No-Reference Image Quality Assessment: An Attention Driven Approach

Diqi Chen(*)& Yizhou Wang(*)& Member, IEEE, and Wen Gao, Fellow, IEEE

Abstract—In this paper, we tackle no-reference image quality assessment (NR-IQA), which aims to predict the perceptual quality of a distorted image without referencing its pristine-quality counterpart. Inspired by the free-energy principle, we assume that, while perceiving a distorted image, the human visual system (HVS) tends to predict the pristine image then estimates the perceptual quality based on the distorted-restored pair. Furthermore, the perceptual quality depends heavily on the way how human beings attend to distorted images, namely, the cooperation of foveal vision and the eye movement mechanism. Inspired by these properties of the HVS, given the distorted-restored pair, we implement an attention-driven NR-IQA method with reinforcement learning (RL). The model learns a policy to attend to several regions parallely. The observations of the fixation regions are aggregated in a weighted average way, which is inspired by the robust averaging strategy. For policy learning, the rewards are derived from two tasks—classifying the distortion type and estimating the perceptual score. The goal of policy learning is to maximize the expectation of the accumulated rewards. Extensive experiments on LIVE, TID2008, TID2013 andCSIQ demonstrate the superiority of our methods.

Index Terms—No-reference image quality assessment, attention model.

I. INTRODUCTION

NOWADAYS, enormous visual signals are making their way to end users through mobile devices, social media, HDTV, etc. Therefore, it becomes increasingly paramount to devise computational models/methods to evaluate the perceptual quality of visual content automatically, i.e., image quality assessment (IQA). There are three main schemas in IQA: full-reference (FR) IQA, reduced-reference (RR) IQA and no-reference (NR) IQA. Reference images are required in FR-IQA, only partial information of reference images is accessible in RR-IQA, and NR-IQA only assessing distorted images. In this work, we focus on learning the NR-IQA models, which automatically predict the perceptual quality of a distorted image without referencing its pristine-quality counterpart, namely, its reference image with perfect quality.

NR-IQA is a challenging problem. (i) Compared with FR-IQA, NR-IQA—as the name tells—has no pristine-quality image to compare with. Hence, it is hard to determine the degree of degradation. (ii) Many distortions are perceptually inhomogeneous (e.g., ringing artifacts and block-wise mosaics [1]). Some regions contribute significantly to the holistic perceptual score of an image, and some contribute less. It is hard to identify the salient regions and quantify the influence of local distortions to predict the image quality. As is shown in Fig. 1, the distortions in both images are perceptually inhomogeneous. In the “parrot” image, the artifacts of ringing and blocking are more salient. In the “plane” image, the plane captures more attention than the sky. Both images are from the TID2008 dataset.

The free-energy principle [2], [3] implies that the human visual system (HVS) infers the environment based on the interior states. Inspired by this, we assume that the HVS predicts the pristine image while perceiving the distorted one, then judges the perceptual quality based on the distorted-restored pair. Furthermore, collaborated with the foveal vision—the HVS shows a roughly linear decrease of visual acuity with eccentricity [4], the HVS iteratively move the visual fixation at the task-informative regions to acquire and process task-related information efficiently. Besides, in subjective experiments, we observe that most eye fixations are closely related to...
more severely distorted regions. Hence, we assume that, while perceiving a distorted image, the attention is closely related to distortion type classification and perceptual quality prediction.

Therefore, in the attention-driven NR-IQA method proposed in our preliminary work [5], reinforcement learning (RL) is utilized to learn the attention mechanism as a simulation of the eye movement mechanism. The model [5] restores the distorted image first. Then given the distorted-restored pair, NR-IQA is formulated as a dynamic attention process, which is implemented by a recurrent neural network (RNN). This dynamic model learns a policy to attend to a sequence of regions step-by-step. At each step, the hidden states of the RNN are updated with the appearance features of the fixation region, which is a simulation that the HVS integrates all the previous observations. And the policy predicts the shift of the fixation, which is represented by a 2-dimensional real vector. Through several attention shifts, the RNN aggregates the observations to predict the distortion type and perceptual quality. During learning, the rewards are derived from these two tasks.

However, the dynamic model has three shortcomings. 1) The model converges slowly because of the massive action space for exploration. 2) At the early stage, training is slow and unstable because the policy learning is guided under inaccurate dual-task estimates. 3) The gradient vanishing problem is caused by the deep recurrent structure. To alleviate these shortcomings, we propose a static attention model in this paper. The idea of both models is to process salient regions subtly, handle the adjacent regions of salient regions coarsely and neglect non-salient regions. This design is consistent with the properties of the attention mechanism of the HVS.

In the first place, the restorative adversarial network (RAN) [6] is adopted to restore the distorted images. The reason is that the free energy principle implies that the HVS tends to minimize the distance between the distribution of restored images and the distribution of pristine images, rather than minimize a pixel-level error between two single images. RAN is a variant of the generative adversarial network (GAN) [7], including a restorator to restore the test image and a discriminator for discerning whether the input belongs to a restored image or a pristine one.

Based on the restored-distorted pair, the static model extracts multi-scale feature maps by a convolutional neural network (CNN). Based on the feature maps, on one side, the model estimates a quality map, a distortion type map and related weight maps. These estimations are processed in a fully convolutional manner. On the other side, the model learns a policy that integrates the observations of the whole image to predict attended regions parallelly. In this policy module, the recurrent connections are removed to alleviate the gradient vanishing problem. And the attended regions are represented by discretized variables, which reduces the exploration space.

