Self-supervised Deep Multiple Choice Learning Network for Blind Image Quality Assessment

Kwan-Yee Lin
linjunyi@pku.edu.cn

Department of Information Science,
School of Mathematical Sciences,
Peking University

Guanxiang Wang
gxwang@math.pku.edu.cn

Department of Mathematics,
School of Mathematical Sciences,
Peking University

Abstract

Blind image quality assessment (BIQA), which is the problem of predicting the perceptual quality of an image with unknown distortion information and no access to the reference image, has been a longstanding task in the low-level computer vision community. In recent years, various methods have been proposed to leverage the powerful representation ability of deep neural networks to solve the problem. However, the extreme lack of labeled training samples makes it is still a challenge for deep BIQA models to make robust predictions.

In this work, we propose a novel method to address the problem by simplifying the solution space in a self-supervised manner. This idea is achieved by generating multiple quality hypotheses, and re-filtering subsequently with an auxiliary decision mechanism. The two-stage work is done through a new convolutional network architecture with two *interacting coupled* sub-networks, *i.e*, a multiple hypotheses network (MH-Net) and an election network (E-Net). Our approach achieves the state-of-the-art performance on the well-known benchmarks with real-time and training from scratch properties. Moreover, we demonstrate the effectiveness and scalability of our method with insightful analyses.

1 Introduction

In this visual-informational explosion era, images have become an important medium of communication in our daily life. While, the generation processes of images are always in company with quality degradation[1], which greatly affects user visual experience in a negative aspect. Also, the quality degradation problem would arise numerous difficulties for image/video processing and computer visual applications, like image super-resolution, image restoration, and person re-identification. Thus, the ability to automatically monitor the perceptual quality of images in a way that coincides with human judgment is of fundamental importance in many fields.

[1] An image could be distorted in any stages in the whole process of its lifecycle from acquisition to storage, and therefore will suffer quality degradation from diverse distortions, like various blur, compression artifact, transmission errors, *etc*.

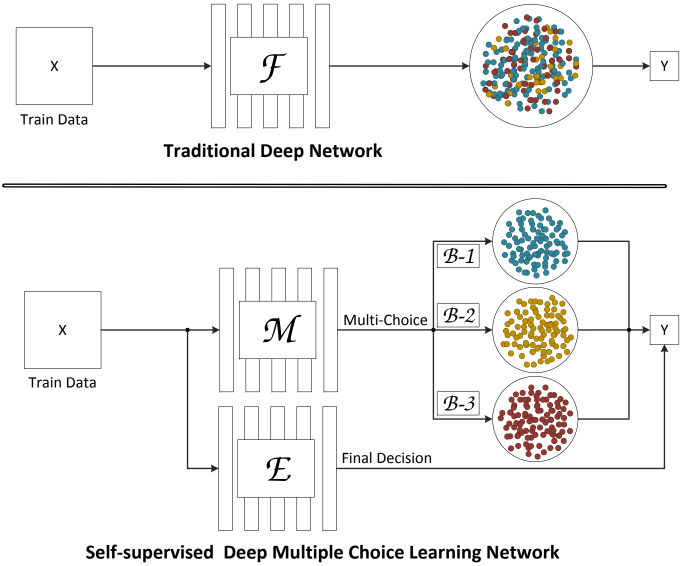


Figure 1: An illustration of the difference between our approach and traditional deep network. Suppose that the raw train data can be split into m subsets (we take $m = 3$ for example) according to their distribution in high-dimensional space. Traditional deep network treats the train data and its corresponding solution space holistically, which leads to difficulty of optimization. In contrast, our approach divides the train data distribution automatically and thus simplify the corresponding solution space (hypothesis space) by m split branches. In $\mathcal{M}$, parameters are optimized by all train data, while in $\mathcal{B}$-*1* to $\mathcal{B}$-*3*, parameters are optimized by automatically divided subsets respectively. In addition, a coupled auxiliary network is proposed to select the final prediction given m candidates. (Best viewed in color)

In the literature of image quality assessment (IQA), many full-reference (FR) methods, such as SSIM [20], FSIM [23] and DeepQA [7], have achieved excellent results with the prior of reference images[2] to capture the difference information. However, the impossible process to obtain ideal reference information in most cases limits the feasibility of FR methods in practical applications. In contrast, Blind image quality assessment (BIQA), which takes only the distorted image to be evaluated as input without any additional information, is more realistic and therefore receives substantial attention in recent years. Nevertheless, BIQA faces unique challenges.

