) luminance layer of the YCbCr image
6	Impulse noise	Noise	Add salt and pepper noise to the RGB image
7	Multiplicative noise	Noise	Add speckle noise to the RGB image
8	Denoise 1	Noise	Apply a denoising DnCNN to the white noise image
9	Denoise 2	Noise	Apply a denoising DnCNN to white noise in the color component image
10	Denoise 3	Noise	Apply a denoising DnCNN to monochromatic image
11	Denoise 4	Noise	Apply a denoising DnCNN to impulse noise image
12	Denoise 5	Noise	Apply a denoising DnCNN to multiplicative noise image
13	Gaussian blur	Blurs	Introduce blur using Gaussian kernel filtering
14	Lens blur	Blurs	Introduce blur using circular kernel filtering
15	Motion blur	Blurs	Introduce blur using line kernel filtering
16	Brighten	Brightness	Adjust the brightness channel in a nonlinear manner while maintaining extreme values
17	Darken	Brightness	Darken the brightness channel while maintaining extreme values
18	Mean shift	Brightness	Add a constant to each value in the image, then trim the values to the original range
19	Unsharp Mask	Sharpness	Perform unsharp masking to sharpen the image features
20	Contrast change	Contrast	Use a Sigmoid-type adjustment curve to modify RGB values nonlinearly
21	Contrast Stretching	Contrast	Perform contrast stretching of RGB image
22	Jitter	Spatial	Create a random distribution of visual data by bending each pixel with arbitrary small offsets
23	Non-eccentricity patch	Spatial	Small areas in the image are erratically offset to surrounding locations
24	Pixelate	Spatial	Using nearest-neighbor interpolation in each instance, reduce the image's size and increase it back to its original size
25	Quantization	Spatial	Image values are quantized using N thresholds determined by Otsu's approach
26	Color block	Spatial	Add homogenous random colored blocks to the image in different locations
27	Color diffusion	Color	Simulate Gaussian blur on the color channels of the Lab color-space
28	Color shift	Color	Translate the green channel at random, then include it into the background of the original image using a gray-level map
29	Color quantization	Color	Employ minimum variance quantization and dithering to convert to an indexed image
30	Color saturation 1	Color	In the HSV color space, amplify the saturation channel by a certain amount
31	Color saturation 2	Color	In the Lab color space, amplify the color channels by a certain amount
32	JPEG Compression of Noisy Image	Multiple	Add additive white noise and then compress via standard JPEG compression
33	JPEG2000 Compression of Noisy Image	Multiple	Add additive white noise and then compress via standard JPEG2000 compression

To address the challenge of subjective scoring, we have adopted a synthetic scoring approach, which provides a reasonably reliable quality score. This allows us to perform pre-training in a weakly supervised manner, considering that the quality ratings obtained may be somewhat noisy.

To generate synthetic quality scores, we've developed a powerful method that draws upon full-reference quality evaluation techniques. To determine the quality score for each image, we conducted a comprehensive literature survey and experimental evaluation using the Kadid-10K dataset [9]. Through this process, we identified 10 full-reference methods that can be used to obtain quality scores.

Table 2 offers information on the chosen full-reference quality assessment algorithms, including the year of publication and the weightage assigned to each of them based on their predictive performance. This weighted approach allows us to derive a synthetic quality score for each image, which serves as a valuable resource for our research. The performance assessment scores of the 10 full-reference methods are combined using a weighted average approach based on linear regression. We train a multiple linear regression model on 80% of the Kadid-10K dataset and evaluate its performance on the remaining 20% of the dataset.

Let

$$W = [0.0324, 0.0398, 0.0549, 0.0643, 0.0951, 0.1878, 0.1920, 0.1085, 0.1097, 0.1153] \quad (16)$$

be the weights vector that each of the 10 full-reference methods has been given, and

$$S = [S1, S2, S3, S4, S5, S6, S7, S8, S9, S10] \quad (17)$$

be the vector of quality assessment scores obtained from the 10 full-reference methods. The quality score Q can be calculated as the weighted average using 18:

$$Q = f(W, S) \quad (18)$$

where f denotes the dot product of the score vector S and the weight vector W. The quality score Q obtained from 18 is used as a label for weakly supervised learning.

Figure 4 illustrates the process used to generate synthetic scores and label the distorted images. This process involves training the multiple linear regression model on a subset of the

TABLE 2. Full-reference methods and their assigned weightage.

Sr.	IQA Method	Year	Weightage
1	GMSD [47]	2013	0.0324
2	FSIM [48]	2011	0.0398
3	VIF [49]	2006	0.0549
4	CW-SSIM [50]	2005	0.0643
5	NLPD [51]	2016	0.0951
6	MAE [52]	2016	0.1878
7	MS-SSIM [3]	2003	0.1920
8	MAD [53]	2010	0.1085
9	DISTS [54]	2020	0.1097
10	LPIPS [55]	2018	0.1153

dataset and utilizing the predictions from the 10 full-reference methods to obtain the output prediction. This synthetic score generation technique is a key step in our methodology and is crucial for training our model in a weakly supervised manner.

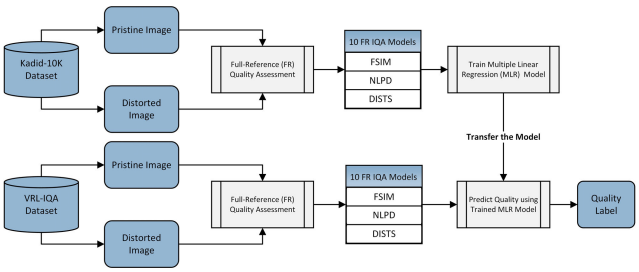


FIGURE 4. Process flow of synthetic score generation to label distorted images.