Then, from the attended regions on the estimated maps, the scores and weights are extracted by Region of Interest (RoI) pooling and aggregated in a weighted average way. The weighted averaging is inspired by a psychological study — the robust averaging strategy [8], which defines a computational mechanism of perceptual judgment regarding how to integrate multiple/step-wise measurements. Since the extraction of the scores and weights are processed on the estimated maps, the static model does not need to utilize CNN to extract features on attended regions repeatedly, as is different from the dynamic model.

The static model is optimized by a hybrid loss function. The dual-task prediction is optimized by ground-truth labels in a supervised way. As an auxiliary task, predicting the distortion type is learned by the cross-entropy loss. As our ultimate goal, predicting quality score is optimized by the $\ell_1$ loss. The attention policy is learned in a reinforcement strategy, in which the task-driven rewards guide the learning of the policy to acquire the image regions highly related to NR-IQA. The gradients of policy learning are estimated by the REINFORCE algorithm [9]. There are two learning stages designed to alleviate the problem of unstable learning. At the pre-training stage, the estimation is calculated from the whole image without sampling attended regions. At the attention learning stage, the model learns the attention policy and dual-task prediction simultaneously.

In experiments, the proposed methods outperform several state-of-the-art methods on LIVE [10], TID2008 [11], TID2013 [11] and CSIQ [12].

II. RELATED WORK

A. Neural Networks for NR-IQA

Patch-based methods predict holistic perceptual quality by assessing uniformly distributed patches. Kang et al. [13] proposed a patch-based NR-IQA method, in which the image patches are uniformly extracted in a predefined size, then a CNN is trained to predict a quality score for each patch and the scores are averaged as the holistic image score. Based on this work, Kang et al. [14] designed a multi-task framework to predict the distortion type and the quality score of each patch simultaneously. Different from patch-based methods, our method makes an image-level prediction which is leveraged by the attention mechanism.

Some methods train neural networks to predict the perceptual score on hand-crafted features. BRISQUE [15] extracts the natural scene statistic (NSS) features in the spatial domain, then trains a support vector regression (SVR) model for quality estimation, since SVR can be regarded as a simple two-layer neural network. Gu et al. [16] introduced an approach training an SVR model on three groups of features: the features of the free energy, the structural degradation information and the NSS model. A Restricted Boltzmann Machine is adopted by Tang et al. [17] to map the LBIQ features [18] to perceptual scores. Hou et al. [19] proposed a method that extracts the NSS features in the wavelet transform domain, in which the IQA problem is regarded as a classification problem and a quality pooling method is used under the Bayesian framework.

Furthermore, there are some end-to-end frameworks straightforwardly mapping the raw images to quality scores. Bosse et al. [20] designed a deep network to predict both the quality and the weight of each patch, then the image score is calculated by weighted averaging. But context information is not considered in [20] while predicting the weights, which
is different from our model. Kim and Lee [21] introduced an approach learning a CNN to predict the overall quality. In this method, as an auxiliary task, the model generates local scores with an FR-IQA metric and trains the CNN to map the extracted patches to the derived scores. Ren et al. [6] proposed a two-stage method based on distorted-restored pairs. At the pre-training stage, the network is trained to predict the pseudo scores and weights of sampled patches. The pseudo labels are generated by a reliable FR method (FSIM [22]) and the patches are sampled from unlabeled images. At the training stage, the method fine-tunes the network to fit the ground-truth scores on labeled data. Liu et al. [23] designed a framework that pre-trains a neural network on self-generated image pairs by a learning-to-rank (L2R) algorithm. The image pairs have the same content and distortion type. This pre-trained network needs to be fine-tuned on cross-type evaluation. In [24] Ma et al. adopt the L2R idea as well. Specifically, the quality-discriminable image pairs are defined on discriminable estimates of reliable FR-IQA methods. The quality-discriminable image pairs are generated from different contents and distortion types so that the learned model performs well on cross-type evaluation without fine-tuning.

B. Saliency in IQA

Although there lacks literature that automatically learns the attention mechanism in NR-IQA, some studies investigate how to combine computational saliency models with NR-IQA methods. Liu and Heynderickx [25] proposed a method which adopts saliency values as weights to average the estimated patch scores. The saliency values are obtained from eye-tracking data. Hou and Gao [26] designed a framework that learns to construct features on salient regions. Zhang et al. [27] investigated different combinations between saliency models and IQA methods, which studies the influence of visual attention on perceptual quality. The computational saliency models used in the above methods are designed initially to compute saliency maps of natural images, while human beings adopt different mechanisms when scanning an image with and without tasks [28]. Furthermore, we believe that image quality assessment depends heavily on the way human beings attend to images. Hence, the attention mechanism is explicitly modeled and learned with reinforcement learning in this work.

