Recurrent Attentional Model for No-Reference Image Quality Assessment

Diqi Chen¹, Yizhou Wang¹, Tianfu Wu², Wen Gao¹
¹Nat’l Engineering Laboratory for Video Technology,
Key Laboratory of Machine Perception (MoE),
Sch’l of EECS, Peking University, Beijing, 100871, China
²Department of ECE and the Visual Narrative Cluster,
North Carolina State University
{cdq, Yizhou.Wang, wgao}@pku.edu.cn, tfwu@ncsu.edu

Abstract

This paper presents a recurrent attentional model (RAM) for general no-reference image quality assessment (NR-IQA), that is to predict the perceptual quality score for an input image without using any reference image and/or prior knowledge regarding underlying distortions. The proposed RAM is inspired by the well known visual attention mechanism, both covert and overt, which affects many aspects of visual perception including image quality assessment. The attentional mechanism is, however, largely ignored in the NR-IQA literature. The proposed RAM hypothesizes that the attentional scanning path in an image should contain intrinsic information for IQA. The RAM thus consists of three components: a glimpse sub-network analyzing the quality at a fixation using multi-scale information, a location sub-network selecting where to look next by sampling a stochastic node, and a recurrent network aggregating information along the scanning path to compute the final prediction. The RAM is formulated under multi-task learning for the joint prediction of distortion type and image quality score and for the REINFORCE rule [21] used to handle the stochastic node. The RAM is trained through back-propagation. In experiments, the RAM is tested on the TID2008 dataset with promising performance obtained, which shows the effectiveness of the proposed RAM. Furthermore, the RAM is very efficient in the sense that a small number of glimpses are used usually in testing.

1. Introduction

1.1. Motivation and Objective

In the era of visual big data, an enormous amount of visual data is making its way to end consumers through mobile devices, social media, HDTV and IPTV, etc. Since the applications are so broad and diverse, it becomes increasingly important to improve the quality of experience for consumers. A key issue is to conduct image quality assessment (IQA) automatically such that the system is aware of the perceptual quality of degraded images and then perceptually optimizes delivery of multimedia services, and the image compression techniques are also driven to be more efficient.

There are two main schema in IQA: full-reference (FR) IQA [19, 23, 22] and general-purpose no-reference (NR) IQA [15, 5, 10, 7, 8, 16, 4, 24]. The former requires a “clean”, pristine reference image with respect to which the quality of the distorted image is assessed, and the latter takes only the distorted image to be assessed as input and thus is more applicable.

This paper focuses on general-purpose NR-IQA. Compared to FR-IQA, the challenges in NR-IQA include many factors: the “what” issue-unknown types of potential distortion (some are local, e.g., an image patch is distorted, some are global, e.g., certain additive noise affecting all the pixels in an image), the “where” issue-unknown underlying distributions of different types of distortion, and the “how” issue-unknown internal mechanism for aggregating all the information collected through time for image quality assessment (e.g., some distortion which does not affect the perceptual quality when observed in the background clutter of an image might cause critical issue when in the salient foreground objects).

In general, NR-IQA can be posed as a regression problem or a classification problem with discrete perceptual quality levels used in assessment. We focus on the regression setting. In the literature of IQA, relying on hand-crafted features in regression [5, 16, 10] and adopting end-to-end deep learning framework [7, 8] are both adopted. One important aspect of assessing perceptual quality is to respect how human visual system evaluate image quality. However, the visual attentional mechanism is largely ignored in the literature of NR-IQA. This paper addresses this
1.2. Method Overview

Fig. 1 illustrate the proposed RAM. Given an input image to be assessed, the RAM computes both the score of perceptual quality and the distortion type. It adopts the recurrent neural network (RNN) framework to capture the sequential essence of attentional scanning path in viewing an image. The RAM consists of three components:

