Learning from rankings for no-reference image quality assessment by Siamese Network

Xia-Lei Liu

Abstract

In this thesis we present a no-reference image quality assessment (NR-IQA) approach based on deep Siamese networks. One of the major challenges to apply deep learning techniques to the problem of image quality assessment is the absence of large data sets. To address this problem, we train our Siamese Network to rank images in terms of image quality by using ranked image sets for which relative image quality is known. These ranked image sets can be automatically generated without the use of laborious human labelling. We then use fine-tuning to transfer the knowledge represented by the trained Siamese Network to a traditional CNN that is able to estimate absolute image quality from single images. To solve the difficulty of pair selection for Siamese network training, we demonstrate how our approach can be made significantly more efficient than traditional Siamese Networks by forward propagating a batch of images through a single network and backpropagating gradients derived from all pairs of images in the batch. We evaluate our approach on the LIVE dataset. Our approach is demonstrated to be superior to the existing NR-IQA techniques. Furthermore, we are the first NR-IQA method to surpass the state-of-the-art full-reference IQA (FR-IQA) methods. Experiments on TID2008 and Places2 datasets show the generalization ability of our approach.

Index Terms

No-reference image quality assessment, Convolution neural networks, Siamese network, Speed-up

I. Introduction

WITH the rapid development of mobile digital devices and Internet media, images are everywhere in our life. Unfortunately, images are often distorted by the processes of acquisition, transmission, storage, and external conditions like illumination, camera motion, etc. Image Quality Assessment (IQA) [1] is a technique developed to automatically predict the perceptual quality of images. The result of any IQA estimate should be highly correlated with quality assessments made by a range of very many human subjects (commonly referred to as the Mean Opinion Score (MOS) [2], [3]). IQA has been applied to image restoration [4], image super-resolution [5], image retrieval and ranking [6], image recognition especially for face recognition [7], and in other fields like medical imaging [8], remote sensing imaging [9] and infrared image processing [10], where the quality of image is essential to the success or safety of these applications.

IQA approaches are widely divided into three categories based on whether the undistorted image or information about it is available [11]: full-reference IQA (FR-IQA) [12], [13], [14], reduced-reference IQA (RR-IQA) [15], and no-reference IQA (NR-IQA), which is also known as blind IQA [16], [17], [18], [19]. The two main approaches are FR-IQA and NR-IQA methods, the differences of which are shown in Fig. 1. In realistic application scenarios, the no-reference scenario is the most common and therefore research on NR-IQA has been the most developed. In NR-IQA, many methods focus on a specific distortion [20], [21], which limits the applicability of these methods. Other methods focus on IQA for a wide range of distortions. Such methods are also known as distortion-generic NR-IQA, which is most difficult because no reference image is available. In this thesis, we proposed an approach of NR-IQA based on deep CNNs which outperforms the best reported FR-IQA.

To evaluate the performance of IQA methods, several image databases have been collected by choosing “perfect” images also known as original images of high quality, and then distorting them by adding varying levels of distortions like Gaussian blur, Gaussian noise, JPEG compression, and JPEG2000 compression. After selecting high-quality reference images and generating distorted versions, multiple human assessors are then asked to score each image in terms of quality. Further statistical post processing is then performed on the raw quality scores to ensure annotator agreement and arrive at a single, ground truth image quality score for each image. This process of dataset generation is illustrated in Fig. 2. Some of the most well-known IQA datasets constructed in this way are LIVE [22], TID2008 [3], and CSIQ [23]. This data collection and annotation process is labour-intensive and costly. Each image in the dataset requires *multiple* subjective estimates of image quality by humans. This fundamentally limits the size of available datasets for IQA research.

Convolutional Neural Networks (CNNs) have shaken the computer vision research and practice. CNNs were first proposed by LeCun [24], however after some initial success the method was abandoned by the wider research community. In 2012, Krizhevsky et al. [25] achieved spectacular results with a CNN in the ImageNet competition. This resulted in renewed interest

Author: Xia-Lei Liu, Computer Vision Center, UAB (xialei@cvc.uab.es)
Advisor 1: Joost van de Weijer, Computer Vision Center, UAB (joost@cvc.uab.es)
Advisor 2: Andrew D. Bagdanov, University of Florence (andrew.bagdanov@unifi.it)
Thesis dissertation submitted: September 2016

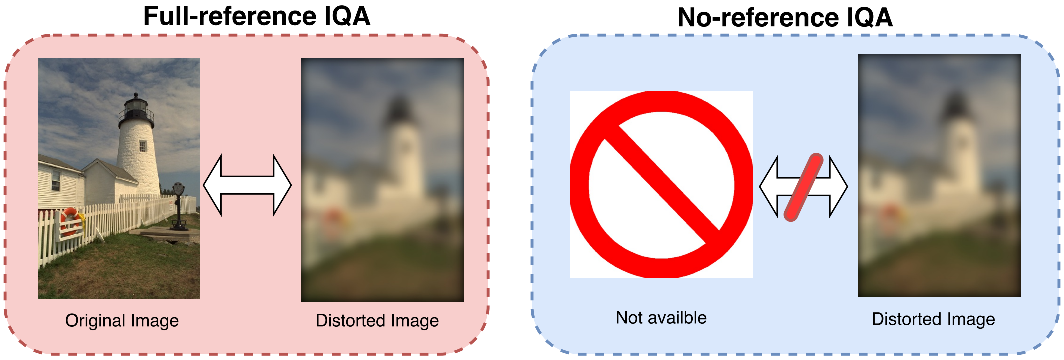


Fig. 1: The difference between Full-reference IQA and No-reference IQA. The aim of IQA is to provide an estimate of the image quality. In NR-IQA, which is the more realistic situation, the original undistorted image is not available. In this thesis we show that with deep CNNs we can obtain better results on NR-IQA than the best reported results in literature on FR-IQA.