C. Attention Models

Recently, deep learning models with attention mechanisms receive a lot of interest. Implemented in the soft attention models [29], [30], the deterministic attention mechanism is differentiable and trained by backpropagation. Sorokin et al. [29] proposed a method which adopts the element-wise multiplication to learn the importance of each area in video games. Kuen et al. [30] introduced an approach utilizing the differentiable spatial transformer [31] to implement attention mechanism to refine the estimated saliency map step-by-step. Stochastic attention mechanisms are implemented in the hard attention models [32], [33], in which RL is an alternative tool for training. Mnih et al. [32] proposed an attention model for object recognition. This model is learned by a policy gradient method, namely, REINFORCE [9]. Ba et al. [33] designed an attention model learned by maximizing a variational lower bound to recognize and localize multiple objects. The model learning can be regarded as a REINFORCE algorithm with specified rewards. Caicedo et al. [34] introduced a top-down analysis method to localize the objects. The localization is processed through a sequence of observations, which are sampled from a learned class-specific policy. Mathe et al. [35] also proposed an approach to localize objects, in which jumping between bounding boxes is adopted as the action. In [36], a tree-structured RL method is proposed to detect multi-scale bounding boxes as object proposals. In [37], a hierarchical attention architecture is designed to scan salient regions sequentially for multiset classification.

Compared with the above models, our model is different from three main perspectives. (i) Our model is multi-task, i.e., it jointly optimizes the performance of two closely related tasks to learn representation. Consequently, the reward is enriched by both tasks. Such enriched rewards empower the learned policy to capture the task-related regions. By doing this, the observations are aggregated more accurately and efficiently. (ii) The robust averaging strategy of perceptual judgment is implemented into the network architecture and learning. (iii) A fully convolutional structure with two learning stages is implemented. Thus the model learning is more computationally stable and efficient.

III. THE DYNAMIC MODEL

The dynamic model consists of four main components — the RAN, a multi-scale image analysis module, a location sampling module and a feature aggregation module. To begin with, the restorator of RAN restores the distorted image. Then distorted-restored pair is treated as the input of the attention model. The attention process starts from an initial location. At each time step, the dynamic model extracts multi-resolution patches on the attended location, then normalizes the patches into the same size. In the multi-scale image analysis module, the CNN extracts features from the normalized patches to update the recurrent layers. Based on the hidden states of the recurrent layer, the location sampling module predicts the next location and the feature aggregation module predicts the temporary quality score and weight. This procedure is repeated until the attention process ends. Then the distortion type prediction and quality score estimation are calculated in a weighted average way. The rewards are derived from dual tasks. The goal of policy learning is to maximize the expectation of the accumulated rewards.

Please refer to our preliminary work [5] for more details of the dynamic model.

IV. THE STATIC MODEL

The static model is proposed to alleviate the weaknesses of the dynamic model. Firstly, a fully convolutional structure takes the place of the RNN structure. This design shortens the length of the gradients backpropagation path. Thus, the gradient vanishing problem is alleviated. Secondly, discrete
actions are used to reduce the search space. Thirdly, a two-stage training scheme is implemented to alleviate the problem of unstable training, which is caused by the inaccurate dual-task estimates at the early learning stage.

The static model aims to map a distorted-restored pair \( x \) to a quality score \( s \). Fig. 2 shows four components of the static model: the RAN, a parallel multi-scale image analysis module, a parallel location sampling module and a parallel score aggregation module. The distorted-restored pair \( x \) is generated by the restorer of RAN and treated as the input of the attention model. The image analysis module extracts feature maps \( f_{\text{conv}} \) from \( x \) by a CNN using multi-scale convolution kernels. Based on \( f_{\text{conv}} \), on one hand, the score aggregation module predicts the quality map \( M_s \), the distortion type map \( M_y \) and related weight maps \( W_s \), \( W_y \). This prediction is processed in a full convolution manner. On the other hand, the location sampling module samples \( K \) fixation regions parallelly. Then each region is represented by a score \( s_k \) and a weight \( \alpha_k \) by applying the Region of Interest (RoI) pooling on the estimated maps, where \( k \) is the index of sampled regions. Then the holistic quality score is calculated in a weighted average way \( s = \frac{\sum_{k=1}^{K} s_k \alpha_k}{\sum_{k=1}^{K} \alpha_k} \). The distortion type is calculated in the same way. The details of these modules are described below.

A. The Model Components

1) The Restorative Adversarial Nets (RAN): This module restores distorted images and generates distorted-restored pair \( x \) as the input of the attention model. In this work, we adopt the design of the RAN from [6]. Fig. 2(a) shows the detailed network structures. The RAN includes a restorer and a discriminator. The discriminator is trained to discern whether the input belongs to a restored image or a pristine one. The restorer restores the test images to fool the discriminator. The restorer and the discriminator are built by the Residual blocks [38] and the strided blocks, respectively. In both networks, batch normalization [39] is adopted to avoid undesired initialization. The restorer is trained through optimizing the perceptual loss [40] and adversarial loss [7]. To define perceptual loss, the module utilizes the VGG19 [41] network pre-trained on ImageNet [42]. In VGG19, the method selects the convolution layers before the previous three max-pooling layers and extract two batches of feature maps from the restored patch and the pristine patch. The perceptual loss is defined by the Euclidean distance between these two batches. To define adversarial loss, inspired by the Wasserstein GAN [43], the Wasserstein distance is adopted to measure the divergence between the distribution of restored images and the distribution of pristine images. This design remits the gradient vanishing problem and makes learning more stable. During
training, the inputs of the restorator and the discriminator are $64 \times 64$ patches. When restoring images, the restorator can be applied to arbitrary-sized images. The reason is that the restorator network only includes the layers that do not affect the size of inputs.

