BLIND IMAGE QUALITY ASSESSMENT BY LEARNING FROM MULTIPLE ANNOTATORS

Kede Ma[⋆]*, Xuelin Liu*[†]*, Yuming Fang*[†]*, Eero P. Simoncelli*[⋆]

[⋆]Center for Neural Science / Courant Institute of Mathematical Sciences
and Howard Hughes Medical Institute, New York University
[†]School of Information Technology, Jiangxi University of Finance and Economics

ABSTRACT

Models for image quality assessment (IQA) are generally optimized and tested by comparing to human ratings, which are expensive to obtain. Here, we develop a blind IQA (BIQA) model, and a method of training it without human ratings. We first generate a large number of corrupted image pairs, and use a set of existing IQA models to identify which image of each pair has higher quality. We then train a convolutional neural network to estimate perceived image quality along with the uncertainty, optimizing for consistency with the binary labels. The reliability of each IQA annotator is also estimated during training. Experiments demonstrate that our model outperforms state-of-the-art BIQA models in terms of correlation with human ratings in existing databases, as well in group maximum differentiation (gMAD) competition.

***Index Terms*—** Blind image quality assessment, convolutional neural networks, gMAD competition

1. INTRODUCTION

Most methods for automatic image quality assessment (IQA) are "full-reference", relying on specification of the original reference image [1]. But in many practical settings, the reference image is not available (or may not even exist), necessitating the use of no-reference or blind IQA (BIQA) models. Such a model should be possible because humans are able to judge the perceptual quality of distorted visual images without comparison to any reference image. Presumably our visual systems have, through evolutionary and developmental processes, learned to preferentially represent and recognize "natural" visual images - those that are likely to arise from the physical interactions of light, surfaces, and optics. It is thus reasonable to assume that a BIQA model should compute values related to the probability of occurrence of an observed image. Existing BIQA models have been developed from models of natural scene statistics. Spatially normalized coefficients [2] and codebook-based representations [3] are two examples that have demonstrated impressive performance on common distortion types [4, 5].

In recent years, there has been a surge of interest in developing data-driven BIQA models based on convolutional neural networks (CNN). These are trained on human data (mean opinion scores - MOS), which are often insufficient to constrain the large set of model parameters (often in the order of millions). One method of compensating for the lack of training data is to start with a CNN pre-trained for object recognition [6], and refine its parameters for BIQA. Alternatively, one can constrain the parameters by training a CNN on image patches [7], but local quality generally depends on global context, and in addition, it is not obvious how to combine spatially varying local scores to obtain a single global score. Some methods have augmented training data by including scores of full-reference IQA models as additional annotators [8, 9]. Here, we further pursue this strategy.

In this paper, we develop a CNN-based BIQA model, and train it on quality scores computed from a set of existing IQA models [10], without reliance on any human data. We first generate a large number of image pairs, and use multiple IQA annotators to compute binary labels indicating which of the two images is of higher quality. We then train a CNN to compute a quality score and associated uncertainty, using a pairwise learning-to-rank algorithm [11]. The reliability of each IQA annotator and the CNN parameters are jointly optimized by maximizing their likelihood. In comparison with eight BIQA models, on data from three standard IQA databases [4, 12, 13], we find that our model achieves high correlation with human perception. We further verify the generality of our model using group maximum differentiation (gMAD) competition [14].

2. METHODS

Let $q(\boldsymbol{x})$ represent the (true) perceptual quality of image $\boldsymbol{x}$. Our method relies on a group of objective IQA models $\{q^j\}_{j=1}^M$, each of which computes an estimate of the perceived quality, $q^j(\boldsymbol{x})$. These can be blind or full-reference IQA models; in the latter case, the reference image must also be available to compute the quality score. In general, these IQA annotators provide noisy nonlinear approximations to the true perceived quality. As such, we adopt them to obtain pairwise ranking information. Specifically, we use each q^j to assign a binary label r^j to image pair $(\boldsymbol{x}, \boldsymbol{y})$, where $r^j = 1$ if $q^j(\boldsymbol{x}) > q^j(\boldsymbol{y})$ and $r^j = 0$ otherwise. Given training data consisting of labels from M IQA annotators computed on N image pairs, *i.e.*, $\{(\boldsymbol{x}_i, \boldsymbol{y}_i), r_i^1, \cdots, r_i^M\}_{i=1}^N$, our goal is to optimize two differentiable functions, $f_{\boldsymbol{w}}(\boldsymbol{x})$ and $\sigma_{\boldsymbol{w}}(\boldsymbol{x})$, parameterized by a vector $\boldsymbol{w}$, that estimate the perceptual quality and its uncertainty, respectively.