- A glimpse sub-network extracting features around a fixation using multi-scale information. We extract three patches centered at a fixation and a convolutional neural network is used to learn useful features for quality assessment. The glimpse sub-network can resolve the “what” issue in NR-IQA stated above in the end-to-end training. It learns to map the raw image data to informative features capturing underlying distortion type and effects.
- A location sub-network selecting where to look next by sampling a stochastic node. To take into account the strong center bias in both visual attention and statistics of multimedia content, the location sub-network starts from the center of an image. The stochastic node is introduced accounting for the “where” issue in NR-IQA stated above. It learns to select where to look next based on integrated information on what it has seen so far and how well they predict such that the next selected fixation will be sufficiently informative.
- A recurrent network aggregating information along the scanning path to compute the final prediction. It captures both local information and global information in the sequential unfolding. It learns to resolve the “how” issue in a principled way. The intuitive ideas lie in two straightforward observations: what we see will determine how we assess the quality and to make better assessment we should pay attention to what are important in an image. How can we explore these two aspects jointly in learning? We introduce visual attention mechanism [17, 18] in formalizing NR-IQA. Visual attention has been studied extensively in both human and computer vision and plays important roles in many aspects of visual perception and cognition. We hypothesize that the attentional scanning path in an image should contain intrinsic information for NR-IQA and propose the recurrent attentional model (RAM). We briefly overview the proposed RAM in the following.
issue in NR-IQA stated above.

The learning of RAM is formulated under multi-task learning and REINFORCE rule using the back-propagation technique. In experiments, the RAM is tested on the TID2008 dataset with promising performance obtained, which shows the effectiveness of the proposed RAM. Furthermore, the RAM is very efficient in the sense that a small number of glimpses are used usually in testing.

2. Related Work

We briefly review the application of deep learning for NR-IQA, IQA methods taking objectness or saliency into consideration and relevant recurrent attentional models.

Neural Networks for NR-IQA: Deep learning provides an approach to learning a mapping from raw input or low-level features into image perceptual quality, which avoids delicate design on hand-crafted feature extraction. Kang et al. [7] train a CNN mapping a patch to corresponding quality score for the first time and average scores of patches as the holistic image score. They further propose a multi-task CNN [8] with fewer parameters and representation constrained by distortion type identifying shows excellent generalization performance. Patch-based model is compact and efficient but not in a global view, so combining low-level hand-crafted features with deep networks is an alternative approach for general application. Performing semi-supervised learning, Tang et al. [16] extract LBIQ features [15], wisely pre-train a Restricted Boilzman Machine with auxiliary data and fine-tune the network achieving the state-of-the-art performance. Hou et al. [5] propose a reasonable classification-based model with quality pooling treating IQA problem in a different view.

Objectness and Saliency in IQA: To the best of our knowledge, attentional models have not been applied to general-purpose NR-IQA, but semantic objectness or saliency is widely accepted. Objectness or saliency is the intrinsic property of image regions and attention is the active process of an observer. Zhang and Li [22] state that visual saliency and perceptual quality are highly related, and utilize the relationship to predict quality score. Zhang et al. [24] process IQA algorithms on object-like regions, and their method is able to be applied with other existing IQA methods to obtain better performance. Hou and Gao [4] propose a saliency-guided framework whose idea is similar to [24]. More commonly, saliency maps from computational models are exploited for post-processing, i.e., adopt saliency-weighted average score rather than average score as final prediction. Zhang et al. [25] research on different combinations between computational saliency models and IQA methods, which aims to find out strength and weakness for this approach and thoroughly make use of the capabilities.

Recurrent Attentional Models: Recently, deep learning models with attentional mechanism receive a lot of interest because it is natural, inspiring and also computationally efficient. How the model assigns attention is one of the key problems. Soft attentional models [13, 9] implement deterministic attention mechanism trained by normal backpropagation. Kuen et al. [9] realize attention mechanism through the differentiable spatial transformer [6] and recurrent connections to refine saliency map step by step. Stochastic attention happens in hard attentional models [11, 1, 13] optimized by REINFORCE algorithm [21] usually. Implementing the similar approaches, Mnih et al. [11] propose a general well-designed attentional model with recurrent neural networks for object recognition and Ba et al. [1] recognize and localize multiple objects by maximizing variational lower bound. Sorokin et al. [13] propose a soft attention mechanism designed as element-wise multiplication with importance vectors and a hard attention mechanism optimized by REINFORCE algorithm.