2) The Parallel Multi-Scale Image Analysis Module: This module learns a multi-scale representation $f_{\text{conv}}$ for the distorted-restored pair $x$. Fig. 2(b) shows the detailed structure of this module, which is built with multi-scale convolutional layers and fully convolutional layers [44]. The multi-scale convolutional kernels are adopted to make the learning more efficient, which is suggested by [45]. Concerning the kernel size, there are three kinds of kernels in every convolution layer. The sizes of these kernels are $3 \times 3$, $5 \times 5$ and $1 \times 1$ respectively. The ratio of the numbers of these three kinds of kernels is $2 : 1 : 1$. In the first two convolution layers, half of the $3 \times 3$ and $5 \times 5$ kernels are replaced with dilated kernels [46], with dilation factor 3. Formally, the 3-dilated convolution is defined as $(F \ast_d p)(x) = \sum_{s+3d=0}^p F(s)d(t)$, where $F$ denotes the input feature maps, $d$ denotes the dilated kernel, $\ast_d$ denotes the 3-dilated convolution operation. And $p, s, t$ indicates a specific location of the output feature map, the input feature map, and the dilated kernel, respectively. The dilated kernels are utilized to enlarge the receptive fields of the neurons in $f_{\text{conv}}$. By doing this, the contextual information is collected, which is a benefit to not only multi-scale image analysis but also the saliency/attention prediction. Through this operation, the salient regions are processed subtly and the adjacent regions of salient regions are processed coarsely. The fully convolutional layers apply more dense sampling than the patchwise sampling and can handle arbitrary-sized inputs [44]. All these convolutional operations do not change the sizes of the feature maps due to appropriate paddings.

3) The Parallel Location Sampling Module: Based on $f_{\text{conv}}$, this module samples several fixation regions parallelly. To handle arbitrary-sized images, an adaptive average pooling layer is adopted to map $f_{\text{conv}}$ into rescaled feature maps with predefined size $N_c \times N_x \times N_y$, where $N_c$ is the number of channels and $N_x \times N_y$ is the size of each map. Then a convolution layer, a flatten operation and a fully connected layer are stacked on the rescaled feature maps to compute the hidden states $h_l$. The receptive field of each neuron in $h_l$ covers the whole image. Hence, in the static model, the fully connected layer in this module takes the place of the recurrent connections for integrating local observations.

Based on $h_l$, the module predicts $K$ attended regions $l = \{l_k\}_{k=1}^K$. For simplicity, the superscript $k$ is omitted while describing each region. Each region is assumed to be a square area on the rescaled map. The square area is represented by a triplet $(l_x, l_y, l_{\text{scale}})$, where $(l_x, l_y)$ indicates the center of the attended region and $l_{\text{scale}}$ indicates the length. To reduce the exploration space, the actions are discretized, so these three elements are positive integers. Specifically, $l_x$ ranges from 1 to $N_x$, $l_y$ ranges from 1 to $N_y$ and $l_{\text{scale}}$ ranges from 1 to 4 empirically. Each element follows a Categorical distribution predicted by feeding $h_l$ to a linear layer and a softmax layer. For example, $l_x$ is sampled from $\text{Cat}(\mu_x)$, where $\mu_x = \text{SoftMax}(W_x h_l + b_l)$. Cat denotes the Categorical distribution, $\mu_x$ denotes the estimated variables of the distribution, $W_l$ denotes the weight and $b_l$ denotes the bias. $l_x$ and $l_{\text{scale}}$ are calculated in the same way. By doing this, the module estimates a group of variables $\mu = (\mu_x, \mu_y, \mu_{\text{scale}})$ which denote three Categorical distributions for defining an attended region. When testing, the attended locations are selected according to the maximum probability. The details of the learning method are discussed in Section IV-B.

4) The Parallel Score Aggregation Module: Based on $f_{\text{conv}}$ as well, this module estimates four maps by fully convolutional operation: a map of quality scores $M_s$, a map of weights of quality $W_s$, a map of distortion types $M_y$, a map of weights of distortion types $W_y$. Specifically, $M_s, W_s$ and $W_y$ are $N'_x \times N'_y$ matrices and $M_y$ is a $C \times N'_x \times N'_y$ matrix, where $C$ denotes the number of distortion types. Each element in these matrices represents a specific estimation of a local region. Then the $K$ sampled regions are proportionally projected onto the predicted maps as RoI regions. For instance, the X-coordinate of the projected location is calculated by $l'_x = l_x \times N'_x / N_x$. On $K$ projected regions, the RoI pooling [47] is adopted to aggregate estimations. By doing this, each region is represented by an averaged quality $s^k$ and an averaged weight $\alpha^k$. Hence, the holistic quality score is calculated by weighted averaging all the extracted estimates: $s = \sum_{k=1}^K \alpha^k s^k$. The distortion type $y$ is estimated in the same way. This weighted average way is inspired by [8], which states that an optimal agent makes judgments based on the strength and reliability of decision-relevant evidence. In our case, the observation of the attended region denotes the decision-relevant evidence, the predicted quality score denotes the strength, and the predicted weight denotes the reliability.