This stage has returned 24.7 million images with quality labels using a weighted average of 10 good-performing full-reference quality assessment algorithms. The pre-training of the NASNet-Large model is performed using this dataset, and the process of pre-training is explained further.

a: PRE-TRAINING OF NASNET-LARGE

The pre-training of the NASNet-Large model is performed by using the generated VRL-IQA dataset. As the objective of the trained model is to correlate well with human judgment, therefore, the loss function used in this study incorporates error as well as correlation terms.

b: DATA AUGMENTATION

The training of deep neural networks frequently makes use of data augmentation. In order to expand the amount of data and add perturbations that are not present in the data but may frequently occur in the testing images, controlled perturbations are introduced into the images. The perceived quality of the image may be compromised by some of the perturbations used for visual recognition, so it is vital to keep this in mind when training IQA models. As a result, the only image enhancement techniques used in this study are horizontal flips, translation, rotation, and random cropping.

TABLE 3. Training hyperparameters.

Sr.	Hypermeter	Value
1	Optimizer	Adam
2	Initial Learning Rate	5x10-3
3	Learn Rate Schedule	Piecewise
4	Learn Rate Drop Period	20
5	Learn Rate Drop Ratio	0.5
6	Batch Size	16
7	Epochs	100
8	Validation Patience	3 epochs

c: TRAINING HYPERPARAMETERS

The process of selecting optimal learning parameters is known as hyperparameter tuning. These settings, which are made by the user and may have an impact on the model's performance and training outcomes, are not learned from the training data. Table 3 contains the training parameters that were used to train the model. It should be noted that although the maximum epochs are initially set at a high value, overfitting is prevented by using a validation check. The training is stopped, and the weights are kept as the final model if the loss of the model stops reducing for three epochs. Additionally, a piecewise learning rate scheduler is employed, and the starting learning rate is set to 5×10^{-3}. After 20 epochs, the learning rate is cut in half, allowing the model to converge to an optimal point and prevent destabilization.

C. DOWNSTREAM FINE-TUNING

As the model pre-training is performed in a weakly supervised manner by using distorted images labeled with synthetic scores, the fine-tuning of the model weights on benchmark datasets is performed for evaluation. To perform model fine-tuning, we have used two benchmark datasets containing synthetic distortions and two datasets containing authentic distortions.

1) DATASET FOR DOWNSTREAM FINE-TUNING

Various datasets are used for benchmarking IQA, but there are two distinct classes of these datasets. The datasets designed for full-reference IQA contain images distorted using simulation and therefore contain pristine as well as distorted images. Two popular datasets in this category are TID2013 [10], and Kadid-10K [9], which are described in Table 4. These are large datasets with synthetic distortion and are publicly available for train-test evaluation. The other category of the dataset contains authentically distorted images in which the distortion is not simulated, but the images with distortion are chosen to constitute the dataset. The two largest and latest releases of datasets in this category are KonIQ-10K [11], and BIQ2021 [12] containing 10,073 and 12,000 images, respectively, and are described in Table 4. To evaluate downstream performance, we have used four datasets. These datasets differ in the type of distortions, image resolution, nature of the content, number of images, and way of quality scoring. The MOS or DMOS of these

TABLE 4. Dataset for downstream evaluation.

Sr.	Dataset	Year	Distortion Type	Reference Information	Resolution	No. of Images
1	TID2013 [10]	2013	Synthetic	Available	512 × 384	3,000
2	Kadid-10K [9]	2019	Synthetic	Available	512 × 384	10,125
3	KonIQ-10K [11]	2020	Authentic	Not available	512 × 384	10,073
4	BIQ2021 [12]	2022	Authentic	Not available	512 × 512	12,000

datasets is rescaled to a range of 0-1, and the image size is not rescaled. With the exception of the learning rate and the schedule, the pre-trained model is fine-tuned on each of these datasets using the identical training hyperparameters as shown in Table 3. The drop factor is set at 0.5, the drop period is 10 epochs, and the learning rate is 3×10^{-4}. To avoid overfitting, the training is stopped when the loss stops reducing further. Additionally, since the model's input image size is different and lower than the image sizes offered in the dataset, random cropping is used during training instead of resizing. Resizing is observed to affect the perceptual quality of images, whereas random cropping provides a different portion of the image to model and acts as a regularization method [7]. Since an image's quality is constant throughout the image while its content varies, this enables the model to learn quality-related representations rather than the image's actual content.

IV. RESULTS & DISCUSSION

This section describes the experimental evaluation of transfer learning performance for four IQA datasets, which are listed in Table 5. The trials are carried out independently for each dataset, and comparisons are made for both synthetic and authentic distortion datasets. The execution environment for these trials and the evaluation metrics used for performance quantification are discussed further in section IV-A and IV-B.

A. EXECUTION ENVIRONMENT

The experiments are conducted using a Dell Precision T3610 workstation with Intel Xeon E52687 v2 with 32 GB of RAM and NVIDIA GeForce RTX 3060 with 12 GB GRRD6 memory. The workstation was operating with Windows 10 Pro 64-bit and MATLAB®2022b for implementation. An onboard SATA SSD with 512GB of storage space is used for the operating system, MATLAB, and dataset to minimize latency and speed up computations.

B. EVALUATION METRICS

To rigorously evaluate the performance of our models, it is imperative to establish robust metrics. The choice of an evaluation metric is contingent upon the specific task requirements and the inherent characteristics of the dataset under consideration. In the context of no-reference IQA, the primary objective is to predict image quality with a high degree of correlation to human judgments. To quantify this correlation, we have employed two widely accepted measures: the Pearson Linear Correlation Coefficient (PLCC) and the Spearman's Rank Order Correlation Coefficient (SROCC). These metrics are particularly well-suited for assessing perceptual IQA.