3. Proposed Model

In this section, we introduce the problem definition, illustrate each component of the recurrent attentional model in detail and explain how to jointly learn knowledge about distortion type, perceptual quality, and attention policy.

As illustrated in Figure 1, the proposed model is constructed with three main parts — a glimpse sub-network, a location sub-network and a recurrent network. The ultimate goal of proposed RAM is to predict the quality score $s$ given input image $x$. Start from initial focused location $l^0 = [0, 0]$ which is in the right middle of the image, in each time $t$, RAM analyzes three normalized multi-resolution patches $x^l$ sampled from $l^{t−1}$ and then predict next location $l^t$. Repeat this procedure until $T$ steps, we obtain a sequence of locations $l = \{l^1\}_{t=0}^{T−1}$ and the information on each location is aggregated into recurrent layer $h$. Then we compute the final quality score $s$ based on the state of last time $h^T$. Label $y$ which denotes distortion type of $x$ is also predicted based on $h^T$ as an auxiliary task. Except for two supervised signals for learning knowledge about image quality and distortion type, we also design a classification reward to force our RAM to efficiently sample distinct location sequence for IQA task.

3.1. the Recurrent Attentional Model for IQA

Glimpse Sub-network: The glimpse sub-network mainly resolves the “what” issue. In time step $t$, the output $g^t$ of glimpse sub-network is computed by the original image $x$, location $l^{t−1}$ and parameters $\theta_g$: $g^t = f_g(x, l^{t−1}; \theta_g)$.

We extract glimpse patches inspired by what really happens in human visual system. At each moment, individuals just focus on a very small region called fovea, where has
high-resolution representation in the middle of the retina. Regions outside of fovea named peripheral regions are perceived without details and the degradation of details grows with eccentricity [3]. Human beings adopt eye movements to make up this defect and form an efficient system. Instead of completely simulating HVS’s behavior which is also next to impossible, we simply extract three patches with the same center and normalize them into $32 \times 32$ patches. The normalized patches are combined together represented by $x^t$.

In order to analyze image content, we accept some ideas from [14] and implement our CNN with multi-scale convolution kernels. Multi-scale convolution is better for model learning in the perspective of computation. We do not apply other mature networks like the residual net because a very deep network would be easily overfitting, considering that lack of data in IQA problem. Through CNN, we got $h_{g_1} = CNN(a^t; \theta_g)$.

Futhermore, information about where we are also matters, distortion in the middle of the image is much more annoying than in the corner. We deal with location with a linear layer with ReLU activation in glimpse sub-network: $h_{g_2} = \varphi(Linear(l^t; \theta_{g_2}))$. Finally, we concatenate two hidden vectors about image content and attention location through one linear layer, then the output $g^t = \varphi(Linear(h_{g_2}, h_{g_2}; \theta_{g_2}))$ is obtained.

**Location Sub-network:** The location sub-network is introduced accounting for the “where” issue. The output $l^t$ of location sub-network is influenced by the hidden state $h^t$ and parameters $\theta_l$: $l^t = f_l(h^t; \theta_l)$. We assume that each dimension of the next location satisfies Gaussian distribution independently with the same fixed standard deviation. The real situation is much more complicated than this assumption. Considering that optimization on stochastic node converges slow, adding more variables may be not a wise choice as a start.

We sample locations stochastically in the training stage and deterministically during testing. Stochastic policy is a common resource in reinforcement learning which improves the ability of exploration of proposed model. Firstly, we predict mean of the Gaussian distribution by $\mu^t = \phi(W_{r_1}h^t + b_1)$. $\phi$ is the activation function limiting $\mu^t$ into appropriate range, and $h^t$ is hidden state in recurrent layer. Then sample next location from Gaussian distribution $a^t \sim p(\cdot|\mu^t, \sigma)$, where $\sigma$ is the standard deviation for each independent dimension of Gaussian distribution. In the end, in case of bounds overflow, we still need an activation function $l^t = \phi(a^t)$ to limit the bound of output value among $[-1, 1]$. Location bounds overflow may cause limited samples which tends to be overfitting, so we use Tanh activation rather than HardTanh in our model. Because HardTanh function has zero derivatives when the absolute value of variable is larger than one. This character results in the difficulty to draw locations back.