B. Learning of the Static Model

The loss function includes three terms:

$$L = L_{\text{reg}} + L_{\text{cls}} - \alpha J_{\text{rein}},$$

where $\alpha$ is the free parameter, the cross-entropy loss $L_{\text{cls}}$ is minimized for distortion type classification and the $\ell_1$ loss $L_{\text{reg}}$ is minimized for quality score prediction. The RL term $J_{\text{rein}}$ is the expectation of accumulated rewards following the current policy. In the perspective of RL, $f_{\text{conv}}$ represents the internal states, location prediction denotes the action, and the location sampling module represents the policy. The reward function is defined as

$$R = \begin{cases} 1 & y = y' \text{ or } \|s - s'\| < \delta \\ 0 & \text{otherwise} \end{cases},$$

where $\delta$, $y'$, $s'$ are a threshold, the ground-truth distortion type and the ground-truth quality score, respectively. The reward is equal to one if the classification is correct or the score prediction is accurate enough; zero otherwise.

In one minibatch, the derivative of $J_{\text{rein}}$ is approximated by

$$\nabla_{\theta} J_{\text{rein}} \approx \frac{1}{M} \sum_{i=1}^M \sum_{k=1}^K \nabla_{\theta} \log p(l'_i | \mu_k) R_i,$$
where \( i \) denotes the index of the training images, \( M \) denotes the number of images in one mini-batch, \( k \) denotes the index of the predicted fixation regions, \( \mu_i^k \) denotes the estimated variables denoting the \( k \)-th fixation region of the \( i \)-th image, \( I_i^k \) denotes the \( k \)-th sampled region of the \( i \)-th image, and \( \theta_R \) represents the parameters of the parallel multi-scale analysis module and the parallel location sampling module. Then the model is optimized by stochastic gradients descent method end-to-end. Intuitively, the attention policy is learned by maximizing the above log likelihood guided by the reward. Therefore, the goal of the policy learning is to classify distortion type and predict score accurately.

To alleviate the problem of unstable training, the learning process of the static model includes two stages: the pre-training stage and the attention learning stage. At the pre-training stage, the quality score and the distortion type is calculated over the whole image rather than attended regions. The RL term is disabled by setting \( \alpha \) to 0. At the attention learning stage, the attention policy and the dual-task estimation are learned simultaneously.

V. EXPERIMENTAL RESULTS

A. Implementation Details

1) Programming Environment: Our codes are implemented by PyTorch (version 0.2.0) and tested on Nvidia K40 GPU.

2) Preprocessing: The images are represented in three channels (RGB). The local contrast normalization method is applied to each channel of the input images, which is suggested in [13].

3) Training of the RAN: The training of the RAN includes two stages. At the first stage, only the restorator is trained to reconstruct the pristine patches. The Adam optimizer [52] is utilized at a learning rate of \( 10^{-4} \) for 300,000 iterations. The pixel-wise mean square error is also optimized to reserve more low-level details, which is different from [6]. Then the restorator and the discriminator are trained simultaneously using RMSProp [53] with a learning rate \( 10^{-4} \) for 300,000 iterations and a lower learning rate \( 10^{-5} \) for another 300,000 iterations. In each iteration, the discriminator is trained five times and the restorator once.

4) Training of the Attention Model: The network parameters are initialized by the Kaiming normal method [54]. The Adam [52] algorithm with momentum 0.9 is used for optimization. In the Equation 1, parameter \( \alpha \) in the RL term is equal to 0.01. The reason is that small \( \alpha \) slows the policy training, then the policy is more random at the early stage, which encourages exploration. During training on TID2008 or TID2013, \( \delta \) in Equation 2 is set to 0.5, which are chosen based on the results of pre-experiments on a random train/validation split. And \( \delta \) is set to 7.0 on the LIVE dataset. The learning rate is initialized with 0.001 and linearly decayed to 0.0001. The model is trained by 300 epochs at both the pre-training stage and the attention learning stage. The number of sampled locations equals to 5, which are also chosen through the pre-experiments.

5) Training/Testing on Images With Different Sizes: As far as we know, in previous works, there are no discussions about applying attention models to arbitrary-sized images. The reason may be the training/testing images could be normalized into the same size in their problems. For example, regarding object recognition/detection, there is a slight influence when images are rescaled. However, the perceptual quality is vulnerable for rescaling, so IQA methods should reserve the size of images when training and testing. In this scenario, it is critical to discuss how to apply our attention models to arbitrary-sized inputs. Firstly, the dynamic model is considered. The dynamic model samples several fixations sequentially and extracts multi-resolution patches with pre-defined size on attended locations. The fixation is represented by a 2-dimensional vector. Therefore, it is necessary to guarantee the consistent meaning of this vector. Our approach is to adopt adaptive sizes for low-resolution patches during the extraction of multi-resolution patches. The adaptation is conditioned on a constant ratio of the width/height of the low-resolution patch to the width/height of the distorted image. Secondly, in the static model, the parallel location sampling module adaptively rescales the feature maps to a pre-defined size. Hence, the static model is not sensitive to the size of the inputs.