2.1. Probabilistic Formulation

To model the uncertainty, we make use of the Thurstone's model [15] and assume that the perceptual quality is Gaussian with mean $f_{\boldsymbol{w}}(\boldsymbol{x})$ and standard deviation $\sigma_{\boldsymbol{w}}(\boldsymbol{x})$. Assuming the variability of quality across images is uncorrelated, the quality difference of two images $q(\boldsymbol{x}) - q(\boldsymbol{y})$ is also Gaussian with mean $f_{\boldsymbol{w}}(\boldsymbol{x}) - f_{\boldsymbol{w}}(\boldsymbol{y})$ and standard deviation $\sqrt{\sigma_{\boldsymbol{w}}^2(\boldsymbol{x}) + \sigma_{\boldsymbol{w}}^2(\boldsymbol{y})}$. The probability that $\boldsymbol{x}$ has higher quality than $\boldsymbol{y}$ (*i.e.*, the probability of $r = 1$) is then

$$\begin{aligned}\Pr(r = 1|\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{w}) &= \Pr(q(\boldsymbol{x}) > q(\boldsymbol{y})|\boldsymbol{w}) \\ &= \Phi\left(\frac{f_{\boldsymbol{w}}(\boldsymbol{x}) - f_{\boldsymbol{w}}(\boldsymbol{y})}{\sqrt{\sigma_{\boldsymbol{w}}^2(\boldsymbol{x}) + \sigma_{\boldsymbol{w}}^2(\boldsymbol{y})}}\right), \end{aligned} \quad (1)$$

where $\Phi(\cdot)$ is the standard Normal cumulative distribution function.

To model the reliabilities of the IQA annotators, we assume they do not depend on the input image $\boldsymbol{x}$, and can be characterized entirely by probabilities of correct answers (known as "hits" and "correct rejections" in signal detection theory). Specifically, if the ground-truth label $r = 1$, the hit rate of the j-th IQA annotator q^j is defined as

$$\alpha^j = \Pr(r^j = 1 | r = 1). \tag{2}$$

Similarly, if $r = 0$, the correct rejection rate is defined as

$$\beta^j = \Pr(r^j = 0 | r = 0). \tag{3}$$

The parameters $\{\alpha^j, \beta^j\}$ are estimated along with the model parameters $\boldsymbol{w}$.

2.2. Maximum Likelihood Model Estimation

Given the assumption that the variability across training pairs is uncorrelated, we factorized the likelihood function of the full set of parameters $\{\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}\}$ as

$$\Pr(\{\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{r}_i\} | \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^{N} \Pr(r_i^1, \cdots, r_i^M | \boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}). \tag{4}$$

Assuming r_i^j is conditionally independent of everything else given α^j, β^j and the ground-truth label r_i, we decomposed the probabilities in the likelihood by conditioning on r_i

$$\begin{aligned} \Pr(r_i^1, \cdots, r_i^M | \boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \\ \Pr(r_i^1, \cdots, r_i^M | r_i = 1, \boldsymbol{\alpha}) \Pr(r_i = 1 | \boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{w}) + \\ \Pr(r_i^1, \cdots, r_i^M | r_i = 0, \boldsymbol{\beta}) \Pr(r_i = 0 | \boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{w}). \end{aligned} \tag{5}$$