The PLCC (Equation 19) measures the linear relationship between predicted quality scores ($\hat{y}_P$) and ground truth scores (y_P). It essentially quantifies how well the predicted and actual quality scores align in terms of a linear correlation.

The SROCC (Equation 20) is based on the ranks of the quality scores rather than their actual values. It evaluates the monotonic relationship between the predicted rank order ($\hat{y}_S$) and the ground truth rank order (y_S). In essence, it assesses how well the predicted scores maintain their relative ordering with respect to the ground truth scores.

These two metrics provide a comprehensive evaluation of the model's ability to predict image quality in line with human perceptions, covering both linear and monotonic aspects of correlation. They serve as robust indicators of the model's performance in the domain of perceptual IQA.

$$\text{PLCC} = \frac{\sum_i (y_P - \hat{y_P})(y_S - \hat{y_S})}{\sqrt{\sum_i (y_P - \hat{y_P})^2 \sum_i (y_S - \hat{y_S})^2}} \tag{19}$$

$$\text{SROCC} = 1 - \frac{6 \sum_n d_n^2}{N(N^2 - 1)} \tag{20}$$

C. EVALUATION ON TID2013

The TID2013 dataset [10] is a commonly utilized benchmark for IQA evaluation. It comprises 3,000 distorted images, each with a corresponding high-quality reference image. These reference images serve as exemplars of superior perceptual quality, while the distorted images have been intentionally manipulated using a variety of distortion simulations, including noise, blur, compression, and various forms of artifacts. Figure 5 showcases a selection of sample images from this dataset.

For the evaluation process, the dataset underwent subjective scoring involving human observers who rated the perceived quality of the distorted images on a scale ranging from 0 to 9. The resulting data was then used to generate a histogram depicting the distribution of image quality ratings for all 3,000 images, as shown in Figure 6. This thorough evaluation process ensures a comprehensive assessment of the IQA capabilities of the proposed model on a diverse range of image distortions.

The quality ratings of the images are scaled to a range of 0 to 9, and the dataset is divided into a train-test split in order to fine-tune the VRL-IQA model. The data is partitioned into an 80/20 split to perform testing on the 20% holdout dataset. The pre-trained VRL-IQA model is fine-tuned using 2400 images and the training settings provided

FIGURE 5. Nine sample images from TID2013 dataset.

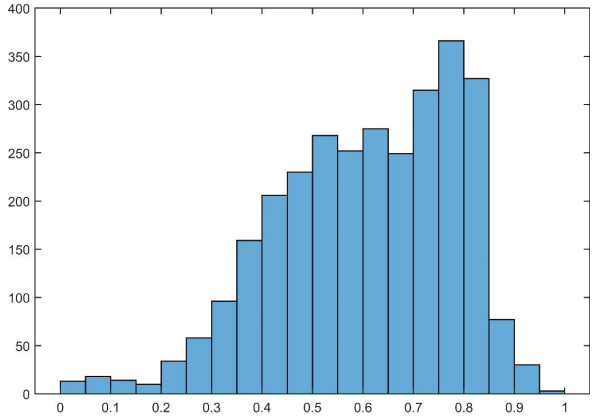


FIGURE 6. MOS distribution of TID2013 dataset.

in Section III-C. The pre-trained model's input image size is 331 × 331, which is not the same as the image sizes in the dataset (Table 4). The final testing score is an average of ten crops, with random cropping used for both training and testing. The RMSE, PLCC, and SROCC of the fine-tuned model on the test set (600 images) of the TID2013 dataset are shown in Table 5.

D. EVALUATION ON KADID-10K

The Kadid-10K dataset, publicly released by the VQA Group at Universität Konstanz [9], is an extensive collection of images that have been synthetically distorted. This dataset encompasses 25 distinct types of distortions and was derived from 81 original, pristine images. A selection of sample images from this dataset is illustrated in Figure 7. In total, the dataset comprises 10,125 distorted images, each of which has been evaluated through subjective scoring using a pairwise comparison on a scale ranging from 1 to 5.

Unlike the TID2013 dataset [10], the subjective scoring for Kadid-10K was conducted through a crowdsourcing experiment. This approach ensured that the reliability and consistency of the ratings were rigorously assessed through carefully designed qualification tests. The distribution of subjective quality assessment scores is presented in the form of a histogram in Figure 8. This dataset provides a rich resource for evaluating and benchmarking IQA models on a diverse range of synthetic distortions.

FIGURE 7. Nine sample images from the Kadid-10K dataset.

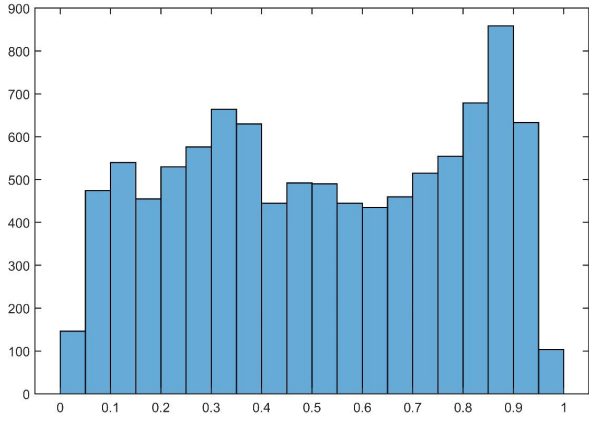


FIGURE 8. MOS distribution of Kadid-10K dataset.