**Recurrent Network:** The recurrent network is learned to solve the “how” issue trying to figure out the internal mechanism of information aggregation. The subjects judge perceptual quality of images after scanning a path. It is a sequential procedure. We choose a recurrent neural network to forward information at each time step then predict distortion type and image quality at the last time step. Distortion type prediction is an auxiliary task for learning efficient scan paths for IQA. Simultaneously, our RAM should choose next location $l^t$ at time $t$ which is described previously. The recurrent layer is computed as

$$h^t = \varphi(W_{gh}g^t + W_{hh}h^{t-1} + b_h)$$

(1)

where $W_{gh}$ denotes the connection weights from the output of glimpse sub-network $g^t$ to the hidden layer $h$ and $W_{hh}$ denotes the connection of the hidden layer to itself, $b_h$ is the bias and $\varphi$ is ReLU activation function.

Distortion type classification and quality prediction are two different tasks. We solve them based on the same representation $h^T$ with multi-task learning:

$$s = \varphi(Linear(\varphi(Linear(h^T))))$$

(2)

where $s$ is predicted quality score, $\hat{s}$ and predicted distortion type $\hat{g}$ have the same representation but no shared parameters in the recurrent network, since these two tasks are conflicting to some extent. Predicting quality score is our ultimate goal and classification task not only help to generalize but also give birth to a reward for learning attention mechanism which we will discuss more in next subsection.

### 3.2. Learning

There are three terms for final loss function $L = L_{cla} + \lambda L_{reg} - \alpha J_{rein}$, where $L_{cla}$ is log softmax loss for distortion type classification and $L_{reg}$ is mean average error for quality score prediction. Both of them are trained with Back-Propagation Through Time (BPTT) algorithm. As shown in Figure 2, black arrows denote the forward computation flow. The red arrows represent the backpropagation flow based on $L_{cla}$ and $L_{reg}$, since gradients flow for there two terms are exactly the same except for the last two layers.

Even though data forward from location sub-network to glimpse sub-network based on the location, we do not back propagate gradients through that connection, so we use dash lines to connect locations.

Noticing that $\theta_l$, parameters of location sub-network, is not able to be updated through straight BPTT algorithm because there is one stochastic node in location sub-network. We use REINFORCE rule [21] to optimize reinforce term $J_{rein}$ and tune parameters in location sub-network. In the perspective of reinforcement learning, the hidden layer in
which means that if classification is right, our model tune parameters to make the mean of Gaussian close to successfully sampled location. $J_{\text{rein}}$ is related to the output of stochastic node $\alpha^t$ rather than $l^t$, so location sub-network can try different activation function without worrying about correctness for learning. As shown in Figure 2, blue arrows represent the back-propagation flow of $J_{\text{rein}}$, parameters in location sub-network are only influenced by reinforcement loss. The model is tuned by mini batch gradient descent algorithm. We just use one sample to compute the gradients based on $J_{\text{rein}}(\theta_R)$ every iteration, namely, $M$ equals to one in each mini batch.

The recurrent layer is the last shared layer for different tasks, so it is very fragile and sensitive. However, location prediction task needs accurate and stable output, this feature also adds insult to injury. At the beginning of training stage, the model has wrong knowledge about distortion type, learning for predicting beneficial locations is complete chaos. We find that sometimes the overflow of locations is out of control seriously. We apply a small trick to make learning more stable at the very start. We watch on $\mu^t$, the input of stochastic node, at first hundreds of epochs. If we detect the sum of $\mu^t$ is larger than a threshold, we randomly reset the parameters in location sub-network, make RAM learn to assign locations again with learned knowledge about distortion and quality.

3.3. Implementation Details

For preprocessing data, we first turn RGB images into gray scale images, then apply a local contrast normalization method on data.