B. Experimental Protocol

1) Datasets for Generating Distorted Images:

- **CSIQ [12]:** The CSIQ database consists of 30 reference images, five types of distortions. Differential Mean Opinion Scores (DMOS) in the range [0, 100] are provided for distorted images.
- **TID2008** [11]: This dataset consists of 25 reference images, 17 types of distortions. Each distortion type has four levels. There are in total 1700 distorted images, each of which is labeled with a Mean Opinion Score (MOS) between 0 and 9.
- **TID2013** [1]: TID2013 is an extended version of TID2008. It consists of the same 25 reference images as TID2008, while there are 24 types of distortions and five levels of each type.
- **LIVE** [10]: This dataset consists of 29 reference images, five types of distortions. Differential Mean Opinion Scores (DMOS) in the range [0, 100] are provided for distorted images.

**Datasets for Generating Distorted Images:** In order to train the RAN, distorted images are generated from the Waterloo dataset [55], which consists of 4,744 pristine natural images. We generate 11 out of a total of 17 distortions on TID2008. The 6 unimplemented distortions are #3, #4, #6, #7, #12, #13.
C. Experimental Results on LIVE/TID2008/TID2013

To compare our attention models with previous methods on the LIVE/TID2008/TID2013 database, all images with different distortions are trained simultaneously. The overall results are presented in Table I.

Four popular FR-IQA methods (PSNR, SSIM [48], FSIM [22] and VSI [49]) are tested for reference. Our models are compared against three NR-IQA methods without using deep networks (CORNIA [50], BRISQUE [15], HOSA [51]) and four deep learning-based methods (CNN [13], CNN++ [14], RankIQA [23] and RAN4IQA [6]). Regarding the methods without learning/fine-tuning, a non-linear mapping method is adopted to map the predicted scores to labeled scores before evaluation, which is suggested by the Video Quality Experts Group (VQEG) [56]. Our proposed methods are indicated in bold.

For a fair comparison, the experiments are conducted on the same random splits for all methods and the source codes of the compared methods are downloaded from the home pages of the related authors. The best results of FR-IQA and NR-IQA are indicated in bold separately as shown in Table I. Concerning the SROCC results on TID2013, BRISQUE [15] and HOSA [51] and four deep learning-based methods (CNN [13], CNN++ [14], RankIQA [23] and RAN4IQA [6]). Regarding the methods without learning/fine-tuning, a non-linear mapping method is adopted to map the predicted scores to labeled scores before evaluation, which is suggested by the Video Quality Experts Group (VQEG) [56]. Our proposed methods are indicated in bold.

For a fair comparison, the experiments are conducted on the same random splits for all methods and the source codes of the compared methods are downloaded from the home pages of the related authors. The best results of FR-IQA and NR-IQA are indicated in bold separately as shown in Table I.

Concerning the SROCC results on TID2013, BRISQUE [15] and HOSA [51] and four deep learning-based methods (CNN [13], CNN++ [14], RankIQA [23] and RAN4IQA [6]). Regarding the methods without learning/fine-tuning, a non-linear mapping method is adopted to map the predicted scores to labeled scores before evaluation, which is suggested by the Video Quality Experts Group (VQEG) [56]. Our proposed methods are indicated in bold.

For a fair comparison, the experiments are conducted on the same random splits for all methods and the source codes of the compared methods are downloaded from the home pages of the related authors. The best results of FR-IQA and NR-IQA are indicated in bold separately as shown in Table I.

Concerning the SROCC results on TID2013, BRISQUE [15] and HOSA [51] and four deep learning-based methods (CNN [13], CNN++ [14], RankIQA [23] and RAN4IQA [6]). Regarding the methods without learning/fine-tuning, a non-linear mapping method is adopted to map the predicted scores to labeled scores before evaluation, which is suggested by the Video Quality Experts Group (VQEG) [56]. Our proposed methods are indicated in bold.
D. Ablation Experiments

1) Ablation Experiments of the Dynamic Model: Each of the components listed below is represented by a capital letter in the bracket. The concatenation of abbreviations represents the model constructed by related components. (1) (R)einforcement Learning. This component represents a dynamic attention model implemented by RL. (2) (M)ulti-Task Learning. This component denotes that the model is learned through optimizing dual tasks. Without this component, the model only predicts the perceptual quality. (3) Robust (A)veraging Strategy. This component denotes that the model predicts temporary quality scores and weights, then the holistic score is estimated in a weighted average way. Without this component, the model estimates the holistic score based on the hidden states of the last time step. (4) Reward from (D)istortion Map. Another reward function is investigated. This reward function is not task-driven. Specifically, a distortion map is calculated from the distorted image and the reference image by an FR-IQA metric (VSI). Then the distortion map is smoothed by a Gaussian kernel. The values of the smoothed map are used as rewards. (5) Distorted-Restored (P)air. The inputs of the attention model are distorted-restored pairs rather than only distorted images. Furthermore, P′ represents the model using distorted-restored pairs generated by the restorator trained without adversarial distance.

For instance, RMA contains the dynamic model implemented by RL, the multi-task learning and the robust averaging strategy. The results of R, RM, RMA, RMAD, RMAP′ and RMAP are demonstrated in Table III. Some conclusions are listed below. (1) RM outperforms R, which implies that the auxiliary task is beneficial to representation learning and policy learning. (2) Through comparing RMA with RM, the results show that the performance is improved by adding the robust average strategy. The reason is that the weighting strategy is more consistent with the strategy of the HVS. (3) The results of RMAP are better than the results of RMA, which demonstrates the superiority of using distorted-restored pairs as inputs. (4) RMA outperforms RMAD, which implies that the regions generated by the task-driven policy are more related to the image quality. (5) The comparison between RMAP′ and RMAP shows that Learning a restorator without adversarial distance degrades the quality assessment results.