The fine-tuning of the VRL-IQA model is performed by rescaling the quality scores to a range of 0 to 1. Additionally, the dataset is divided into an 80/20 train-test set, with 8,100 images utilized for model training and the remaining 2,025 images used for model testing. Additionally, cropping and fine-tuning are applied similarly to TID2013. The PLCC and SROCC of the fine-tuned model on the test set (2,025 images) of the Kadid-10K dataset are shown in Table 5.

E. EVALUATION ON KONIQ-10K

The KonIQ-10K dataset, generously made available by Universität Konstanz's VQA Group [11], features a collection of images that encompass authentic, real-world distortions. This dataset offers images in two spatial resolutions: 1024 × 768 and 512 × 384 pixels. Figure 9 showcases a selection of sample images from this dataset. In total, the dataset comprises 10,073 images, each of which underwent subjective scoring through online crowdsourcing experiments.

Remarkably, the dataset gathered evaluations from 1,459 distinct crowd workers, resulting in 1.2 million quality ratings. This extensive and diverse feedback enriches the dataset with a robust and comprehensive set of quality assessments. Figure 10 provides a visual representation of the distribution of quality scores in the form of a histogram. The KonIQ-10K dataset stands as a valuable resource for evaluating and benchmarking IQA models under real-world, authentic distortions.

FIGURE 9. Nine sample images from KonIQ-10K dataset.

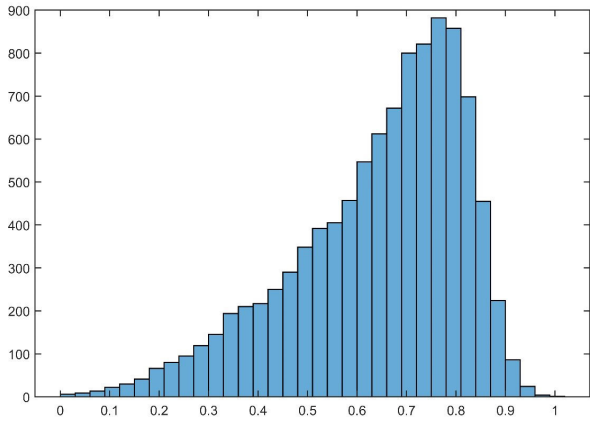


FIGURE 10. MOS distribution of KonIQ-10K dataset.

The fine-tuning of the VRL-IQA model is performed by rescaling the quality scores to a range of 0 to 1. In addition, the dataset is split into a train-test set of 80% and 20%, meaning that 8,058 images are utilized for model training while the remaining 2,015 images are used for testing. Similar to other datasets, images are adjusted and cropped. The PLCC and SROCC of the fine-tuned model on the test set (2,015 images) of the KonIQ-10K dataset are shown in Table 5.

F. EVALUATION ON BIQ2021

BIQ2021 [12] stands as a recent and comprehensive addition to the landscape of image datasets, comprising authentically distorted images. This dataset encompasses a substantial collection of 12,000 distorted images, meticulously curated with a keen emphasis on content diversity, quality assessment, and the authenticity of distortions. Figure 11 offers a visual glimpse of select images from this dataset, showcasing the breadth of distortions captured.

Notably, the assessment process for this dataset differs from that of KonIQ-10K [11] and Kadid-10K [9]. In the case of BIQ2021, images were subjectively scored in a controlled laboratory setting, providing a distinct evaluation environment. Each image in the dataset underwent evaluation for perceptual quality by 30 unique observers, resulting in an impressive aggregate of 0.36 million individual quality ratings. This extensive feedback ensures a robust and detailed assessment of image quality. Figure 12 presents the distribution of quality scores in the form of a histogram, offering a visual representation of the dataset's comprehensive quality assessment. BIQ2021, with its emphasis on authenticity and meticulous evaluation, serves as a valuable resource for advancing IQA in the domain of authentically distorted images.

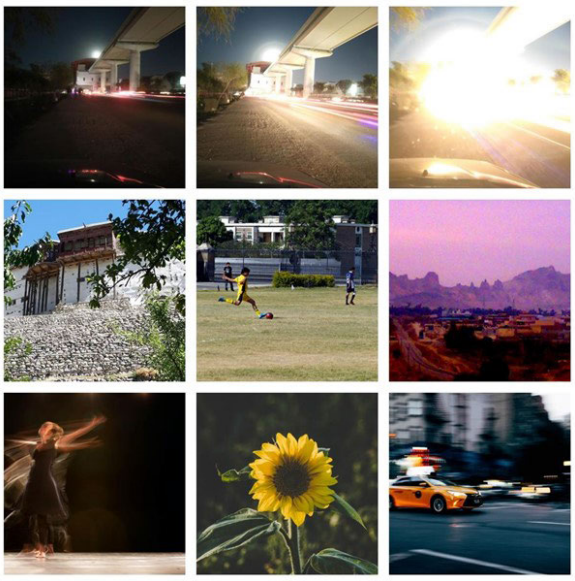


FIGURE 11. Nine sample images from BIQ2021 dataset.

To perform fine-tuning, the rescaling of the images is not required as the dataset provides the MOS, which is already scaled in the range of 0 to 1. The BIQ2021 provides the train-test split of the data, which is used by the author and other researchers to report the performance, and therefore the same train-test split is used for our purpose. The model is trained using the train set's 10,000 photos, while the test set's 2,000 images are used to validate the model's performance. Table 5 contains a report on the performance of the model after it has been fine-tuned.