One of the challenges is that BIQA is highly ill-posed for the absence of reference information. In addition, the ground-truth quality values are labeled by human based on visual perception on distorted images, which are subjective and related to image context and structure. Thus, the ill-posed definition forces BIQA methods to possess more powerful feature representation ability to make robust predictions.

A straightforward way to ease the ill-posed nature of BIQA is to utilize the powerful feature representation ability of deep neural networks (DNNs). However, followed by the second and the critical challenge, the insufficient training data[3] limits the effectiveness of DNNs. Recent methods, *e.g.,* [5], [1] and [8], attempt to address this problem through exploiting various multi-task and data augmentation strategies with extra annotated rank-

[2] The original high-resolution image without any distortion is referred to *reference image*.

[3] The subjective scoring of images quality is very expensive and cumbersome. Thus, even the largest subject-rated IQA database (TID2013) only includes 3000 annotations.

ing, proxy quality scores, or distortion information sophisticatedly, which are unavailable in practical BIQA applications, and hence lack of feasibility for modeling unknown distortion types.

Instead, might there be a way to learn the robust representation with less effort? If the solution space could be simplified, the existing training samples may become sufficient to the problem. Intuitively, we can simplify the solution space by dividing the dataset into sub-datasets and extracting features from several deep networks. However, dividing the dataset artificially will introduce large bias and therefore may lead the results into sub-optimal values. Besides, the overfitting problem will be aggravated due to smaller sub-datasets.

Motivated by these observations, we present a novel deep neural network architecture, named Self-supervised Deep Multiple Choice Learning Network (SDMCL-NET), which consists of two closely coupled sub-networks, to address above issues. The MH-Net divides the input dataset in an adaptive manner based on their underlying similarity to generate multiple competitive hypotheses, and addresses the overfitting problem with its main body weight-sharing and non-interactive branches learning architecture. The E-Net, trained in a self-supervised manner, functions as a "filter" to determine the final hypothesis with the best quality estimation, which is generated by one of these MN-Net originated branches. Comparing with the existing approaches, the proposed framework has features that require neither any designed availability of prior information nor extra data for the sufficiency of BIQA hypothesis space description. Figure. 1 illustrates the key idea of the proposed network and its difference with a traditional deep network.

The main contributions are three folds:

1) We propose a multiple hypotheses network (MH-Net) for BIQA, which exploits inherent similarity and diversity in training data as a guidance for dividing the dataset into several subsets automatically, to simplify the hypothesis space (*i.e*, solution space) without overfitting.

2) We introduce a coupled election network (E-Net) to help MH-Net make a final decision among several candidates. A novel self-supervised manner is proposed to train E-Net with pseudo-labels obtained from the existing data without any further annotation.

3) A comprehensive evaluation of our methods on four challenging datasets shows the superior performance over most of state-of-the-art methods. Also, The light-weight CNN-based architecture makes our framework an efficient real-time solution.

2 Related Work

Blind Image Quality Assessment. In recent years, advances in Deep Neural Networks have motivated researchers to develop models that could, on the one hand, utilize its great feature representation property, and on the other hand, avoid the obstacle of its optimization caused by limited training samples, to solve BIQA problem. For instance, Kang *et al.* propose a shallow CNN model, trained on patch-wise inputs to perform BIQA [5]. This approach is refined to the multi-task CNN [6], where the neural network learns both distortion type and quality scores simultaneously. According to a specific dataset, Ma *et al.* propose generating a mass of ranked image pairs to train the deep BIQA model [11]. Kim *et al.* follow the FR-IQA behavior using the local quality maps predicted by state-of-the-art FR-method as intermediate targets for conventional neural networks to help solve the task[8].