The distribution of noisy estimates provided by the set of IQA annotators may be written

$$\begin{aligned} \Pr(r_i^1, \cdots, r_i^M | r_i = 1, \boldsymbol{\alpha}) &= \prod_{j=1}^{M} \Pr(r_i^j | r_i = 1, \alpha^j) \\ &= \prod_{j=1}^{M} (\alpha^j)^{r_i^j} (1 - \alpha^j)^{1 - r_i^j}, \end{aligned} \tag{6}$$

and

$$\Pr(r_i^1, \cdots, r_i^M | r_i = 0, \boldsymbol{\beta}) = \prod_{j=1}^{M} (\beta^j)^{1 - r_i^j} (1 - \beta^j)^{r_i^j}. \tag{7}$$

Denoting (1), (6) and (7) by $p(\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{w})$, $a(\boldsymbol{r}_i, \boldsymbol{\alpha})$ and $b(\boldsymbol{r}_i, \boldsymbol{\beta})$, respectively, and substituting them into (4), we obtained a likelihood function for the parameters

$$\begin{aligned} \Pr(\{\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{r}_i\} | \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^{N} [a(\boldsymbol{r}_i, \boldsymbol{\alpha}) p(\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{w}) \\ + b(\boldsymbol{r}_i, \boldsymbol{\beta})(1 - p(\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{w}))]. \end{aligned} \tag{8}$$

We maximized the log of this function using stochastic gradient descent to obtain the optimal parameters $\{\hat{\boldsymbol{w}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}\}$.

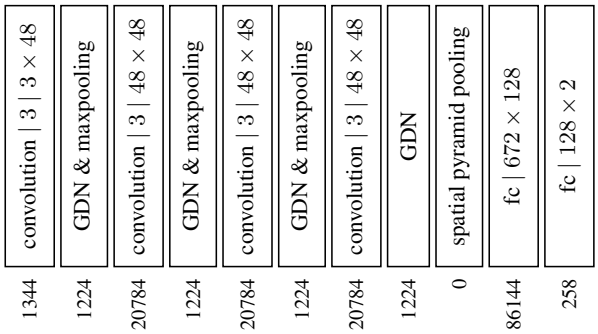

Fig. 1. The network architecture of our BIQA model. The parameterization of the convolution layers is denoted as "filter support | input channel × output channel". The number of parameters for each layer is given at the bottom, yielding a total of $154,994$.

2.3. CNN Architecture

We implemented functions $f_{\boldsymbol{w}}(\boldsymbol{x})$ and $\sigma_{\boldsymbol{w}}(\boldsymbol{x})$ with a CNN, whose architecture is specified in Fig. 1. The network consists of four stages of convolution and generalized divisive normalization (GDN) [16], a nonlinearity that is inspired by models of sensory neurons [17] and has proven effective in density modeling [18, 19], image compression [16], and IQA [20, 21]. The number of convolution filters and their support are fixed to 48 and 3×3 for all stages, respectively. The filter weights are included in the parameter vector $\boldsymbol{w}$. The GDN transform is defined as

$$v_i = \frac{u_i}{\left(\omega_i + \sum_j \gamma_{ij} u_j^2\right)^{\frac{1}{2}}}, \tag{9}$$

where ω_i and γ_{ij} are also included in $\boldsymbol{w}$. The GDN responses for the first three stages are max-pooled over 2×2 blocks. At the final GDN layer, we used spatial pyramid pooling [22] to summarize the spatial statistics and generate a fixed-length representation regardless of image size. Last, we appended two fully connected layers with a halfwave-rectification (ReLU) nonlinearity in between, to generate the values $f_{\boldsymbol{w}}(\boldsymbol{x})$ and $\sigma_{\boldsymbol{w}}(\boldsymbol{x})$.

3. RESULTS

3.1. Model Training

We assembled a training dataset based on reference images from the Waterloo Exploration Database [5], which contains $4,744$ high-quality natural images with diverse content. We simulated nine distortion types, each at five levels of severity - Gaussian blur, additive Gaussian noise, additive pink noise, JPEG compression, JPEG2000 compression, contrast change, color quantization with dithering, over-exposure, and under-exposure - yielding a total of $213,480$ distorted images. To generate training labels, we used six full-reference IQA models: SSIM [1], MS-SSIM [23], VIF [24], MAD [12], VSI [25] and NLPD [26]; and three BIQA models: NIQE [27], ILNIQE [28] and dipIQ [9]. Implementations of all nine models were obtained from the respective authors, and parameters were set independently of the test IQA databases used in Section 3.2. We generated four types of training pairs $(\boldsymbol{x}, \boldsymbol{y})$: (1) same reference image and distortion type, with different distortion levels; (2) same reference image, but different distortion types and levels; (3) two