G. COMPARISON WITH IMAGENET PRETRAINED MODELS

To assess the efficacy of weakly supervised pre-training to perform visual representation learning, this section presents

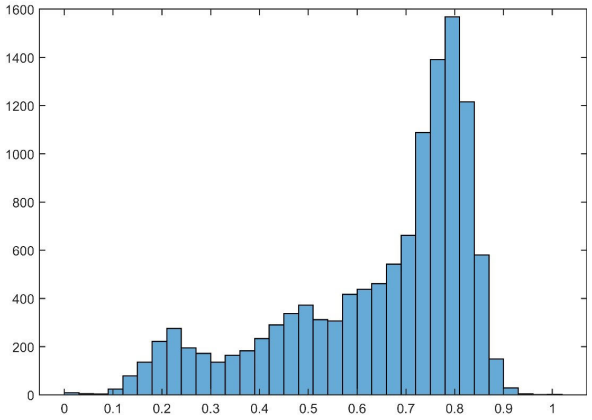


FIGURE 12. MOS distribution of BIQ2021 dataset.

TABLE 5. Performance of the VRL-IQA model on four benchmark datasets.

Sr.	Dataset	Distortion Type	PLCC	SROCC
1	TID2013	Synthetic	0.913	0.921
2	Kadid-10K	Synthetic	0.878	0.893
3	KonIQ-10K	Authentic	0.887	0.884
4	BIQ2021	Authentic	0.786	0.793

TABLE 6. Performance Comparison for ImageNet pre-trained model and VRL-IQA model.

		ImageNet			VRL-IQA		
Sr.	Dataset	PLCC	SROCC	Eps	PLCC	SROCC	Eps
1	TID2013	0.903	0.896	40	0.913	0.921	21
2	Kadid-10K	0.871	0.874	51	0.878	0.893	27
3	KonIQ-10K	0.790	0.785	62	0.887	0.884	41
4	BIQ2021	0.778	0.750	59	0.786	0.793	38

the experimental results of transfer learning performance. The experiments are conducted to perform fine-tuning of the ImageNet pre-trained model and VRL-IQA model, which is trained via weakly supervised learning. The fine-tuning of each of these models is performed on four benchmark datasets to establish the superiority of the fine-tuning performance using the proposed approach. The outcomes of the two modeling methodologies are presented in Table 6 in terms of PLCC, SROCC, and the number of Epochs required till validation criteria are met. The conditions for validation criteria are listed in Table 3.

In contrast to the pre-trained ImageNet model, the experimental findings show a higher correlation between the predicted quality score and the ground truth for VRL-IQA. Furthermore, compared to the pre-trained ImageNet model, the number of epochs needed to do fine-tuning until validation conditions are met is much lower for VRL-IQA. These findings show that the proposed framework is particularly effective in teaching visual representations for IQA and may be applied to increase prediction performance in terms of correlation with fewer training epochs. Additionally, the VRL-IQA is fine-tuned using two synthetic distortion datasets and two authentic distortion datasets, showing that even though the model was initially trained on synthetic distortion datasets, it excels when used to fine-tune authentic distortion datasets.

TABLE 7. Comparison of the proposed approach with 22 existing approaches using SROCC.

Sr.	Method	TID2013	Kadid-10K	KonIQ-10K	BIQ2021
1	NIQE [56]	0.263	0.338	0.4	0.356
2	GWH-GLBP [57]	0.315	0.285	0.698	0.602
3	BIQI [58]	0.468	0.294	0.662	0.564
4	Rb BRISQUE [56]	0.487	0.301	0.668	0.605
5	PIQE [59]	0.491	0.237	0.246	0.213
6	IL-NIQE [60]	0.516	0.63	0.447	0.461
7	BLIINDS-II [61]	0.521	0.534	0.575	0.496
8	DIIVINE [62]	0.521	0.436	0.693	0.617
9	CurveletQA [63]	0.56	0.442	0.718	0.63
10	OG-IQA [64]	0.564	0.447	0.635	0.371
11	BRISQUE [65]	0.565	0.398	0.677	0.603
12	GM-LOG-BIQA [66]	0.596	0.57	0.696	0.617
13	ENIQA [67]	0.596	0.641	0.745	0.634
14	SSEQ [63]	0.615	0.434	0.572	0.528
15	BMPRI [68]	0.692	0.534	0.619	0.494
16	NBIQA [69]	0.695	0.615	0.749	0.642
17	SGL-IQA [70]	0.713	0.774	0.794	0.71
18	PIQI [71]	0.818	0.893	0.824	-
19	RAN4IQA [72]	0.82	-	0.763	-
20	DB-CNN [73]	0.865	0.801	0.875	-
21	AIGQA [74]	0.871	0.864	0.766	-
22	DeepEns [7]	0.891	0.884	0.864	-
23	NASNet-Large [7]	0.896	0.874	0.785	0.75
24	VRL-IQA	**0.921**	**0.893**	**0.884**	**0.793**

TABLE 8. Comparison of the proposed approach with 22 existing approaches using PLCC.