In glimpse sub-network, we extract three patches according to center location and scale them to $32 \times 32$. We use four layers of naive inception modules [14] to stack our CNN. Each naive inception with $3 \times 3$, $5 \times 5$ and $1 \times 1$ convolution filters with number ratio $2 : 1 : 1$, and the number of all three kinds of filters are $32-64-64-128$. Spatial pooling size is $8 \times 8$ in the last convolution layer and $2 \times 2$ for other convolution layers except for the second convolution layer. Both $h_{g_1}$ and $h_{g_2}$ are 128-dimensional vectors. We assign ReLU rectifier for all convolution and linear layers in glimpse sub-network and the recurrent network. The dimension of the recurrent layer is 128, and two hidden layers for final decision both have 256 neurons. The standard deviation $\sigma$ for locator is 0.1, a larger value may cause instability and smaller value may reduce exploration of distortion. And the number of recurrent iterations is set to be 5.

For loss function, parameter $\lambda$ for regression task is set to 1 and $\alpha$ for reinforcement loss equals to 0.01.

Training neural networks is a tiresome task. For convenient, we use a kind of adaptive gradient descent methods Adam [2] with momentum 0.9 as our optimization method. The initial learning rate is 0.001. We run programs for 1200

Figure 2: Illustration of backpropagation gradients flow, red arrows denote gradients based on supervised signals and blue arrows denote gradients by REINFORCE rule.
epochs and decay learning rate linearly to make learning rate achieve minimum learning rate 0.0001 after training 1000 epochs.

4. Experimental Results

4.1. Experimental Setup

TID2008 [12]: It consists of 25 reference images. Each reference image assigned with a Mean Opinion Score (MOS) between 0 and 9 is distorted manually into 17 different distortion types with 4 different levels, so there are in total 1700 distorted images in TID2008. There are four kinds of local distortion during all the 17 types.

Evaluation: We choose Spearman Rank Order Correlation Coefficient (SROCC) to measure the prediction monotonicity. Where \( d_i \) is the difference between the ranks of image \( i \) in the subjective and objective evaluations. So SROCC only measures consistency on the rank of the data points and ignores the relative distance.

Local contrast normalization applied in our model makes the last two distortions mean shift and contrast change intractable. And the another reason is that mean shift and contrast are more related to image aesthetics in some ways, so we just deal with the first 15 distortion types in our experiments. And considering that the last reference image is not a natural image which is designed for specific analysis, we ignore it in our experiments.

We select 60% of reference images and associated distorted images as the training set, 20% and last 20% as the validation set and testing set. We choose the model with highest SROCC in the validation set as the final model. As a reference in our method, we set reference images as score 8 rather than 9 to avoid being regarded as outliers, because the highest MOS in chosen images is just 7.5.

4.2. Evaluation on TID2008

We train images of all the 15 selected distortion types together. The overall results are presented in Table 1 and SROCC evaluation for each specific distortion type is presented in Figure 3. We compare our model with classic FR-IQA methods PSNR, SSIM [19], VSI [23] and well performed NR-IQA methods CNN [7], CNN++ [8], Tang et al.’s method [16]. Results for Tang et al.’s method [16] are copied from the original paper, because some commercial reasons they can not provide both the codes and data for semi-supervised learning. Other models are implemented by ourselves basically following the original settings. Except for proposed RAM, we also implement a multi-task CNN (CNN_MT) with similar structure compared to our RAM but it is trained and tested following the framework in [8], namely, train patch-score mapping and average scores as a prediction. Our CNN_MT has the same number of layers and almost the same number of parameters with proposed RAM. As shown in Table 1, our RAM achieves promising results in non-distortion-specific experiments. The results of our CNN_MT and RAM are comparable to the state-of-the-art Tang et al.’s method [16].

As shown in Figure 3, our RAM performs better on images with local distortion, especially for the type 14 non eccentricity pattern noise and the type 15 local block-wise distortions of different intensity. We also observe that our CNN_MT outperforms RAM in most situation but achieves
Figure 4: Location of the last glimpse for images with masked noise of four levels. Our RAM locates on or near the most salient region, around black letters, in the last glimpse.