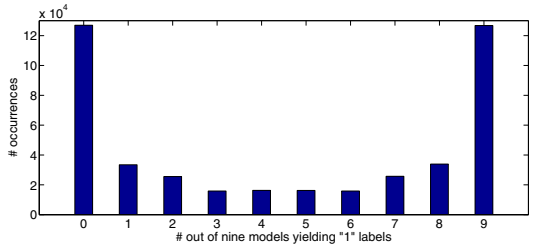

Fig. 2. Histogram of the number of IQA models assigning a "1" to an image pair in the training dataset.

different reference images, distortion types and levels; (4) two different reference images, with one undistorted. Altogether, we obtained more than $400,000$ image pairs. Consistency of the IQA annotations over all pairs is summarized by the histogram in Fig. 2, which shows the number of models that chose the first image in each pair as the higher quality one. Overall, the nine IQA annotators are in reasonable agreement, with complete consistency on roughly 60% of the image pairs.

Training was performed by jointly maximizing the log of the likelihood function (Eq. (8)), using stochastic gradient descent on mini-batches randomly sampled from the training data. We used the Adam stochastic optimization package [29] with a mini-batch size of 16, and learning rates for $\boldsymbol{w}$ and $\{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$ set to 10^{-4} and 10^{-3}, respectively. The parameters $\boldsymbol{\omega}$ and $\boldsymbol{\gamma}$ in GDN were projected to nonnegative values after each step. Additionally, we forced $\boldsymbol{\gamma}$ to be symmetric as suggested in [16]. We trained the network to predict the log variance, $s_{\boldsymbol{w}}(\boldsymbol{x}) = \log \sigma_{\boldsymbol{w}}(\boldsymbol{x})^2$, so as to avoid a potential division by zero in Eq. (1). The learning took roughly one day, running on an NVIDIA GeForce GTX 1080 Ti machine, when the epoch number was set to eight. In all experiments, we tested on images of original size.

3.2. Model Performance

We first examined the correlation of the learned model with human ratings from three standard IQA databases: LIVE [4], CSIQ [12], and TID2013 [13]. We computed both the Spearman rank correlation coefficient (SRCC) and the Pearson linear correlation coefficient (PLCC). For the latter, we fit a four-parameter monotonic function $\hat{f} = (\eta_1 - \eta_2)/(1 + \exp(-(f - \eta_3)/|\eta_4|)) + \eta_2$ to linearize the predictions. Table 1 provides comparisons of our model with five BIQA models (none of which was optimized for data in any of the three IQA databases), two state-of-the-art full-reference IQA models, and three CNN-based BIQA models that were trained on the full TID2013 database.

Several aspects of the results are worth noting. First, our method achieves significant improvements on CSIQ [12] and TID2013 [13] over the five competing BIQA models, three of which have been used to provide noisy labels during training. This indicates the effectiveness of our learning scheme, which takes into account the reliability of each annotator. Second, with a substantially smaller number of model parameters, our method exhibits better generalizability on LIVE and CSIQ than recent CNN-based BIQA models deepIQA [7] and DB-CNN [30]. We believe this improvement arises because our method is trained on a much larger database with a wider variety of image content. Third, our method is comparable to the two full-reference methods MS-SSIM [23] and NLPD [26] on the CSIQ database, but underperforms them on the other two. Performance on TID2013 is particularly weak, presumably because that database includes fifteen more distortion types than those in the training set.

Table 1. Correlation (SRCC and PLCC) of IQA models against human ratings from three different IQA databases. Top section contains two state-of-the-art full-reference models. Second section contains three CNN-based BIQA models trained on data in TID2013. Third section contains five BIQA models. The top two correlations are highlighted in boldface.