Sr.	Methods	TID2013	Kadid-10K	KonIQ-10K	BIQ2021
1	NIQE [56]	0.277	0.302	0.319	0.301
2	GWH-GLBP [57]	0.296	0.302	0.688	0.644
3	BIQI [58]	0.315	0.375	0.718	0.683
4	Rb BRISQUE [56]	0.357	0.302	0.723	0.664
5	PIQE [59]	0.364	0.289	0.208	0.255
6	IL-NIQE [60]	0.411	0.426	0.707	0.694
7	BLIINDS-II [61]	0.452	0.527	0.652	0.403
8	DIIVINE [62]	0.456	0.588	0.463	0.541
9	CurveletQA [63]	0.471	0.471	0.73	0.698
10	OG-IQA [64]	0.487	0.429	0.709	0.684
11	BRISQUE [65]	0.49	0.553	0.574	0.555
12	GM-LOG-BIQA [66]	0.52	0.454	0.589	0.603
13	ENIQA [67]	0.545	0.637	0.761	0.703
14	SSEQ [63]	0.583	0.555	0.637	0.633
15	BMPRI [68]	0.627	0.59	0.705	0.699
16	NBIQA [69]	0.628	0.646	0.771	0.718
17	SGL-IQA [70]	0.808	0.864	0.874	-
18	PIQI [71]	0.815	0.782	0.815	0.77
19	RAN4IQA [72]	0.82	-	0.752	-
20	DB-CNN [73]	0.854	-	-	-
21	AIGQA [74]	0.865	0.806	0.868	-
22	DeepEns [7]	0.893	0.863	0.773	-
23	NASNet-Large [7]	0.903	0.871	0.79	0.77
24	VRL-IQA	**0.913**	**0.878**	**0.887**	**0.786**

H. COMPARISON WITH EXISTING SCHEMES

In the results section, the proposed VRL-IQA method is subjected to evaluation using distinct training and testing sets to assess its predictive capability. The assessment is based on key metrics, including PLCC and SROCC, which serve as benchmarks for predictive accuracy. The comparison highlights the superior performance of the proposed strategy when compared to existing approaches. Table 7 provides

a comprehensive overview of the VRL-IQA method's performance alongside 22 existing approaches, focusing on SROCC. Similarly, Table 8 presents results based on PLCC, offering a detailed evaluation of the proposed VRL-IQA method.

The proposed approach is superior to existing approaches in terms of PLCC and SROCC, according to a comparison with 22 alternatives. Moreover, the ImageNet pre-trained NASNet-Large model has provided a higher correlation between predicted scores and ground truth in comparison to existing approaches. These outcomes indicate that a larger and more complex model can serve as a suitable choice for predictive modeling to perform IQA. Moreover, model pre-training using a quality-aware dataset can minimize the number of epochs required to fine-tune the model and increase the correlation of the predicted quality score.

V. CONCLUSION

The assessment of distortion in generic image quality without reference information presents a significant challenge that has piqued the interest of the research community. Deep learning-based solutions had the highest correlation between predicted and actual quality scores. These solutions, however, have limitations due to the scarcity of annotated data. The need for diverse images with varying content, distortion severity, and the complexity of obtaining mean opinion scores from subjective IQA experiments make gathering a large amount of annotated data difficult. To address these issues, many existing studies have relied on ImageNet data for model pre-training, which is a poor solution. In our study, we propose a pre-training strategy that involves simulating 165 distortion scenarios on 0.7 million pristine-quality images to generate distorted images. We chose ten full-reference models to predict the quality of 24.75 million distorted images because full-reference quality assessment is a well-established field with consistent image quality predictions. We obtain average-quality scores that serve as the ground truth for pre-training by creating an ensemble that provides a weighted average of these models.

This study used the NASNet-large, a larger and more complex CNN architecture, for pre-training upstream data and fine-tuning downstream data to effectively learn representations from large-scale datasets. This study used four benchmark datasets for downstream fine-tuning and model evaluation: TID2013 and Kadid-10K, both synthetic distortion datasets, as well as KonIQ-10K and BIQ2021, both synthetic distortion datasets. We introduced a quality-aware loss function with an adjusted correlation term to improve robustness and correlation with human judgment. This loss function is used during training to improve the model's performance. Then it demonstrated the effectiveness of the proposed technique by achieving impressive prediction performance on benchmark datasets. The quality-aware pre-training allows the model to fine-tune new IQA datasets with fewer epochs and higher prediction accuracy. The proposed VRL-IQA model produces excellent Spearman's correlation values of 0.921, 0.893, 0.884, and 0.793 in our experiments for the TID2013, Kadid-10K, KonIQ-10K, and BIQ2021 datasets, respectively. These findings demonstrate the effectiveness of our novel pre-training strategy, which combines full-reference models, an ensemble approach, and a quality-aware loss function. Our findings make an important contribution to the field of IQA by presenting a promising approach for robust and accurate quality prediction in a variety of applications.

REFERENCES

[1] M. K. Mandal and M. K. Mandal, "The human visual system and perception," *Multimedia Signals Syst.*, pp. 33–56, 2003.

[2] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[3] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, 2003, pp. 1398–1402.

[4] C. Li and A. C. Bovik, "Three-component weighted structural similarity index," *Proc. SPIE*, vol. 7242, pp. 252–260, Mar. 2009.

[5] M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik, and M. K. Markey, "Complex wavelet structural similarity: A new image similarity index," *IEEE Trans. Image Process.*, vol. 18, no. 11, pp. 2385–2401, Nov. 2009.

[6] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, May 2011.

[7] N. Ahmed, H. M. Shahzad Asif, A. R. Bhatti, and A. Khan, "Deep ensembling for perceptual image quality assessment," *Soft Comput.*, vol. 26, no. 16, pp. 7601–7622, Aug. 2022.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[9] H. Lin, V. Hosu, and D. Saupe, "KADID-10k: A large-scale artificially distorted IQA database," in *Proc. 11th Int. Conf. Quality Multimedia Exper. (QoMEX)*, Jun. 2019, pp. 1–3.

[10] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. Jay Kuo, "Image database TID2013: Peculiarities, results and perspectives," *Signal Process., Image Commun.*, vol. 30, pp. 57–77, Jan. 2015.

[11] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Trans. Image Process.*, vol. 29, pp. 4041–4056, 2020.

[12] N. Ahmed and S. Asif, "BIQ2021: A large-scale blind image quality assessment database," *J. Electron. Imag.*, vol. 31, no. 5, Sep. 2022, Art. no. 053010.