Multiple Outputs Structured Prediction. Existing works on learning multiple outputs prediction architecture from a single model mainly focus on probabilistic models, *e.g.*, [3] and [4]. These two works pose the multi-output prediction as a "Multiple Choice Learning" (MCL) paradigm. MCL explicitly minimizes oracle loss over the outputs of an ensemble, makes the outputs explain different but relevant parts of the data. Stefan Lee *et al.* extend MCL to the deep architecture over an ensemble of deep networks to solve image classification task [10]. However, since these three works all select final prediction with respect to an *Oracle*, it is hard to directly apply them to BIQA, or any task that required to be self-contained. Thus, we transfer deep multi-choice learning into BIQA task on a general deep architecture with developing the architecture into a self-supervised prediction mechanism. Instead of interacting with Oracle, an auxiliary network (*i.e.* E-Net) is proposed in this work to select the final regression result automatically.

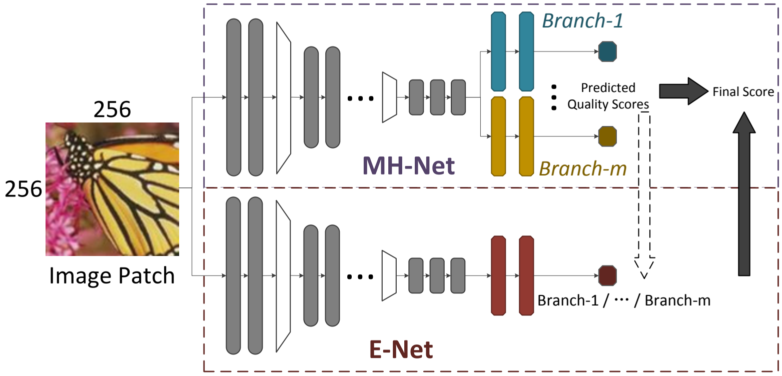


Figure 2: The architecture of our proposed SDMCL-NET. On the training process, MH-Net with m branches is trained end-to-end. Then, MH-Net is able to give m predictions for each image patch. These m predictions along with its ground truth label are used to generate the ground truth class label to facilitate the training of E-Net (white dotted arrow). In other words, the E-Net training is supervised by the trained MH-Net. On the testing process, each image patch will pass simultaneously through MH-Net and E-Net, and E-Net will make the final decision from the m candidates given by MH-Net (black solid arrow). (Best viewed in color)

3 The Proposed Algorithm

In this section, we describe the proposed blind image quality assessment system that learns from both annotated data with plain supervised strategy and pseudo-labeled information with self-supervised manner. As illustrate in Figure. 2, the SDMCL-Net architecture comprises two interacting coupled sub-networks (*i.e*, MH-Net and E-Net) to interactively and collectively provide a precise quality assessment. We detail the modules and the training procedures as follows.

3.1 Network Architecture

Multiple Hypotheses Network. The multiple hypothesis network (MH-Net) is designed to learn multiple potential values of the perceptual quality for an input image.

Given an input distorted image x, the ordinary practice for most previous methods is to forward the x to a feature extraction module to obtain a feature representation that encodes distortion information of the input image, and then infer the perceptual score s with a regression module. Specifically, under the framework of DNNs, the common process is to use a stack of fully-connected (FC) layers over the feature maps from previous convolutional layers, which could be formulated as:

$$s = \mathcal{DNN}(x;\theta) = r(f(x;\theta_f);\theta_r), \tag{1}$$

where θ_f and θ_r denotes the parameters of convolutional layers and FC layers respectively.

However, as we discussed in the introduction, the common *single* regression module will easily lead to overfitting problem, due to the mismatch of training data and model capacity. One feasible way to ease this problem is to utilize MCL to partition the hypothesis space. Many works on other research fields (*e.g,* [26] [17] and [10]) usually choose model-level structure (*i.e,* diversity between member models as a means to improve performance or combine ensemble members under a joint loss) to achieve the effects of MCL. While, this kind of structure will aggravate the over-fitting problem to deep BIQA due to incorporating large magnitude of parameters. Thus, we adopt *branch-level* ensemble strategy to make multiple predictions, which applies ensemble strategy in *FC* stage by forking after the feature vector of last convolutional layer into multiple branches of FC layers, namely *branch-k* (k ranges from 1 to *m*). Such a structure is more effective than normal CNN ensemble, since the parameters will be decreased thanks to sharing weights of previous feature extraction layers. Therefore, Eq. 1 can be modified to:

$$\vec{s} = [s_1, \cdots, s_m]^\top = \{r^k(f(x;\theta_f);\theta_r^k)\}_{k=1}^m, \tag{2}$$

where $r^k(\cdot)$ denotes the mapping function of *k-th* branch, $\vec{s}$ is a vector with s_k represented as the predicted score of *k-th* branch. The proposed MH-Net ensures each branch with the properties of *similarity* and *diversity*, on account of input vector sharing but individual parameters learning, and no interactive connection.