2) Ablation Experiments of the Static Model: To investigate the contributions of the dilated kernels and the static attention mechanism, four models concerning whether using or not using these two modules are compared in Table III. Taking the results on TID2008 as an example, compared with the model without these two modules, either the dilated convolution or the attention mechanism brings some performance gain. The contribution of the attention mechanism is slightly higher than the dilated convolution but both models show limited improvement. Furthermore, the cooperation between the dilated convolution and the attention mechanism outperforms other baseline methods with a considerable margin. This observation suggests that the cooperation between the dilated convolution and the attention mechanism is the key factor for better performance.

E. External Experiments

1) Cross Dataset Test: In the cross-dataset test, our model is trained on the full TID2008 dataset then tested on CSIQ. Only the images of four common distortion types of two datasets are considered. The four shared distortions are JPEG compression, JPEG2000 compression, Gaussian blur and additive white Gaussian noise. Since the images on CSIQ are labeled by DMOS([0, 1]) and the images on TID2008 are labeled by MOS([0, 9]), a non-linear mapping method is applied to eliminate the disparity of two datasets before evaluation. This process is suggested by the Video Quality Experts Group (VQEG) [56]. As shown in Table IV, the dynamic model outperforms CNN [13], CNN++ [14], RAN4IQA [6] and RankIQA [23], and the static model outperforms the dynamic model considering SROCC and KROCC.

2) Experiments of the Dynamic Model on Different Ts: In the dynamic model, the number of steps $T$ is determined by the results on the validation set of a pre-experiment. We still conduct comparative experiments on different $T$ (4, 5 and 6), which are shown in Table V. The results manifest that Dynamic($T = 5$) outperforms Dynamic($T = 4$) in three out

<table>
<thead>
<tr>
<th>Dataset</th>
<th>TID2008</th>
<th>TID2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methods</td>
<td>SROCC</td>
<td>KROCC</td>
</tr>
<tr>
<td>Dynamic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>0.737</td>
<td>0.548</td>
</tr>
<tr>
<td>RM</td>
<td>0.824</td>
<td>0.636</td>
</tr>
<tr>
<td>RMA</td>
<td>0.864</td>
<td>0.648</td>
</tr>
<tr>
<td>RMAD</td>
<td>0.839</td>
<td>0.643</td>
</tr>
<tr>
<td>RMAP′</td>
<td>0.851</td>
<td>0.653</td>
</tr>
<tr>
<td>RMAP</td>
<td>0.865</td>
<td>0.679</td>
</tr>
<tr>
<td>Static</td>
<td></td>
<td></td>
</tr>
<tr>
<td>w/o dilated, w/o att.</td>
<td>0.829</td>
<td>0.642</td>
</tr>
<tr>
<td>w dilated, w/o att.</td>
<td>0.832</td>
<td>0.636</td>
</tr>
<tr>
<td>w/o dilated, w att.</td>
<td>0.835</td>
<td>0.643</td>
</tr>
<tr>
<td>w dilated, w att.</td>
<td>0.862</td>
<td>0.681</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE IV</th>
<th>CROSS DATASET TEST</th>
<th>TID2008 -&gt; CSIQ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SROCC</td>
<td>KROCC</td>
</tr>
<tr>
<td>CNN [13]</td>
<td>0.756</td>
<td>0.554</td>
</tr>
<tr>
<td>CNN++ [14]</td>
<td>0.786</td>
<td>0.585</td>
</tr>
<tr>
<td>RAN4IQA [6]</td>
<td>0.827</td>
<td>0.615</td>
</tr>
<tr>
<td>RankIQA [23]</td>
<td>0.823</td>
<td>0.621</td>
</tr>
<tr>
<td>Dynamic</td>
<td>0.850</td>
<td>0.660</td>
</tr>
<tr>
<td>Static</td>
<td>0.870</td>
<td>0.686</td>
</tr>
</tbody>
</table>
Fig. 3. The last attended regions of four images with the local block-wise distortions of different levels. Our model attends to the block masks.