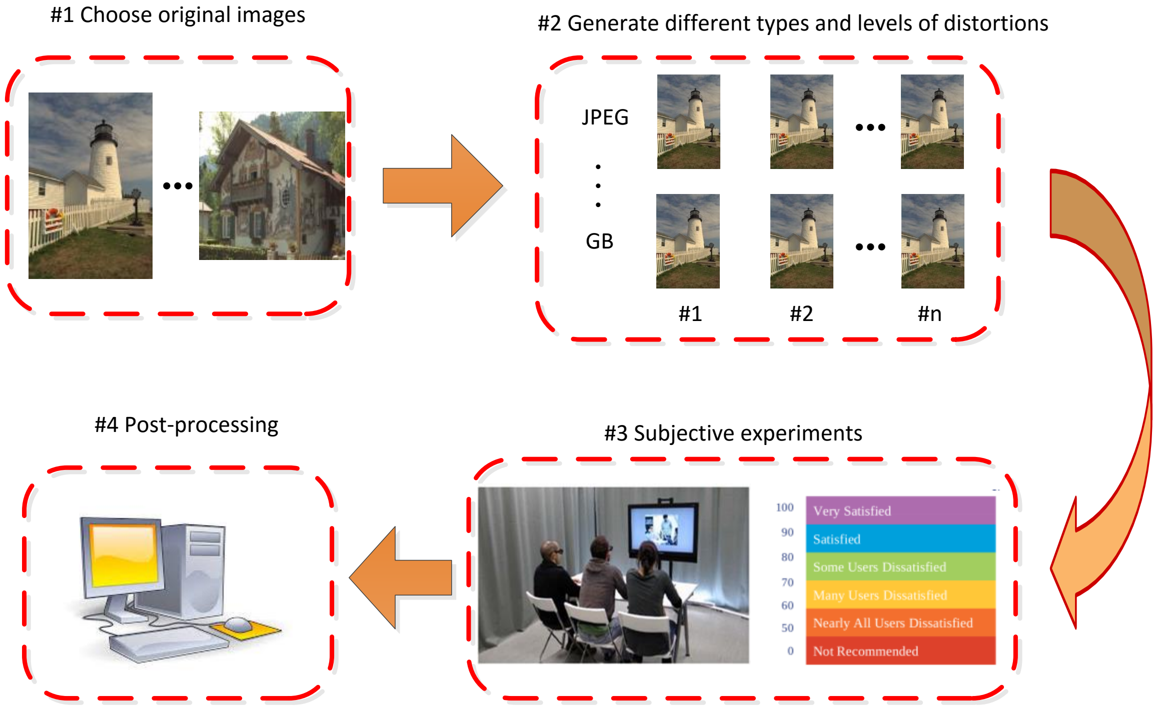

#1 Choose original images
#2 Generate different types and levels of distortions
JPEG
GB
#1
#2
#n
#4 Post-processing
#3 Subjective experiments
100 Very Satisfied
90 Satisfied
80 Some Users Dissatisfied
70 Many Users Dissatisfied
60 Nearly All Users Dissatisfied
50 Not Recommended
0

Fig. 2: The process of generating IQA datasets, which is very costly. As a result, only small datasets exist. In this thesis we address the absence of large datasets with the aim to learn deep CNNs.

in deep learning research. Two of the main factors in the recent success of CNNs are the *availability of large labelled data sets*, and improved Graphics Processing Units (GPUs). Since their reintroduction, CNNs have achieved remarkable success in many areas of computer vision, such as image classification, object recognition and semantic segmentation. The architectures of networks are getting deeper and deeper with respect to the original AlexNet, with ResNet being an example of very deep network architecture [26].

The success of CNNs encouraged us to explore the application of deep models to the problem of NR-IQA. However, as networks grow deeper and wider, the number of parameters increases dramatically. As a consequence, larger and larger annotated datasets are required for training. As mentioned above, the annotation process for IQA image datasets require multiple human annotations for every image, and thus the collection process is extremely labour-intensive and costly. As a results, most of the available IQA datasets are far to small in size to be effective for training CNNs.

The main idea behind our approach, and the main contribution of this thesis, is the observation that while human annotated IQA data is difficult to obtain, it is relatively easy to *generate* images that are *ranked* according to their image quality. That

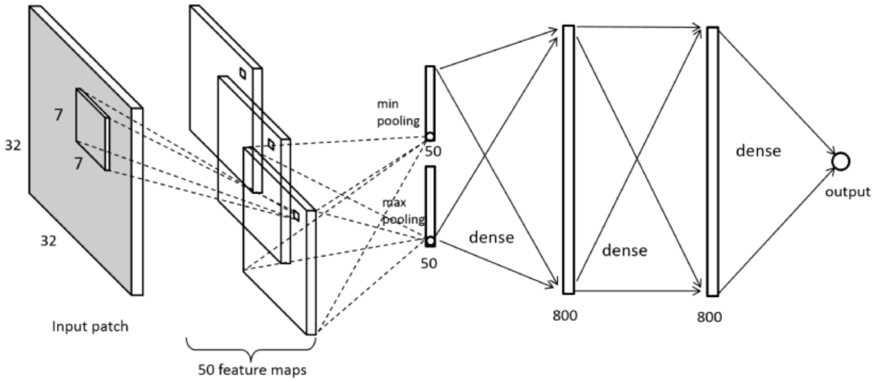


Fig. 3: The CNN architecture proposed for NR-IQA in [32]. The proposed network consists of six layers: a 32×32 input image, followed by a $7 \times 7 \times 50$ convolutional layer, followed by parallel, global min and max pooling layers, followed by two 800 neuron fully connected layers, and finally a single quality assessment output unit. This was the first successful application of CNNs to NR-IQA.