SRCC	LIVE	CSIQ	TID2013
MS-SSIM [23]	0.951	0.913	0.787
NLPD [26]	0.941	0.921	0.800
deepIQA [7]	0.814	0.688	—
MEON [21]	0.792	0.655	—
DB-CNN [30]	0.903	0.769	—
QAC [31]	0.868	0.490	0.372
NIQE [27]	0.906	0.627	0.312
ILNIQE [28]	0.898	**0.815**	**0.494**
BLISS [8]	0.908	0.602	0.460
dipIQ [9]	**0.938**	0.527	0.438
Proposed	**0.919**	**0.915**	**0.578**
PLCC	LIVE	CSIQ	TID2013
MS-SSIM	0.949	0.899	0.833
NLPD	0.941	0.924	0.830
deepIQA	0.837	0.745	—
MEON	0.787	0.739	—
DB-CNN	0.895	0.813	—
QAC	0.863	0.708	0.437
NIQE	0.904	0.716	0.398
ILNIQE	0.903	**0.854**	**0.589**
BLISS	0.905	0.750	0.557
dipIQ	**0.935**	0.779	0.477
Proposed	**0.917**	**0.926**	**0.640**

We also performed a more direct comparison of our model against two other BIQA methods using the gMAD competition method [14]. gMAD is a discrete instantiation of the maximum differentiation (MAD) competition method [32] that aims to falsify a model by synthesizing the strongest possible counterexamples. gMAD performs a discrete optimization over a fixed set of examples, seeking pairs of images that are of nearly equal quality according to one model, while being as different as possible according to the other. To build the dictionary for gMAD, we collected $1,000$ high-quality images (none from the training dataset) and generated $45,000$ distorted images by applying the nine distortions with five levels (described in Section 3.1). We first compared our method with ILNIQE [28] (the current best BIQA model) in Fig. 3. The images in the first row exhibit similar perceptual quality (in agreement with our method) and those in the second have drastically different perceptual quality (in disagreement with ILNIQE), suggesting that our method is better able to generalize to novel content than ILNIQE. A similar result is obtained in comparison with DB-CNN [30] (the most recent CNN-based BIQA model), as shown in Fig. 4. Qualitatively, we find these results to be consistent across all quality levels.

We examined the learned uncertainty $\sigma_{\boldsymbol{w}}(\boldsymbol{x})$ as a function of $f_{\boldsymbol{w}}(\boldsymbol{x})$ on the LIVE database [4] (Fig. 5). Overall, $\sigma_{\boldsymbol{w}}$ is relatively small compared to the range of predicted quality scores $f_{\boldsymbol{w}}$. Moreover, $\sigma_{\boldsymbol{w}}$ increases with decreasing $f_{\boldsymbol{w}}$ regardless of image content,

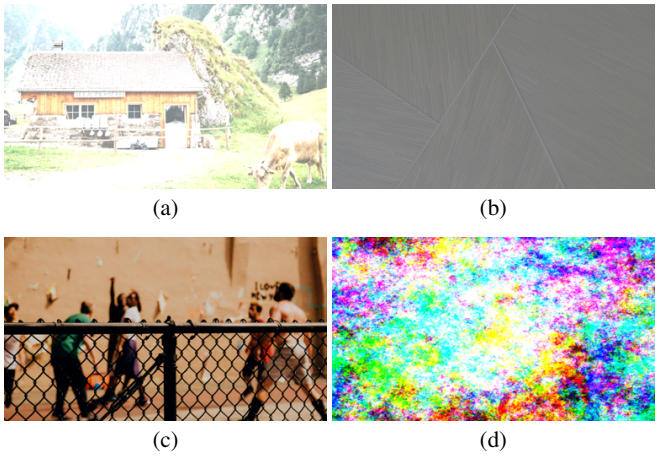

Fig. 3. gMAD competition between our method and ILNIQE [28]. (a/b) Best/worst quality images according to ILNIQE (respectively), with near-identical quality reported by our model. (c/d) Best/worst quality images according to our model with near-identical quality reported by ILNIQE. Images are cropped for improved visibility.