Election Network. To obtain the final prediction based on *m* candidates, a selection mechanism is required to be incorporated into the framework. Simply applying voting methods like averaging, minimization or maximization could not satisfy the complicated hypothesis space. To this end, a novel auxiliary mechanism is proposed to decide the best solution among the multiple hypotheses.

Since the MH-Net separates the solution space into several subspaces, the auxiliary mechanism is expected to determine which subspace is more likely to contain the ground-truth solution, and select the prediction in that subspace to be the final prediction score, given a distorted image and its *m* hypotheses. Thus, E-Net is trained to *discriminate* the best quality estimation, given the input image and the hypotheses of it from MH-Net, which could be represented as:

$$\hat{y} = \mathcal{E}(x,\vec{s};\omega) = \mathcal{E}(x,\{r^k(f(x;\theta_f);\theta_r^k)\}_{k=1}^m;\omega), \tag{3}$$

where $\mathcal{E}$ denotes the mapping function of E-Net, ω denotes the corresponding parameters. The formula indicates the precision of the final prediction is highly related to the classification accuracy of E-Net. Meanwhile, as we mentioned in the related work, BIQA should be a self-contained model without extra human intervention and annotation. Thus, seeking a learning strategy that could satisfy maintaining the performance of E-Net while only leveraging existing data is crucial to our framework.

To solve this problem, we propose to learn E-Net in a *self-supervised* manner. For E-Net, rather than predicting labels extra annotated by humans, it predicts *pseudo-labels* computed from the existing data itself. We will discuss the self-supervised manner in detail in Section 3.2.

3.2 Training Procedure

Learning for MH-Net. The SGD winner-take-all learning scheme could be used to optimize the MH-Net with batch updates in stochastic gradient descent. Given a training set of distorted image and ground truth quality score pairs $B = \{(x_i, y_i) | x_i \in X, y_i \in Y\}$, our goal is to learn a function $\mathcal{M} : X \to Y^m$ which maps each input to m outputs. The loss over a batch B is

$$\mathcal{L}_{MN}(B) = \sum_{i=1}^{n} \min l(y_i, \mathcal{M}(x_i)) = \sum_{i=1}^{n} \min_{k \in [1,\cdots,m]} l(y_i, r^k(f(x_i))), \tag{4}$$

where n represents the batch size. Thus, the objective function can be written as

$$\begin{aligned} &\arg\min_{r^k, f, w_{i,k}} \sum_{i=1}^{n} \sum_{k=1}^{m} w_{i,k} l(y_i, r^k((f(x_i)))) \\ &s.t. \quad \textstyle\sum_{k=1}^{m} w_{i,k} = 1, \quad w_{i,k} \in \{0, 1\} \end{aligned} \tag{5}$$

We adopt L1 loss as the loss function in our implementation. The SGD winner-take-all learning scheme ensures only one (the prediction of which is closest to the ground truth) of the branches is optimized each iteration, and the other's backward gradient are blocked. Consequently, at least one of the branches will produce a precise prediction and larger hypothesis space will be covered due to the competition among m branches.

During the experiments, we found only utilizing the SGD winner-take-all strategy will easily lead the optimization process to an extreme situation. Thus, a constraint is added to the initialization period. A guidance mechanism is further introduced to help increase diversity among the branches. Thus, MH-Net could divide the hypothesis space in a better way.