is, we can generate image sets in which, though we do not have an absolute quality measure for each generated image, for any *pair* of images we know which is of higher quality. We call this approach *learning from rankings*, and with it we will show how to learn from rankings using Siamese Networks, and then transfer knowledge learned from ranked images to IQA datasets in order to improve the accuracy of image quality assessment.

As a second contribution we propose how our approach can be made significantly more efficient than traditional Siamese Networks by forward propagating a batch of images through a single network and backpropagating gradients derived from all pairs of images in the batch.

This thesis is organized as follows. In the next section we discuss work from the literature on IQA. We present our approach to learning from ranked image sets in section III, and report on a range of experiments we performed to quantify the performance of our approach in section IV. We conclude in section V with a discussion of our contribution and directions for ongoing and future work.

II. Related work

In this section we briefly review some of the literature related to our approach to exploiting image quality *rankings* to improve No-Reference Image Quality Assessment (NR-IQA). We focus on distortion-generic NR-IQA since it is more generally applicable than the other research lines of IQA.

A. Traditional IQA approaches

Most of the traditional NR-IQA can be classified into Natural Scene Statistics (NSS) methods and learning-based methods. In NSS methods, the assumption is that the images of different quality vary in the statistics of responses to certain filters. Wavelets [16], DCT [17] and Curvelet [18] are commonly used to extract the features in different sub-bands. These feature distributions are parametrized, for example with the generalized Gaussian distribution (GGD) [27]. Hence, the aim of these methods is to estimate the distributional parameters, from which a quality assessment can be inferred. A drawback of these methods is the complexity of the above transforms. The authors of [19] proposed to extract features in the spatial domain based on NSS, and they obtain significant speed-ups.

In learning-based methods, local features are extracted and mapped to the MOS using, for example, Support Machine Regression (SVR) or Neural Networks (NN) [28]. The codebook method [29], [30] combines different features instead of using local features directly. Datasets without MOS can be exploited to construct the codebook by means of unsupervised learning, which is particularly important because of the small size of existing datasets. Beyond these methods, saliency maps [31] can be used to imitate the human vision system (HVS) to improve precision.

B. Deep learning and IQA

The technique described in [32] was the first to apply CNNs to the NR-IQA problem. The architecture of their network is shown in Fig. 3. It takes non-overlapping 32×32 patches from large images as inputs to the network. This technique of sampling patches from large images is used to compensate for the lack of images in IQA datasets, and to satisfy the need for very many training sample needed to train deep neural networks. The authors of [33] follow the same pipeline and designed a multi-task CNN to learn the type of distortions and image quality simultaneously. The approach in [34] (shown in Fig. 4) extracts features from a pre-trained model fine-tuned on an IQA dataset. These features, along with ground truth MOS quality scores, are then used to train a support vector regression model to map features to IQA scores. The use of pre-trained networks trained on massive datasets helps avoid the need to sample small patches from IQA dataset images.

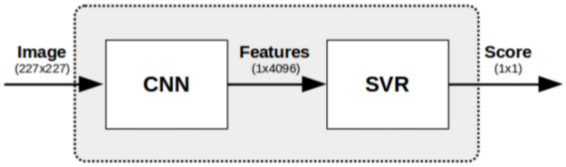


Fig. 4: Graphical representation of the approach proposed in [34]. The input image is fed into a CNN which is used as a *feature extractor* by capturing the activations from the fully-connected layer at the end of the CNN. Then, support vector regression (SVR) is used to map the extracted features to the perceived quality scores.

C. Learning to rank

As mentioned before, one of the main drawbacks of deep networks is the need for large labelled datasets. This problem has been partially addressed by fine-tuning existing pre-trained networks to new tasks [35]. Another approach to mitigate the absence of training data is to use data augmentation, e.g. by flipping, rotating, cropping the images in the dataset. The labelling of data for IQA is laborious and existing data sets are small. To address this problem we exploit *ranked image sets* to train networks. The advantage is that these ranked image sets can be generated automatically just by applying image distortions of different levels to the image.

Learning to rank has become a hot topic in many fields, such as information retrieval and image comparing. The main idea is to learn a function from the given rankings by minimizing a ranking loss function [36]. This function can then be applied to rank test objects. These methods can be divided into three categories: the pointwise approach [37] in which single objects are seen as training samples, the pairwise approach [38] which regards pairs of objects as learning samples, and the listwise approach [39] which takes the whole ranking lists as samples.

The authors of [40] adapt the Stochastic Gradient Descent (SGD) method to perform pairwise learning to rank. This has been successfully applied to large datasets. Combining ideas from ranking and CNN, the Siamese network architecture achieved great success on the face verification problem [41], and in comparing image patches [42]. The Triplet Network [43], [44] is able to learn image similarity, and obtains excellent results in image retrieval.