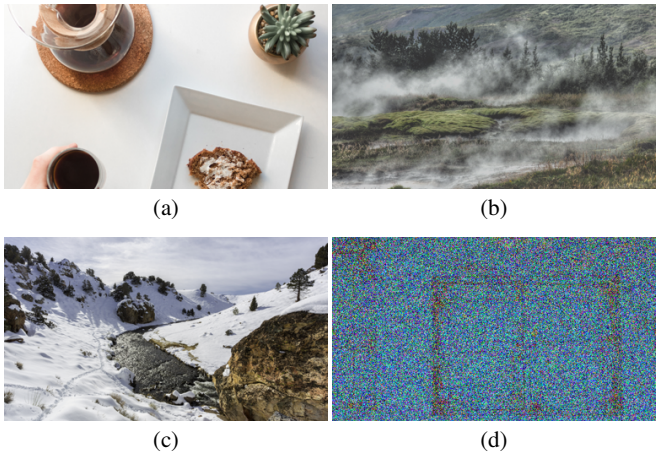

Fig. 4. gMAD competition between our method and DB-CNN [30]. (a/b) Best/worst quality images according to DB-CNN (respectively), with near-identical quality reported by our model. (c/d) Best/worst quality images according to our model with near-identical quality reported by DB-CNN. Images are cropped for improved visibility.

Table 2. The learned hit rate ($\boldsymbol{\alpha}$) and correct rejection rate ($\boldsymbol{\beta}$) of the nine IQA annotators.

$\boldsymbol{\alpha}$	SSIM	MS-SSIM	VIF	MAD	VSI
	0.933	0.958	0.921	0.952	0.959
	NLPD	NIQE	ILNIQE	dipIQ	
	0.945	0.805	0.886	0.735	
$\boldsymbol{\beta}$	SSIM	MS-SSIM	VIF	MAD	VSI
	0.935	0.958	0.918	0.954	0.958
	NLPD	NIQE	ILNIQE	dipIQ	
	0.944	0.800	0.884	0.730	

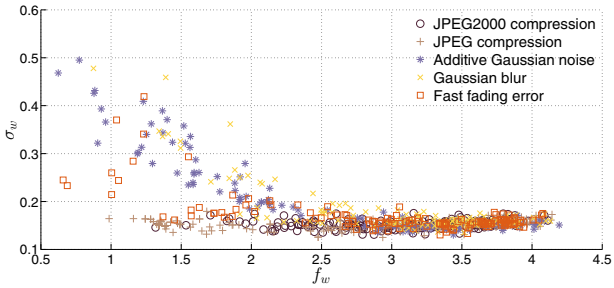

Fig. 5. Scatter plot of the learned uncertainty $\sigma_{\boldsymbol{w}}$ as a function of $f_{\boldsymbol{w}}$ on the LIVE database [4].

for some distortion types. This seems counterintuitive, because humans tend to assess images at the two ends of the quality range with higher confidence than images in the mid-quality range. However, from the (full-reference) IQA models' perspective, they are in closer agreement on higher-quality images (with the help of the reference images) and may give dramatically different penalties to lower-quality images. This discrepancy leads to increasing uncertainty, which is reflected in our model. From the figure, the chosen IQA models appear to perform more consistently for JPEG and JPEG2000 compression than the other three distortion types.

The learned hit rate $\boldsymbol{\alpha}$ and correct rejection rate $\boldsymbol{\beta}$ of the nine IQA annotators are shown in Table 2. As expected, our method trusts full-reference IQA models more than BIQA ones. In other words, the more accurate an IQA model is in predicting image quality, the more influence it has on the learning process. In addition, the learned $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are highly correlated with each other, suggesting that they could be replaced by a single parameter vector during training.

4. CONCLUSION

We have presented a CNN-based BIQA model by training on noisy labels from multiple IQA models. We jointly optimized the parameters of the network and IQA annotators for quality prediction and uncertainty estimation. When trained on a large number of image pairs, the optimized model performs favorably against current BIQA models, outperforming even those trained on human ratings, and generalizes reasonably to novel content and distortion types.

Although the current method benefits from large-scale training data, it is distortion-aware in the sense that a set of distortion types need to be specified when generating training data. As suggested by the poor performance on the TID2013 database, this can limit the generalizability of the model to novel distortion types. We expect that performance on TID2013 could be improved by incorporating additional distortion types into the training set.