The constraint to Eq.5 is formulated as:

$$\begin{cases} \arg\min_{r^k, f, w_{i,k}} \sum_{i=1}^{n} \sum_{k=1}^{m} w_{i,k} l(y_i, r^k((f(x_i)))) \text{ if } (l_{min} - l_{min-1}) > \gamma \\ \arg\min_{r^k, f, w_{i,k}} \sum_{i=1}^{n} \sum_{k=1}^{m} random[w_{i,k}] l(y_i, r^k(f(x_i))) \quad \text{otherwise.} \end{cases} \tag{6}$$

where $\sum_{k=1}^{m} w_{i,k} = 1$. According to Eq.6, the constraints on the parameter initialization stage are used for preventing extreme local optima. Specifically, when the difference between the minimum loss and subminimum loss among the branches is greater than the threshold, the parameter updating of the iteration is considered as "safe", and therefore follows winner-take-all strategy to update. Otherwise, a random initialization is introduced to enforce the network only updating parameters in some of the branches randomly.

As for the guidance mechanism, we use the network parameters which perform well in one dataset to initialize the bad one. The reason behind this operation is that, on the base of the experimental phenomena, we believe the distortion information and the correlation of different distortion types/levels in high-dimensional space have been learned from the successful training dataset. Although the definition of distortion information and labels vary from different IQA datasets, the relative perceptual differences are similar. Therefore, knowledge of the relative perceptual differences learned from the successful training could be transferred to the training of other datasets[4].

[4] For datasets with different label definitions, DMOS value could be mapped to the MOS range by a logistic function, and the introduced bias will be diluted by the network.

Learning for E-Net In the training phase, the training set of E-Net, denoted as D, consists of the distorted image, corresponding m hypotheses and the best quality estimation $r^*(f^*(x_i))$ from MH-Net as triple: $D = \{((x_i, r^j(f(x_i)), k^*)|x_i \in X, r^j(f(x_i)) \in \mathcal{M}\}$, where k^* is the index of optimal branch. That is, the pseudo label to E-Net is corresponding best quality estimation produced by MH-Net. We adopt *softmax loss* to optimize the E-Net:

$$\mathcal{L}_E(D) = \arg\min_{\omega} \sum_{i=1}^{n} \sum_{k=1}^{m} \sum_{j=1}^{m} -log(p(c_i^j|r^k(f(x_i)); \omega)) \tag{7}$$

where p represents the predicted discrete probability distribution from E-Net, c_i^j refers to the category $(0, ..., m-1)$ given k-th hypothesis of x_i.

Since the training data of E-Net comes from MH-Net, it is a network-level interactional supervision. Therefore, it could better exploit the characteristics of the input image, and selects a more proper branch of MH-Net for the input image to predict the final result than normal secondary voting methods. Moreover, no extra annotation is needed in this self-supervised training manner.

3.3 Implementation Detail

The network structure is illustrated in Figure 2. Our implementation for the main CNN body is a modified VGG [18] architecture. We use the Caffe framework with SGD and a mini-batch size of 256. The learning rate starts from 0.1 and is divided by 10 when the error plateaus. The weight decay is of 0.0001 and a momentum is of 0.9. In testing, for comparison studies, we adopt 10 random train-test splits according to [21].

As for the number of m in practice, although more branches lead to performance improvement on complex datasets (*e.g.*, TID2013), there will be obvious performance degradation on simple dataset like LIVE. Therefore, the number of branches is relevant to the complexity and diversity of one particular dataset. We choose $m = 2$ on all dataset evaluations for simplification. Note that the performance can be further improved with m increasing as shown in Table. 1 (a).

m	1	2	3	4	5
LIVE	0.941	0.965	0.940	0.922	0.917
TID2013	0.648	0.731	0.744	0.745	0.750

(a)

	FR-IQA		BIQA		
Method	PSNR	SSIM [20]	LBIQ [19]	BRISQUE [17]	**SDMCL-Net**
SRCC	0.525	0.645	0.74	0.61	**0.83**

(b)

Table 1: (a) The relationship between m and performance (SRCC) in LIVE and TID2013. (b) Performance evaluation (SRCC) on the entire TID2008 database.

Databases	Number of Reference Images	Number of Distorted Images	Number of Distortion Types	Judgment Range	Judgment Type	Number of Judgments
LIVE	29	779	5	[1,100]	DMOS	25k
CSIQ	30	866	6	[0,1]	DMOS	25k
TID2008	25	1700	17	[0,9]	MOS	250k
TID2013	25	3000	24	[0,9]	MOS	500k

Table 2: The information of different IQA datasets evaluated in our work.