For the NR-IQA problem, the main hurdle of using CNNs is that the available datasets are too small to train the deep models. In our work, we are partially inspired by the work of [45] in which they combine different hand-crafted features to represent image pairs from the IQA dataset based on MOS. To use Siamese Networks, we can learn the rankings of image quality from image pairs generated by adding different types and levels of distortions, which does not require any hand labelling of the data.

D. Contributions of this thesis with respect to the start-of-the-art

In this thesis we describe an approach to automatically generating ranked image sets that are used to train a Siamese Network to rank images in terms of image quality. We then show how to use fine-tuning to transfer the knowledge represented by the trained Siamese Network to a traditional CNN able to estimate absolute image quality from single images. Finally, we demonstrate how our approach can be made significantly more efficient than traditional Siamese Networks by forward propagating a batch of images through a *single* network (instead of two) and backpropagating gradients derived from all *pairs* of images in the batch.

To the best of our knowledge, Siamese Networks have not been applied to NR-IQA. Use of Siamese Networks for *ranking* as opposed to absolute IQA estimation allows us to train deep networks on small IQA datasets even when there is no ground truth MOS available. After training on the large ranking dataset, fine-tuning can be used to make the model more precise on a specific dataset that for which MOS are available. This will in turn allow us to design an end-to-end network for NR-IQA (in contrast to [45]).

III. Learning from rankings for NR-IQA

In this section we describe our approach to learning from image quality rankings. We first introduce the general NR-IQA problem and discuss existing IQA datasets.

TABLE I: Overview of publicly available IQA datasets.

Name	Year	Original images	Distorted images	Distortion types	Human annotations
LIVE	2006	29	779	5	161
TID2008	2008	25	1,700	17	838
CSIQ	2010	30	866	6	35
TID2013	2013	25	3,000	24	971
Waterloo	2016	4,744	94,880	4	–

A. The general NR-IQA problem

As mentioned before, the objective of NR-IQA is to automatically predict image quality scores without knowing the distortion types and levels. After introducing the process of generating IQA datasets, we know that laborious human labelling limits the size of existing IQA datasets for training deep CNNs. An overview of available datasets for evaluating IQA methods are given in Table I. We briefly describe the distortion types, the size of dataset and the human resources.

- The LIVE dataset is the most well-known IQA dataset. It consists of 808 images generated from 29 original images by distorting them with five types of distortion: Gaussian blur (GB), Gaussian noise (GN), JPEG compression (JPEG), JPEG2000 compression (JP2K) and fast fading (FF). The ground-truth Mean Opinion Score (DMOS) for each image is in the range [0, 100] and is estimated using annotations by 161 human annotators.
- The TID2008 dataset consists of 25 reference images with 1700 distorted images from 17 different distortion types at 4 degradation levels. Mean Opinion Scores in the range [0, 9] are derived from annotations by 838 humans for each image. Four types of distortion are shared with LIVE dataset: GB, GN, JPEG, and JP2K.
- The CSIQ dataset consists of 866 images distorted from 30 reference images. Six types of distortion are included: GB, GN, JPEG, JP2K, additive Gaussian pink noise and global contrast decrements. MOS estimates are derived from annotations by 35 humans and are available for every image in the range [0, 1].
- The TID2013 dataset enlarges TID2008 by including another 7 types of distortion. Ground-truth MOS estimates are derived from 971 human annotators and are given for each image in the range [0.2,7.3].
- The latest IQA dataset, the Waterloo dataset, contains 4744 reference images, and four types of distortion shared with other datasets are generated at 5 levels for each type. Waterloo is the currently the largest IQA dataset available with a total of 94,880 images. However, it does *not* include Mean Opinion Scores due to the large expense of obtaining them from human annotators.

The first thing to notice about these publicly available IQA datasets is their limited size. As mentioned above, this is largely due to the high cost of annotation. To obtain reliable MOS estimates, each image must be annotated by *multiple human annotators*. In the case of TID2008, for example, each image was annotated by 838 human subjects, for a total of *nearly 1.5 million total annotations.*.

Massive amounts of data are needed to train deep neural networks and take advantage of the recent advances in CNNs in the context of IQA. There are two popular strategies to augment IQA datasets to meet this requirement of large datasets:

- Sample small patches from the original images in the IQA dataset (i.e. 32×32 patches chosen at random). The drawback of this method is that the size of input images must be very small in order to generate enough input samples for training. This is not straightforward for IQA methods where large size of images are needed to robustly estimate image quality. Also, the generalization of this method is hard to ensure since the models are only trained using the databases containing fewer than a thousand real-world images.
- Extract the feature descriptions using pre-trained CNNs or fine-tuned CNNs based on image classification problem. The drawback of this method is that it is not end-to-end trainable, so it is difficult to achieve the optimal results. Fine-tuning helps the training, but the pre-trained models used are usually trained on ImageNet for classification problems, which is not related to IQA problem.

Both methods to address the lack of training data for IQA have their drawbacks. We will propose an alternative method in next section.

B. The framework of our approach

The lack of large IQA datasets for the IQA problem motivates us to propose a new strategy to take advantage of large, *unlabelled* databases from which we can generate images ranked by image quality. Our approach is based on the observation that, given high quality reference images, it is very easy to apply image distortions to generate a *ranking* image dataset, which is strongly related to the final scores of the images. As an example consider distortion with Gaussian blur. Given a high quality reference image we could apply various levels of Gaussian blur. The set of images which is thus generated can be easily ranked because we do know that adding Gaussian blur (or any other distortion) will always deteriorate the quality score. It should be noted though that, in such a ranking image dataset we do not have any absolute IQA scores for any images, but we do know for any pair of images *which is of higher quality*.