[13] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor, "ImageNet-21K pretraining for the masses," 2021, *arXiv:2104.10972*.

[14] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby, "Big transfer (bit): General visual representation learning," in *Proc. 16th Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Aug. 2020, pp. 491–507.

[15] E. Cole, X. Yang, K. Wilber, O. M. Aodha, and S. Belongie, "When does contrastive visual representation learning work?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 01–10.

[16] X. Wang, K. He, and A. Gupta, "Transitive invariance for self-supervised visual representation learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1338–1347.

[17] A. Kolesnikov, X. Zhai, and L. Beyer, "Revisiting self-supervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1920–1929.

[18] X. Zhai, J. Puigcerver, A. Kolesnikov, P. Ruyssen, C. Riquelme, M. Lucic, J. Djolonga, A. Susano Pinto, M. Neumann, A. Dosovitskiy, L. Beyer, O. Bachem, M. Tschannen, M. Michalski, O. Bousquet, S. Gelly, and N. Houlsby, "A large-scale study of representation learning with the visual task adaptation benchmark," 2019, *arXiv:1910.04867*.

[19] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1717–1724.

[20] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, "End-to-end blind image quality assessment using deep neural networks," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1202–1213, Mar. 2018.

[21] S. Yang, Q. Jiang, W. Lin, and Y. Wang, "SGDNet: An end-to-end saliency-guided deep neural network for no-reference image quality assessment," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1383–1391.

[22] J. Wu, J. Ma, F. Liang, W. Dong, G. Shi, and W. Lin, "End-to-end blind image quality prediction with cascaded deep neural network," *IEEE Trans. Image Process.*, vol. 29, pp. 7414–7426, 2020.

[23] N. Ahmed and H. M. S. Asif, "Ensembling convolutional neural networks for perceptual image quality assessment," in *Proc. 13th Int. Conf. Math., Actuarial Sci., Comput. Sci. Statist. (MACS)*, Dec. 2019, pp. 1–5.

[24] W. Zhang, K. Ma, G. Zhai, and X. Yang, "Uncertainty-aware blind image quality assessment in the laboratory and wild," *IEEE Trans. Image Process.*, vol. 30, pp. 3474–3486, 2021.

[25] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "MetaIQA: Deep meta-learning for no-reference image quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14131–14140.

[26] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021.

[27] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

[28] J. Liu, X. Huang, J. Zheng, Y. Liu, and H. Li, "MixMAE: Mixed and masked autoencoder for efficient pretraining of hierarchical vision transformers," 2022, *arXiv:2205.13137*.

[29] T. Yao, Y. Pan, Y. Li, C.-W. Ngo, and T. Mei, "Wave-ViT: Unifying wavelet and transformers for visual representation learning," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 328–345.

[30] G. Yang, P. Rota, X. Alameda-Pineda, D. Xu, M. Ding, and E. Ricci, "Variational structured attention networks for deep visual representation learning," *IEEE Trans. Image Process.*, 2022.

[31] P. Goyal, Z. Hu, X. Liang, C. Wang, E. P. Xing, and C. Mellon, "Nonparametric variational auto-encoders for hierarchical representation learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5104–5112.

[32] J.-H. Kim, Y. Zhang, K. Han, Z. Wen, M. Choi, and Z. Liu, "Representation learning of resting state fMRI with variational autoencoder," *NeuroImage*, vol. 241, Jan. 2021, Art. no. 118423.

[33] J. Pereira and M. Silveira, "Unsupervised representation learning and anomaly detection in ECG sequences," *Int. J. Data Mining Bioinf.*, vol. 22, no. 4, p. 389, 2019.

[34] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*.

[35] J. Li, J. Jia, and D. Xu, "Unsupervised representation learning of image-based plant disease with deep convolutional generative adversarial networks," in *Proc. 37th Chin. Control Conf. (CCC)*, Jul. 2018, pp. 9159–9163.

[36] Y. Wei, X. Luo, L. Hu, Y. Peng, and J. Feng, "An improved unsupervised representation learning generative adversarial network for remote sensing image scene classification," *Remote Sens. Lett.*, vol. 11, no. 6, pp. 598–607, Jun. 2020.

[37] Q. Qian, Y. Xu, J. Hu, H. Li, and R. Jin, "Unsupervised visual representation learning by online constrained K-means," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16619–16628.

[38] B. Pang, Y. Zhang, Y. Li, J. Cai, and C. Lu, "Unsupervised visual representation learning by synchronous momentum grouping," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 265–282.

[39] A. Piergiovanni, A. Angelova, and M. S. Ryoo, "Evolving losses for unsupervised video representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 130–139.

[40] J. Wang and J. Jiang, "An unsupervised deep learning framework via integrated optimization of representation learning and GMM-based modeling," in *Proc. 14th Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2019, pp. 249–265.

[41] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5455–5516, Dec. 2020.

[42] N. Ahmed and H. M. S. Asif, "Perceptual quality assessment of digital images using deep features," *Comput. Informat.*, vol. 39, no. 3, pp. 385–409, 2020.

[43] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8697–8710.

[44] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," 2016, *arXiv:1611.01578*.

[45] Q. Zheng, X. Tian, M. Yang, Y. Wu, and H. Su, "PAC-Bayesian framework based drop-path method for 2D discriminative convolutional network pruning," *Multidimensional Syst. Signal Process.*, vol. 31, no. 3, pp. 793–827, Jul. 2020.

[46] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, Int. Telecommun. Union, Geneva, Switzerland, Tech. Rep., 2002.

[47] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, Feb. 2014.