Dataset	Method	PSNR	SSIM [20]	FSIM [23]	BLIIDNS-II [15]	BRISQUE [12]	CORNIA [22]	QAF [2]	CNN [5]	SOM [25]	HOSA [21]	BIECON [8]	**SDMCL-Net**
LIVE	LCC	0.856	0.906	0.960	0.930	0.942	0.935	0.953	0.953	0.962	0.953	0.962	**0.964**
	SRCC	0.866	0.913	0.964	0.931	0.940	0.942	0.948	0.956	0.964	0.950	0.961	**0.965**
TID2013	LCC	0.675	0.790	0.877	0.628	0.651	0.613	0.662	-	-	-	0.765	**0.768**
	SRCC	0.687	0.742	0.851	0.536	0.573	0.549	0.589	-	-	0.728	0.721	**0.731**

Table 3: Performance on the LIVE and TID2013 dataset. We divide approaches into full-reference (the first three methods on the table) and blind techniques.

4 Experiments

4.1 Experimental Protocol

Datasets We evaluate our method on LIVE [16], TID2008 [13], TID2013 [14] and CSIQ [9] datasets. A summary of these databases is reported in Table 2.

LIVE dataset is introduced in 2003, which is the first public IQA dataset, and has served as the *de-facto baseline* for development and evaluation of FR-IQA/BIQA metrics. TID2008 and TID2013 datasets are currently the most convincing IQA dataset along with several important dimensions, including the number of distorted images and the diversity of distorted types and levels.

Evaluation metrics The performances on above datasets are evaluated by two common metrics for model evaluation: Pearson linear correlation coefficient (LCC) and Spearman rank order coefficient (SRCC). LCC is used to measure linear correlation between the ground-truth and the prediction, which is defined as

$$LCC = \frac{\sum_{i=1}^{N}(y_i - \bar{y}_i)(\hat{y}_i - \bar{\hat{y}}_i)}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}}_i)^2}} \tag{8}$$

where $\bar{y}_i$ and $\bar{\hat{y}}_i$ denote the means of the ground truth and predicted score, respectively. SRCC is an index of monotonicity, which could be formulated as:

$$SRCC = 1 - \frac{6\sum_{i=1}^{N} r_i}{N(N^2-1)} \tag{9}$$

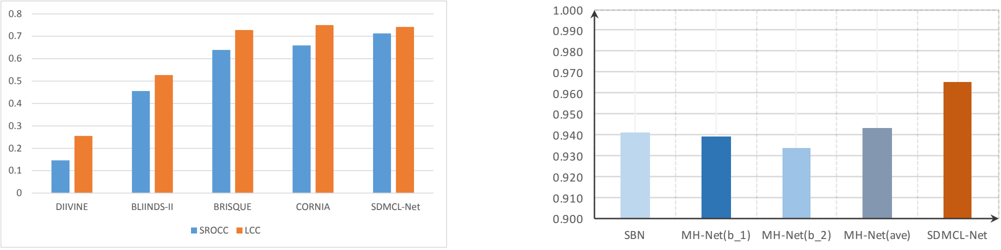
where N represents the number of distorted images, and r is the difference of ranking.

4.2 Results

We compare our SDMCL-Net against with state-of-the-art methods, including BLIINDS-II [15], CORNIA [22], BRISQUE [12], CNN [5], QAF [2], SOM [25], HOSA [21], and BIECON [8]. Besides, we also list classic *FR-IQA* methods, comprising PSNR, SSIM [20], and FSIM [23], to form better assessment.

Comparison with other BIQA methods. We train our proposed SDMCL-Net on LIVE, TID2008 and TID2013 respectively, with randomly selecting 80% reference images and associated distorted images as the training set, 20% testing set for each dataset according to the state-of-the-arts [5, 12, 22, 25]. Table. 3 and Table. 1 (b) show the comparison results with different IQA models and our method achieves promising results on all of the three datasets.