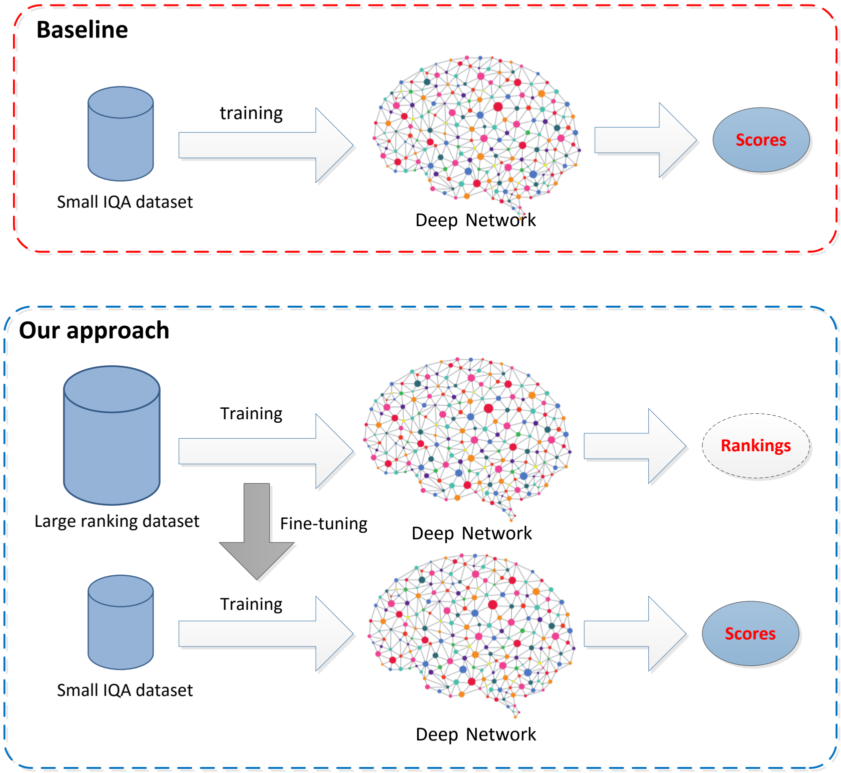


Fig. 5: The Difference between the baseline CNN and our approach. The baseline approach (top) trains a deep Convolutional Neural Network *directly* on the ground-truth MOS annotations. The result is a deep CNN trained as a *regressor* from images to IQA scores. Our approach (bottom) is to train a network on a image *ranking* dataset. These ranked images can be easily generated by applying distortions of varying intensities to high-quality reference images. The network is trained to *rank* pairs of images in terms of image quality.

In the next subsections we will elaborate on the learning method we use to learn from ranked image sets. After learning on these *ranked images*, we can transfer this knowledge via fine-tuning on small image quality datasets in order to solve the IQA problem. The difference between our approach and the baseline is shown in Fig. 5. The baseline usually trains a shallow network directly on the IQA dataset to estimate IQA score from images. Due to the limited data only few layers can be used, which limits accuracy. Since we have access to much larger data sets with ranked images, we can now train deeper and wider networks to learn a distance embedding. Next we follow this by fine-tuning for domain adaptation to the absolute IQA problem. Apart from solving the IQA problem, our ranking network could predict rankings among different qualities of images as well.

C. *Siamese networks for ranking*

In this section we introduce the Siamese network which we will use to learn from image rankings. The Siamese network [41] is a network with two branches and a loss module. The two branches of the network are identical and share weights during the training process. Pairs of images and labels are the input of the network, yielding two outputs which are passed to the loss module (illustrated in Fig. 6). The gradients of the loss function are computed by backpropagation and updated with the stochastic gradient method.

The aim of a Siamese network is to learn a similarity model. They map images to a feature space in which equal class images are close and non-equal class images are far away. These networks have a resemblance with earlier work on distance learning. Therefore, to train a Siamese network pairs of images need to be generated. Sampling randomly from the dataset is a direct way to generate pairs in a mini-batch.

Specifically, given an image x as the input of the network, the output feature representation of x, denoted by $f(x)$, is obtained by capturing the activations in the last layer of the network. The distance between images x_1 and x_2 is then computed as:

$$D(x_1, x_2) = \|f(x_1) - f(x_2)\|_2 \tag{1}$$

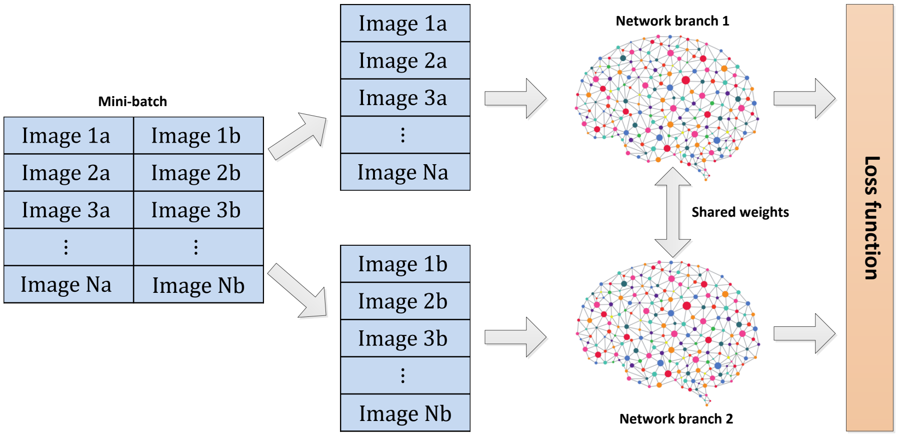


Fig. 6: Standard architecture of Siamese networks

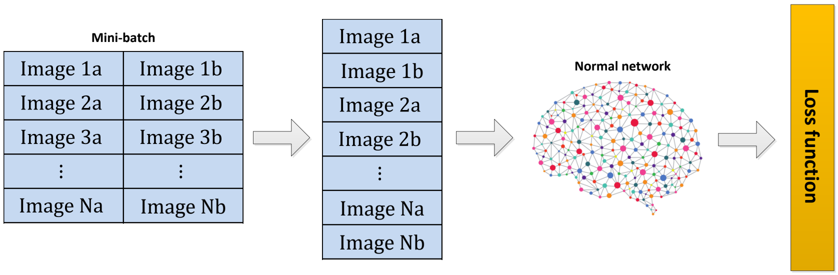


Fig. 7: A variant architecture of Siamese networks. This architecture avoids some of the redundancy in the original model the loss and gradient computation to extract pairs from the mini-batch on which to compute the loss.

The most popular loss function used for training Siamese networks is the contrastive loss, which is computed as:

$$L\left(x_1,x_2,l\right)=\frac{1}{2}lD(x_1,x_2)^2+\frac{1}{2}\left(1-l\right)\max\left(0,\varepsilon-D(x_1,x_2)\right)^2 \tag{2}$$

In Eq. 2, similarity label $l\in\{0,1\}$ indicates whether the input pair of images are similar or not ($l=1$ for similar, $l=0$ for dissimilar), and ε is a margin for determining which images are dissimilar.

Our goal is to estimate the rankings among different levels of distortions. In this case the contrastive loss is inappropriate because we have no same class ($l=1$) images in our dataset. Instead we have rankings of images. Therefore we apply the pairwise ranking hinge loss:

$$L(x_1,x_2)=\max\left(0,-(f(x_1)-f(x_2))+\varepsilon\right) \tag{3}$$

Here we assume without loss of generality that the rank of x_1 is higher than x_2. The gradient of the loss in Eq. 3 is given by

$$\nabla_\theta L=\begin{cases}0 & \text{if } f\left(x_2;\theta\right)\leq f\left(x_1;\theta\right)-\varepsilon\\ \nabla_\theta L\left(x_2;\theta\right)-\nabla_\theta L\left(x_1;\theta\right) & \text{otherwise}\end{cases} \tag{4}$$

It is decomposed in two cases. When the outcome of the network is in accordance with the ranking, the gradient is set to zero. When the outcome of the network is not in accordance we decrease the gradient of the higher and add the gradient of the lower score. Parameter θ can be updated by SGD.

D. Siamese networks and redundant computation

One drawback of the Siamese networks is the redundant computation. Consider the case of three image pairs, image 1 and 2, image 1 and 3 and image 2 and 3. In a naive implementation the three images are passed twice through the network, because they appear all in two pairs. However, one could wonder if for the computation of the last pair (2-3) any image needs to be passed through the network since all computation was already done previously. It is exactly this idea that we exploit in the next section where we propose a new interpretation of Siamese Networks which leads to a faster backpropagation algorithm.

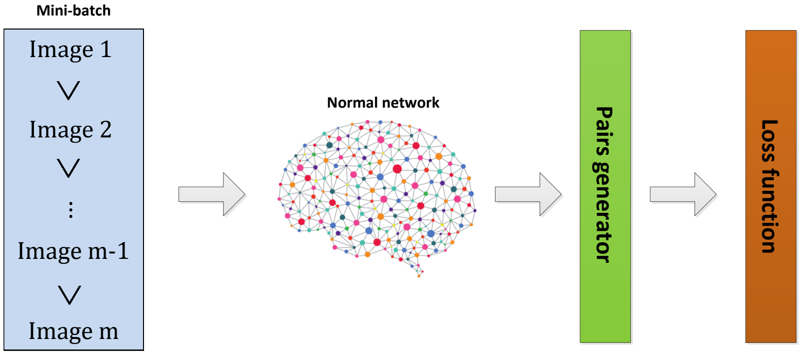


Fig. 8: The fast Siamese network. This network takes a list of ranking images as input and adds a module called "Pairs generator" to select all possible pairs in input list for computing the loss.

To facilitate the understanding of the network which we will introduce in the next section, we will first introduce an intermediate network. This network does the same as the original Siamese network but does not require the two branches. The standard Siamese network architecture passes image x_1 through one branch of the network, and x_2 through the other branch (see Fig. 6). Consider the single-branch architecture shown in Fig 7. The only difference between these two architectures is in how the loss function finds pairs of images in the batch. In the previous architecture, the loss is computed between the outputs of two networks, so the gradient should be backpropagated through two networks respectively. In this architecture, only one branch of network is used, and the pairs are passed through the mini-batch. This alternative way of looking at the Siamese network is further exploited in the next section.

E. *Fast Siamese backpropagation*

In the previous section we proposed a variant of the Siamese network, where images are passed through a single network, and the pairs are combined in the loss. We also discussed the redundancy problem, showing that when the same image is present in multiple pairs it is passed multiple times through the network. This redundancy problem can however be addressed in the architecture of Fig. 7. The loss function could be adapted to use the same image for multiple pairs (only passing it once through the network).