Figure 5: Location of the last glimpse for images with local block-wise distortions of four levels. Our RAM locates on or near the blocks in the last glimpse.

Figure 6: Sampled locations for images with high frequency noise of four levels. The fixation moves out into background only in the image with most serious degradation.

|       | SROCC | Class. Acc. |
|-------|-------|-------------|
| RAM without context | 0.774 | 0.807 |
| RAM | 0.832 | 0.877 |

Table 3. SROCC and classification accuracy through comparison between with and without multi-resolution context information on TID2008.

nearly the same overall results. This is just a compromise that the methods with a global view are stronger for the task with global distortion but failed in the local situation, and our RAM performs better in the local situation but has some shortcomings for global distortion images. Suppose an extreme situation, if all the patches in each image are equally important, then the methods with a global view is definitely more stable during training and testing since it collects information from a large number of patches.

Experiments on Local Distortion Types: We also straightly train our model on data of 4 local distortion types in TID2008. Results of the comparison between CNN_MT and RAM are shown in Table 2. Not surprisingly, the patch-based method still gets high enough SROCC because SROCC measures the prediction monotonicity. RAM outperforms CNN_MT a lot if we take mean absolute error (MAE) as comparison criterion.

Learning without Context Information: Context information matters for location assignment in our framework. In proposed method, we extract three patches with sizes $32 \times 32$, $96 \times 96$ and $288 \times 288$. As a reference, we learn a RAM without context information, it operates on only one $32 \times 32$ patch and is compared with proposed model. As shown in Table 3, both quality assessment results and distortion type prediction results decline on testing dataset when learning without context information.

Classification Task: On testing set, our RAM obtains 84.7 accuracy for 16 distortion types. Images of type 16 are reference undistorted images, and if we ignore that the accuracy comes to 87.7. Considering that limited visual view in our model, it is a challenging task and results are quite positive. The confusion matrix is shown in Figure 7. Half images of type 2 are misclassified into type 1 because
they are both additive Gaussian noise while the second one is specifically operated in color components. The lower right corner of confusion matrix shows that images of the last two distortion types and reference images confuse our model, which is reasonable because several images of type 14 and type 15 are degraded in very small regions and our RAM takes greedy policy so that if it takes a wrong direction then it is likely that distortions will never be found after limited glimpses.

4.3. Attentional Locations

We illustrate some attentional results to reveal where our model is looking at. We present the locations generated from images which are selected from validation set or testing set by our RAM in Figure 4, 5 and 6.

For Figure 4 and Figure 5, we magnify the sampled patches into the bottom right corner of each image. The selected patch is masked with red square and other two scale patches are masked with white squares. Figure 4 shows the last glimpse of four images with masked noise with different intensity. Masked noise is stronger in regions with high spatial frequency. The highest spatial frequency regions should be the areas around front letters of the left “white” cap because the dark letters and light cap form a strong contrast. Our RAM locates on this most salient region for all the four distortion levels even through different scan paths.

Local block-wise distortion degrades image quality by adding some annoying blocks with different intensity. In Figure 5, we find that attention locates on blocks in the last glimpse. Especially for the last two images, this result is indeed an exciting affair.

Attentional scan paths of high frequency noise are illustrated in Figure 6. Along with the noises more intensive, we suppose that our attention may shift from regions with more semantic foreground like the house and rocks into background regions such as the sky. We find that only the image with most heavy degradation moves attention into the sky, which coincides with the meaningful supposition.

However, through observing the scan paths, we also find that sometimes the location jumps out of view, namely jumps into somewhere not in the visual field. Excluding failed cases, we think this phenomenon may be caused by several reasons: firstly, RAM truly learn some stochastic jumping strategy when no distortion found; secondly, the location prediction is inaccurate because of downsampling; thirdly, it is likely that our RAM learns some global knowledge about where to find local distortion instead of inferring from context information.