On LIVE dataset, our model outperforms the previous best results (*i.e,* SOM, BIECON) by more than 0.2 in regards to both SRCC and LCC, and reaches at least 4% improvements than other methods. Since LIVE is simpler than other challenging databases, with 5 common



(a) (b)

Figure 3: (a) Performance of training on the TID2013 dataset and testing on the CSIQ dataset. (b) Performance of different variants of our model on the LIVE dataset.

Method	BIQI	BRISQUE [17]	CORNIA-10K [22]	CORNIA-100 [22]	HOSA [21]	**SDMCL-Net**
Time Cost (ms)	111.9	76.8	1621.6	2227.1	352.9	**7.2**

Table 4: Time cost comparison to state-of-the-arts.

distorted types, and there is little difference (less than 2%) among three of the most recently proposed methods (SOM, HOSA and BIECON), we believe that these experiments on LIVE also demonstrate LIVE is becoming saturated.

On TID2013 dataset, the results of our model are close to BIECON, with about 1% relative improvements. BIECON employes state-of-the-art *full-reference* IQA algorithms to generate proxy scores and then fine-tunes them on BIQA. The impossibility of obtaining reference images in practice, limits the feasibility of BIECON. Our SDMCL-Net provides a flexible *alternative* while maintaining satisfactory precision. These experiments demonstrate that simplifying the solution space with self-supervised manner incorporated in is an effective way to ease the ill-posed definition and insufficient training data problem, and therefore improves the precision of BIQA. Besides, it is remarkable that, as a BIQA method, the precision of our SDMCL-Net is very close to those of the state-of-the-art FR-IQA methods.

On TID2008 dataset, our model significantly outperforms LBIQ, which leverages extra data to extract features. The performance improvement is largely due to the effective multiple hypotheses network and the self-supervised mechanism to prevent over-fitting problem of DNNs while maintaining its powerful feature representation ability.

Cross-Dataset evaluations. For evaluating the generalization ability of our approach, we present the cross-dataset experiment following the setups of [24]. Figure 3 (a) presents result of training on TID2013, and testing on CSIQ. The proposed SDMCL-Net achieves promising results on cross-dataset evaluations, which shows our method is generalizable.

Component analysis. For the propose of providing a further insight into our self-supervised deep multiple choice learning scheme, we evaluate different variants of SDMCL-Net on LIVE database with SRCC metric: **Single Branch Network(SBN)** is a normal VGG network. **MH-Net(b_1/b_2)** refers to a policy that generates assessment from only one of the branches of MH-Net without re-filtering from other schemes. **MH-Net(ave)** follows a typical policy of subsequent processing for generating assessment by averaging hypotheses from branches of MH-Net. **SDMCL-Net** is our final model as described in Sec. 3.

As shown in Figure 3 (b), if we only select the result from one of the branches without any auxiliary mechanism, the performance will reduce significantly, which is worse than a single VGG network. While leveraging the averaging strategy, the SRCC value increases. When the E-Net is incorporated in, the SRCC value has a relative 1% improvement than averaging strategy. These experiments again demonstrate the effectiveness of our method.

We hope above experiments could not only demonstrate that our framework provides a flexible alternative to BIQA, but also towards a new perspective to other computer vision tasks that also suffer from the lacks of training data problem.

4.3 Computational Cost

The computational cost of our model is also evaluated. As shown in Table. 4, when we perform our experiments on a PC with a single core i5-4300u CPU, the cost time per image is 7.2ms. The proposed SDMCL-Net is purely feed-forward, and thus can provide a real-time solution to high-performance BIQA applications. Specifically, our method achieves 10 times faster than the previous fastest BIQA method BRISQUE, and more than 49 times faster than HOSA, which is the current state-of-the-art method. This experiment demonstrates the remarkable practicability of our proposed method.

5 Conclusion

In this paper, we present a novel Self-supervised Deep Multiple Choice Learning Network (SDMCL-Net) for BIQA task. The proposed network consists two sub-networks to generate multiple hypotheses of image quality assessment (MH-Net) in branch-level and select the best hypothesis as final prediction (E-Net) to achieve robust estimation. Our result outperforms the state-of-the-art BIQA methods and is comparable to the FR-IQA methods. We show its great potential as an efficient and scalable solution for this task. Future work will focus on investigating more effective "re-filtering" mechanism for better selection.