A more interesting avenue for future work lies in the development of BIQA frameworks that are distortion-unaware, and thus must rely solely on knowledge of the appearance of natural undistorted images. From a probabilistic perspective, this means that the BIQA method should embody a prior probability model for natural images, and perhaps even that the BIQA values should be monotonically related to such a probability model. Since probability models for natural images form a cornerstone for most classical problems in image processing and visual analysis, a generalized BIQA model could lead to improvements in a wide variety of applications.

5. REFERENCES

[1] 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.

[2] 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.

[3] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *IEEE Conf. Comput. Vis. and Pattern Recognit.*, 2012, pp. 1098–1105.

[4] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[5] K. Ma, Z. Duanmu, Q. Wu, Z. Wang, H. Yong, H. Li, and L. Zhang, "Waterloo Exploration Database: New challenges for image quality assessment models," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 1004–1016, Feb. 2017.

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

[7] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, Jan. 2018.

[8] P. Ye, J. Kumar, and D. Doermann, "Beyond human opinion scores: Blind image quality assessment based on synthetic scores," in *IEEE Conf. Comput. Vis. and Pattern Recognit.*, 2014, pp. 4241–4248.

[9] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "dipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3951–3964, Aug. 2017.

[10] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds," *J. Mach. Learn. Res.*, vol. 11, pp. 1297–1322, Apr. 2010.

[11] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *Int. Conf. Mach. Learn.*, 2005, pp. 89–96.

[12] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imaging*, vol. 19, no. 1, pp. 1–21, Jan. 2010.

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

[14] K. Ma, Z. Duanmu, Z. Wang, Q. Wu, W. Liu, H. Yong, H. Li, and L. Zhang, "Group maximum differentiation competition: Model comparison with few samples," *IEEE Trans. Pattern. Anal. Mach. Intell.*, to appear.

[15] L. L. Thurstone, "A law of comparative judgment," *Psychol. Rev.*, vol. 34, no. 4, pp. 273–286, Jul. 1927.

[16] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in *Int. Conf. Learn. Represent.*, 2017.

[17] M. Carandini and D. J. Heeger, "Normalization as a canonical neural computation," *Nat. Rev. Neurosci.*, vol. 13, no. 1, pp. 51–62, Jan. 2012.

[18] S. Lyu and E. P. Simoncelli, "Nonlinear extraction of 'independent components' of natural images using radial Gaussianization," *Neural Comput.*, vol. 21, no. 6, pp. 1485–1519, Jun. 2009.

[19] J. Ballé, V. Laparra, and E. P. Simoncelli, "Density modeling of images using a generalized normalization transformation," in *Int. Conf. Learn. Represent.*, 2016.

[20] V. Laparra, J. Ballé, A. Berardino, and E. P. Simoncelli, "Perceptual image quality assessment using a normalized Laplacian pyramid," *Electron. Imaging*, vol. 2016, no. 16, pp. 1–6, Feb. 2016.

[21] 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.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Eur. Conf. Comput. Vis.*, 2014, pp. 346–361.

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

[24] 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.

[25] L. Zhang, Y. Shen, and H. Li, "VSI: A visual saliency-induced index for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4270–4281, Oct. 2014.

[26] V. Laparra, A. Berardino, J. Ballé, and E. P. Simoncelli, "Perceptually optimized image rendering," *J. of Opt. Soc. of Am. A*, vol. 34, no. 9, pp. 1511–1525, Sep. 2017.

[27] 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.

[28] 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.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[30] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Deep bilinear pooling for blind image quality assessment," *IEEE Trans. Circuits and Syst. Video Technol.*, to appear.

[31] W. Xue, L. Zhang, and X. Mou, "Learning without human scores for blind image quality assessment," in *IEEE Conf. Comput. Vis. and Pattern Recognit.*, 2013, pp. 995–1002.

[32] Z. Wang and E. P. Simoncelli, "Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities," *J. Vis.*, vol. 8, no. 12, pp. 1–13, Sep. 2008.