5. Conclusion

Inspired by visual attention mechanism, we propose a recurrent attentional model for general no-reference image quality assessment. Our framework predicts image perceptual quality and distortion type through a small number of glimpses. The experimental results and illustration of scan paths show the potential of attentional models for handling complex images with nonuniform distortion. In the future, we plan to accelerate learning and increase stability of the recurrent attentional model.

References

[1] J. Ba, V. Mnih, and K. Kavukcuoglu. Multiple object recognition with visual attention. arXiv preprint arXiv:1412.7755, 2014.
[2] T. Dozat. Incorporating nesterov momentum into adam.
[3] J. Freeman and E. P. Simoncelli. Metamers of the ventral stream. Nature neuroscience, 14(9):1195–1201, 2011.
[4] W. Hou and X. Gao. Saliency-guided deep framework for image quality assessment. IEEE MultiMedia, 22(2):46–55, 2015.
[5] W. Hou, X. Gao, D. Tao, and X. Li. Blind image quality assessment via deep learning. IEEE transactions on neural networks and learning systems, 26(6):1275–1286, 2015.
[6] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In Advances in Neural Information Processing Systems, pages 2017–2025, 2015.
[7] L. Kang, P. Ye, Y. Li, and D. Doermann. Convolutional neural networks for no-reference image quality assessment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1733–1740, 2014.
[8] L. Kang, P. Ye, Y. Li, and D. Doermann. Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks. In Image Processing (ICIP), 2015 IEEE International Conference on, pages 2791–2795. IEEE, 2015.
[9] J. Kuen, Z. Wang, and G. Wang. Recurrent attentional networks for saliency detection. arXiv preprint arXiv:1604.03227, 2016.
[10] Y. Li, L.-M. Po, X. Xu, L. Feng, F. Yuan, C.-H. Cheung, and K.-W. Cheung. No-reference image quality assessment with shearlet transform and deep neural networks. *Neurocomputing*, 154:94–109, 2015. 1

[11] V. Mnih, N. Heess, A. Graves, et al. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems*, pages 2204–2212, 2014. 3

[12] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti. Tid2008-a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, 10(4):30–45, 2009. 6

[13] I. Sorokin, A. Seleznev, M. Pavlov, A. Fedorov, and A. Ignatieva. Deep attention recurrent q-network. *arXiv preprint arXiv:1512.01693*, 2015. 3

[14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015. 4, 5

[15] H. Tang, N. Joshi, and A. Kapoor. Learning a blind measure of perceptual image quality. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 305–312. IEEE, 2011. 1, 3

[16] H. Tang, N. Joshi, and A. Kapoor. Blind image quality assessment using semi-supervised rectifier networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2877–2884, 2014. 1, 3, 6

[17] S. K. Ungerleider and L. G. Mechanisms of visual attention in the human cortex. *Annual review of neuroscience*, 23(1):315–341, 2000. 2

[18] W. Wang, C. Chen, Y. Wang, T. Jiang, F. Fang, and Y. Yao. Simulating human saccadic scanpaths on natural images. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 441–448. IEEE, 2011. 2

[19] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 1, 6

[20] D. Wierstra, A. Foerster, J. Peters, and J. Schmidhuber. Solving deep memory pomdps with recurrent policy gradients. In *International Conference on Artificial Neural Networks*, pages 697–706. Springer, 2007. 5

[21] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992. 1, 3, 4, 5

[22] L. Zhang and H. Li. Sr-sim: A fast and high performance iqa index based on spectral residual. In *2012 19th IEEE International Conference on Image Processing*, pages 1473–1476. IEEE, 2012. 1, 3

[23] L. Zhang, Y. Shen, and H. Li. Vsi: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 23(10):4270–4281, 2014. 1, 6

[24] P. Zhang, W. Zhou, L. Wu, and H. Li. Som: Semantic obviousness metric for image quality assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2394–2402, 2015. 1, 3

[25] W. Zhang, Y. Tian, X. Zha, and H. Liu. Benchmarking state-of-the-art visual saliency models for image quality assessment. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1090–1094. IEEE, 2016. 3