References

[1] Sebastian Bosse, Dominique Maniry, Thomas Wiegand, and Wojciech Samek. A deep neural network for image quality assessment. In *ICIP*, pages 3773–3777, 2016.

[2] Zhongyi Gu, Lin Zhang, Xiaoxu Liu, and Hongyu Li. Learning quality-aware filters for no-reference image quality assessment. In *ICME*, pages 1–6, 2014.

[3] Abner Guzmán-Rivera, Dhruv Batra, and Pushmeet Kohli. Multiple choice learning: Learning to produce multiple structured outputs. In *NIPS*, pages 1808–1816, 2012.

[4] Abner Guzman-Rivera, Pushmeet Kohli, Ben Glocker, Jamie Shotton, Toby Sharp, Andrew Fitzgibbon, and Shahram Izadi. Multi-output learning for camera relocalization. In *CVPR*, pages 1114–1121, 2014.

[5] Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *CVPR*, pages 1733–1740, 2014.

[6] Le Kang, Peng Ye, Yi Li, and David S. Doermann. Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks. In *ICIP*, pages 2791–2795, 2015.

[7] Jongyoo Kim and Sanghoon Lee. Deep learning of human visual sensitivity in image quality assessment framework. In *CVPR*, pages 1969–1977, 2017.

[8] Jongyoo Kim and Sanghoon Lee. Fully deep blind image quality predictor. *J. Sel. Topics Signal Processing*, 11(1):206–220, 2017.

[9] Eric Cooper Larson and Damon Michael Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *J. Electronic Imaging*, 19 (1):011006, 2010.

[10] Stefan Lee, Senthil Purushwalkam, Michael Cogswell, Viresh Ranjan, David J. Crandall, and Dhruv Batra. Stochastic multiple choice learning for training diverse deep ensembles. In *NIPS*, pages 2119–2127, 2016.

[11] Kede Ma, Wentao Liu, Tongliang Liu, Zhou Wang, and Dacheng Tao. dipiq: Blind image quality assessment by learning-to-rank discriminable image pairs. *TIP*, 26(8): 3951–3964, 2017.

[12] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *TIP*, 21(12):4695–4708, 2012.

[13] 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. *Adv Modern Radioelectron*, 10:30–45, 2004.

[14] Nikolay Ponomarenko, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Lukui Jin, Jaakko Astola, Benoit Vozel, Kacem Chehdi, M Carli, and F Battisti. Color image database tid2013: Peculiarities and preliminary results. In *EUVIP*, pages 106–111, 2013.

[15] Michele A Saad, Alan C Bovik, and Christophe Charrier. Blind image quality assessment: A natural scene statistics approach in the dct domain. *TIP*, 21(8):3339–3352, 2012.

[16] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *TIP*, 15(11):3440–3451, 2006.

[17] Wu Shi, Chen Change Loy, and Xiaoou Tang. Deep specialized network for illuminant estimation. In *ECCV*, pages 371–387, 2016.

[18] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[19] Huixuan Tang, Neel Joshi, and Ashish Kapoor. Learning a blind measure of perceptual image quality. In *CVPR*, 2011.

[20] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 13(4):600–612, 2004.

[21] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D Doermann. Blind image quality assessment based on high order statistics aggregation. *TIP*, 25(9):4444–4457, 2016.

[22] Peng Ye, Jayant Kumar, Le Kang, and David Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *CVPR*, pages 1098–1105, 2012.

[23] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. FSIM: A feature similarity index for image quality assessment. *TIP*, 20(8):2378–2386, 2011.

[24] Lin Zhang, Lei Zhang, and Alan C. Bovik. A feature-enriched completely blind image quality evaluator. *Trans. Image Processing*, 24(8):2579–2591, 2015.

[25] Peng Zhang, Wengang Zhou, Lei Wu, and Houqiang Li. Som: Semantic obviousness metric for image quality assessment. In *CVPR*, pages 2394–2402, June 2015.

[26] Shizhan Zhu, Sifei Liu, Chen Change Loy, and Xiaoou Tang. Deep cascaded bi-network for face hallucination. In *ECCV*, pages 614–630, 2016.