[48] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[49] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[50] Z. Wang and E. P. Simoncelli, "Translation insensitive image similarity in complex wavelet domain," in *Proc. ICASSP*, Oct. 2005, p. 573.

[51] V. Laparra, J. Ballé, A. Berardino, and E. P. Simoncelii, "Perceptual image quality assessment using a normalized Laplacian pyramid," in *Proc. HVEI*, 2016, pp. 43–48.

[52] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 47–57, Mar. 2017.

[53] D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, Jan. 2010, Art. no. 011006.

[54] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2567–2581, May 2022.

[55] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[56] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[57] Q. Li, W. Lin, and Y. Fang, "No-reference quality assessment for multiply-distorted images in gradient domain," *IEEE Signal Process. Lett.*, vol. 23, no. 4, pp. 541–545, Apr. 2016.

[58] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 513–516, May 2010.

[59] M. Chandrasekhar, S. S. Channappayya, and S. S. Medasani, "Blind image quality evaluation using perception based features," in *Proc. 21st Nat. Conf. Commun. (NCC)*, Feb. 2015, pp. 1–6.

[60] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2579–2591, Aug. 2015.

[61] M. A. Saad and A. C. Bovik, "Blind quality assessment of videos using a model of natural scene statistics and motion coherency," in *Proc. Conf. Rec. 46th Asilomar Conf. Signals, Syst. Comput. (ASILOMAR)*, Nov. 2012, pp. 332–336.

[62] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011.

[63] L. Liu, H. Dong, H. Huang, and A. C. Bovik, "No-reference image quality assessment in curvelet domain," *Signal Process., Image Commun.*, vol. 29, no. 4, pp. 494–505, Apr. 2014.

[64] L. Liu, Y. Hua, Q. Zhao, H. Huang, and A. C. Bovik, "Blind image quality assessment by relative gradient statistics and AdaBoosting neural network," *Signal Process., Image Commun.*, vol. 40, pp. 1–15, Jan. 2016.

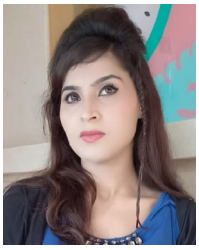
[65] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[66] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and Laplacian features," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4850–4862, Nov. 2014.
[67] X. Chen, Q. Zhang, M. Lin, G. Yang, and C. He, "No-reference color image quality assessment: From entropy to perceptual quality," *EURASIP J. Image Video Process.*, vol. 2019, no. 1, pp. 1–14, Dec. 2019.
[68] X. Min, G. Zhai, K. Gu, Y. Liu, and X. Yang, "Blind image quality estimation via distortion aggravation," *IEEE Trans. Broadcast.*, vol. 64, no. 2, pp. 508–517, Jun. 2018.
[69] F.-Z. Ou, Y.-G. Wang, and G. Zhu, "A novel blind image quality assessment method based on refined natural scene statistics," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1004–1008.
[70] D. Varga, "No-reference image quality assessment using the statistics of global and local image features," *Electronics*, vol. 12, no. 7, p. 1615, Mar. 2023.
[71] N. Ahmed, H. M. S. Asif, and H. Khalid, "PIQI: Perceptual image quality index based on ensemble of Gaussian process regression," *Multimedia Tools Appl.*, vol. 80, no. 10, pp. 15677–15700, Apr. 2021.
[72] H. Ren, D. Chen, and Y. Wang, "Ran4Iqa: Restorative adversarial nets for no-reference image quality assessment," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018.
[73] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 36–47, Jan. 2020.
[74] J. Ma, J. Wu, L. Li, W. Dong, X. Xie, G. Shi, and W. Lin, "Blind image quality assessment with active inference," *IEEE Trans. Image Process.*, vol. 30, pp. 3650–3663, 2021.

MUHAMMAD AZEEM ASLAM received the Ph.D. degree from Northwest Polytechnic University, Xi'an, China. He is currently with the School of Information Engineering, Xi'an Eurasia University, Xi'an. His primary research interests include computer vision and machine learning.

XU WEI received the B.S. degree in mechanical and electronic engineering from Jilin University, Changchun, China, in 2003, and the Ph.D. degree in mechanical and electronic engineering from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, in 2008. Since 2008, he has been with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences. He is a Research Fellow and a Ph.D. Supervisor. His current research interests include the integration technology of satellites and payloads and the highly reliable electronic systems for aerospace.

NISAR AHMED received the master's and Ph.D. degrees in computer engineering from the University of Engineering and Technology, Lahore, Pakistan. With more than 12 years of dedicated research and professional experience, he has made significant contributions to the fields of digital image processing, computer vision, machine learning, and data science. His current research interests include pattern recognition, computer vision, and digital image and video processing.



GULSHAN SALEEM received the master's degree in software engineering from the College of Electrical and Mechanical Engineering (CEME), National University of Science and Technology, Rawalpindi, Pakistan, in 2016. She is currently pursuing the Ph.D. degree in computer science with COMSATS University Islamabad, Lahore Campus, Pakistan. Concurrently, she is a Lecturer with the Department of Computer Science, Lahore Garrison University, Lahore. She is passionate about exploring innovative solutions at the intersection of these fields to address contemporary challenges in computer science and technology. Her primary research interests include computer vision, machine learning, and digital image processing.

TUBA AMIN received the bachelor's degree in information technology from Government College University Faisalabad, in 2014, and the master's degree in computer applications from the University of Agriculture Faisalabad, in 2017. She is currently a Lecturer with the Department of Computer Science, Government College University Faisalabad.

HUI CAIXUE received the bachelor's and master's degrees in information engineering from Xi'an Eurasia University, where she is currently pursuing the Ph.D. degree with the School of Information Engineering.