<table>
<thead>
<tr>
<th>Methods</th>
<th>SROCC</th>
<th>KROCC</th>
<th>LCC</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamic((T = 5))</td>
<td>0.812</td>
<td>0.631</td>
<td>0.845</td>
<td>0.677</td>
</tr>
<tr>
<td>Dynamic((T = 4))</td>
<td>0.809</td>
<td>0.624</td>
<td>0.837</td>
<td>0.669</td>
</tr>
<tr>
<td>Dynamic((T = 6))</td>
<td>0.797</td>
<td>0.612</td>
<td>0.821</td>
<td>0.683</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods</th>
<th>SROCC</th>
<th>KROCC</th>
<th>LCC</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static((K = 5))</td>
<td>0.820</td>
<td>0.640</td>
<td>0.842</td>
<td>0.699</td>
</tr>
<tr>
<td>Static((K = 4))</td>
<td>0.813</td>
<td>0.635</td>
<td>0.844</td>
<td>0.702</td>
</tr>
<tr>
<td>Static((K = 6))</td>
<td>0.817</td>
<td>0.633</td>
<td>0.836</td>
<td>0.694</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods</th>
<th>SROCC</th>
<th>KROCC</th>
<th>LCC</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CM</td>
<td>0.752</td>
<td>0.738</td>
<td>0.111</td>
<td>0.630</td>
</tr>
<tr>
<td>RMA</td>
<td>0.683</td>
<td>0.748</td>
<td>0.106</td>
<td>0.739</td>
</tr>
<tr>
<td>Static</td>
<td>0.696</td>
<td>0.480</td>
<td>0.145</td>
<td>0.817</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SROCC</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>CM</td>
<td>0.702</td>
<td>0.683</td>
<td>0.065</td>
<td>0.383</td>
<td>0.613</td>
</tr>
<tr>
<td>RMA</td>
<td>0.686</td>
<td>0.707</td>
<td>0.071</td>
<td>0.740</td>
<td>0.658</td>
</tr>
<tr>
<td>Static</td>
<td>0.664</td>
<td>0.446</td>
<td>0.094</td>
<td>0.824</td>
<td>0.703</td>
</tr>
</tbody>
</table>

of four metrics. The result is not positive when \( T \) equals to 6 because of the gradient vanishing problem caused by training a deep structure of RNN.

3) Experiments of the Static Model on Different \( K \)s: We conduct comparative experiments on different \( K \)s (4, 5 and 6), which are shown in Table VI. The results show that Static\((T = 5)\) outperforms Static\((T = 4)\) and Static\((T = 6)\) both in three out of four metrics. In this setting, the results of these three models are very close to each other, which shows the static model is more stable than the dynamic model. This observation implies that the gradient vanishing problem is alleviated through removing recurrent connections.

4) Experiments on Spatially Inhomogeneous Distortion Types: To show the benefit of the attention mechanism on the images with spatially inhomogeneous distortions, two comparative baselines are designed. The CM denotes a patch-based method using a multi-task CNN. The RMA represents a multi-task dynamic attention model implemented by RL and the robust averaging strategy. The static model is compared with CM and RMA on the images of four spatially inhomogeneous distortion types on TID2008. This static model only takes distorted images as inputs for a fair comparison. The four tested distortion types are JPEG transmission errors, JPEG2000 transmission errors, non-eccentricity pattern noise and local block-wise distortions of different intensity. In Table VII, the numbers in the first row represent the indexes of these four distortion types. The results show that RMA outperforms CM\((0.691 \text{ v.s. } 0.617 \text{ on SROCC, } 0.658 \text{ v.s. } 0.613 \text{ on LCC})\), which shows the improvement from the dynamic attention mechanism. Furthermore, the static model achieves 0.739 and 0.703 on SROCC and LCC. The static model outperforms RMA, which suggests that the static model shows more superiority on the dynamic model in this challenging problem setting.

5) Computational Time: The testing time between our models and the state-of-the-art method RankIQA [23] are compared. The size of the distorted images is \( 3 \times 384 \times 512 \) and the batch size is 512. Our programs are processed on a single Nvidia K40 GPU. For each image, averagely, RankIQA costs 0.399 seconds, the dynamic model costs 0.232 seconds and the static model costs 0.234 seconds. The restoration of each image takes 0.107 seconds, which is included in the testing time of our models.

F. Qualitative Results

Some qualitative results of the dynamic model are shown in Figures 3, 4 and 5. In Fig. 3 and Fig. 4, the sampled patches (the second scale of the three scales) are magnified and represented at the bottom right corner of each image.

The local block-wise distortion degrades image quality by adding some annoying blocks with different intensity. Fig. 3 shows that our model attends to the salient regions with blocks. Notice that in the last two images, the captured distorted regions contain only a few blocks.

Fig. 4 shows that the last attended regions of four images with JPEG2000 transmission errors. The first image is distorted mildly, while the most salient region of the other three images is around those two people. Our model attends to these salient regions for all the right three images.

Fig. 5 shows the attentional scanpaths of the four images distorted by JPEG2000 transmission errors with different levels. The most salient region of the left three images is around the ball. For the last image, our attention will be scattered because the last image is distorted heavily everywhere.
The movement of the fixations of our model is consistent with this observation.

VI. CONCLUSION

In this paper, we implement an attention-driven model by a fully convolutional structure based on distorted-restored image pairs for NR-IQA. Based on the extracted feature maps, on one side, the model estimates the quality maps, distortion type maps and related weight maps; on the other side, the model predicts the salient regions. The holistic estimation is integrated from the projected regions on the estimated maps by RoI pooling and the weighted averaging strategy. The multi-task prediction is optimized by ground-truth labels in a supervised way. And the sampling policy is learned in a reinforcement strategy, in which the task-driven rewards guide the learning of the sampling policy to acquire the image regions highly related to IQA. Extensive experiments demonstrate the superiority of our methods.

The proposed model may be improved or extended in many ways. Firstly, some constraints may be added to the attended regions to encourage policy exploration. Secondly, some interpolation techniques may be adopted to address the limitation of the discretized representation of fixation regions. Thirdly, the attention mechanism can be extended to video quality assessment tasks.

REFERENCES