In fact, nothing prevents us, from considering all possible pairs in the mini-batch, almost without any additional computation. Therefore, we propose to take the list of ranking images as input to the network by adding a new module that generates all possible pairs in a mini-batch at the end of the network before computing the loss. This will eliminate the problem of pair selection and potentially boost efficiency. Our proposed architecture is shown in Fig. 8.

To compute the speed-up of our network proposal consider the following. If we have one reference image distorted by n levels of distortions, then the total number of passes through the traditional Siamese network is equal to $n^2 - n$, which is the number of pairs you can generate with n images.

Instead of individual pairs, in our new architecture we consider all possible pairs of one ranking. That would reduce computation to just n passes through the network. Therefore, the speed-up would be equal to: $\frac{n^2-n}{n} = n - 1$. In the best scenario $n = M$, where M is equal to the number of images in the mini-batch, and hence the speed-up of our proposed method would be in the order of the mini-batch size.

In practise, if M images are passed through the network in a mini-batch, the gradient for backpropagation is computed as:

$$\nabla_\theta L = \frac{1}{M} \sum_{k=1}^{M} \left\{ \begin{matrix} 0 & a_{12} & \cdots & a_{1j} \\ a_{21} & 0 & \cdots & a_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i1} & \cdots & a_{i(j-1)} & 0 \end{matrix} \right\} \begin{pmatrix} \nabla_\theta L(x_1;\theta) \\ \nabla_\theta L(x_2;\theta) \\ \vdots \\ \nabla_\theta L(x_M;\theta) \end{pmatrix}, \tag{5}$$

where

$$a_{ij} = \begin{cases} 1 & \text{if } f(x_j;\theta) > f(x_i;\theta) - \varepsilon \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad a_{ji} = -a_{ij} \tag{6}$$

We sum up the gradient for each image in mini-batch to update the parameters, which works for only one distortion type when training the network. Suppose instead we have D types of distortions. We can compute the gradients of the loss using a block diagonal matrix as:

$$\nabla_\theta L = \frac{1}{dM}\sum_{k=1}^{dM}\begin{Bmatrix} A^1 & 0 & \cdots & 0 \\ 0 & A^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A^d \end{Bmatrix}\begin{Bmatrix} \nabla_\theta L^1 \\ \nabla_\theta L^2 \\ \vdots \\ \nabla_\theta L^d \end{Bmatrix}, \tag{7}$$

where

$$A^d = \begin{Bmatrix} 0 & a^d_{12} & \cdots & a^d_{1j} \\ a^d_{21} & 0 & \cdots & a^d_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ a^d_{i1} & \cdots & a^d_{i(j-1)} & 0 \end{Bmatrix} \quad \text{and} \quad \nabla_\theta L^d = \begin{pmatrix} \nabla_\theta L(x_{d1};\theta) \\ \nabla_\theta L(x_{d2};\theta) \\ \vdots \\ \nabla_\theta L(x_{dM};\theta) \end{pmatrix}. \tag{8}$$

for $d \in \{1, 2, \ldots, D\}$. In the definition of A^d above, the a^d_{ij} are the gradient coefficients as in Eq. 6 determined by the rank of images i and j in terms of distortion d.

IV. Experimental results

In this section we report on a number of experiments we performed to quantify the performance of our approach. Experiments were conducted on the LIVE, TID2008, Waterloo IQA datasets, and the validation dataset of Places2. After introducing the experimental setup and evaluation protocol (sections IV-A and IV-B), we evaluate several relevant settings of the network (in experiments IV-C & IV-E & IV-F). In experiment IV-D we evaluate our newly proposed fast Siamese backpropagation method. In experiment IV-G we compare our method to state-of-the-art and finally in IV-H we evaluate the generalization of our approach on another dataset.

A. *Experimental setup*

Generation of Ranked Pairs Now that we have the machinery necessary to efficiently compute gradients for backpropagation from a ranked list of images, we now detail precisely how we generate these ranked batches of images. In this thesis, we focus on four types of distortions which are widely used and shared in most of existing IQA databases: Gaussian Blur (GB), Gaussian Noise (GN), JPEG Compression (JPEG), and JPEG2000 Compression (JP2K).

We follow the database generating method in [46]. Source images from the Waterloo dataset are distorted with the four distortion types at five levels each. All distorted images are generated using MATLAB functions provided by [46], and the parameters that control the distortion levels for each type are chosen to uniformly cover the subjective quality scale:

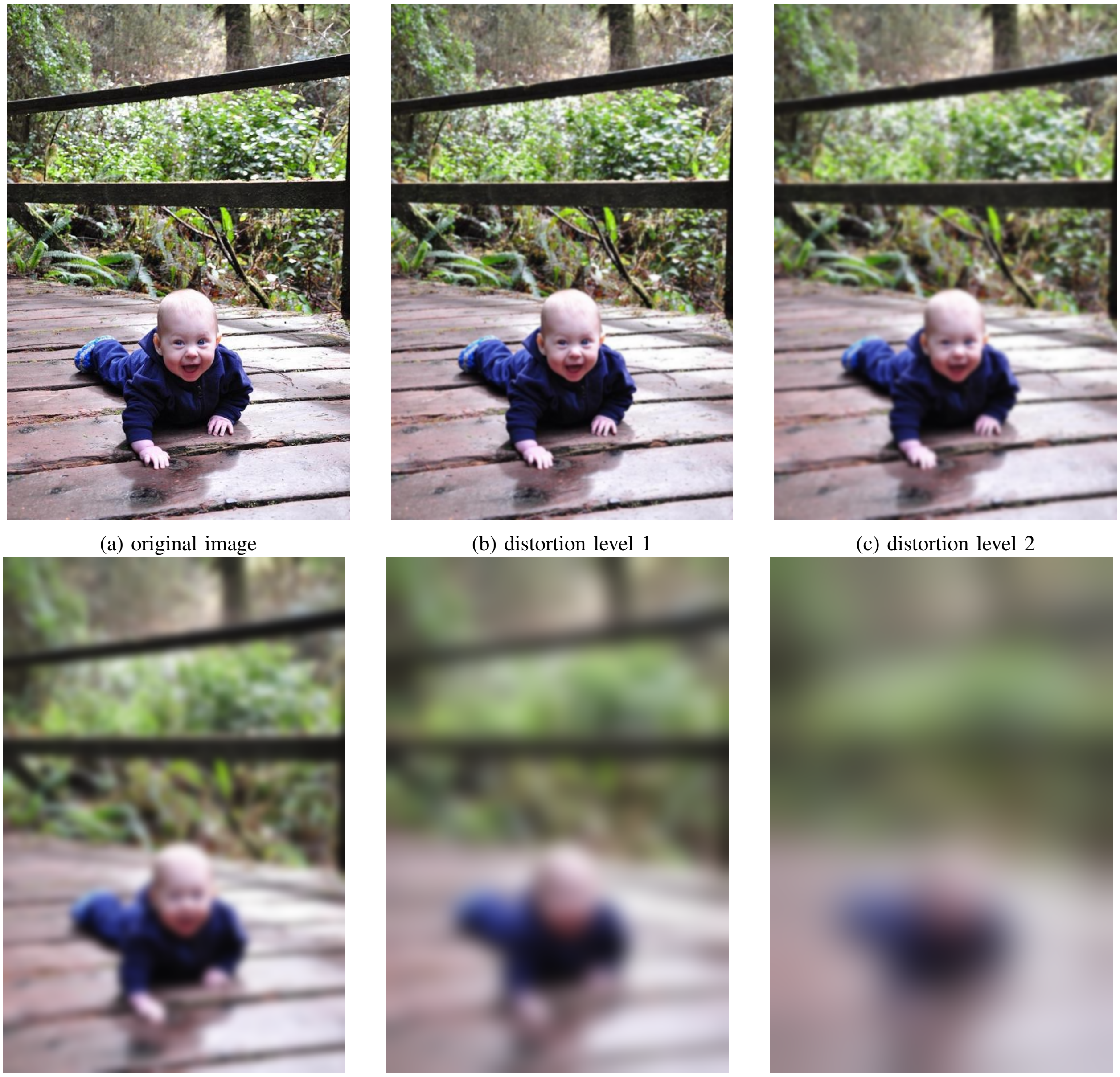
- **JPEG Compression**: The quality factor that determines the DCT quantization matrix is set to be [43, 12, 7, 4, 0] for the five levels, respectively.
- **JPEG2000 Compression**: The compression ratio is set to be [52, 150, 343, 600, 1200] for the five levels, respectively.
- **Gaussian Blur**: 2D circularly symmetric Gaussian blur kernels with standard deviations of [1.2, 2.5, 6.5, 15.2, 33.2] are used to distort the original images.
- **Gaussian Noise**: Gaussian noise is added to the original images, where variances are set to [0.001, 0.006, 0.022, 0.088, 1.000] for the five distortion levels, respectively.

Apart from the Waterloo dataset, we generated another ranking dataset using the validation set of the Places2 dataset of 356 scene categories [47]. There are 100 images per category in the validation set, for a total 36500 images. After distortion, we have a total of 730,000 distorted images for learning a image quality ranking embedding. An example of the ranking images of GB distortion generated on validation dataset of Places2 is shown in Fig. 9. The aim of generating this dataset is to demonstrate the generalization of our ranking network. High-quality ranking embeddings can be learned using datasets not specifically designed for the IQA problem.

TABLE II: Details of our Shallow network used for evaluation.

Blocks	Output sizes	Details of block
Input block	3x227x227	3-channels images
Block 1	32x57x57	conv1(3x3), relu1, pool1(4x4)
Block 2	32x14x14	conv2(3x3), relu2, pool2(4x4)
Block 3	32x12x12	conv3(3x3), relu3
Block 4	32x1x1	conv4(12x12)
Block 5	1	fc1

Network Architectures We evaluate three typical network architectures varying from shallow to deep. We refer to them as: Shallow, AlexNet [25], and VGG-16 [48]. The details of the Shallow network are given in Table II. For details of the AlexNet and VGG-16 networks, please refer to the original publications. The only change we make to these two networks is the number of outputs, since our final objective is to assign one score value to each of the distorted image.



(a) original image

(b) distortion level 1

(c) distortion level 2

(d) distortion level 3

(e) distortion level 4

(f) distortion level 5

Fig. 9: The ranking images of GB distortion generated on validation dataset of Places2.

Strategy for Training and Testing As shown in Fig. 10, we randomly sample sub-images of the appropriate size from the original high resolution images. We do this instead of scaling images to avoid introducing distortions caused by interpolation or filtering. The size of sampled images are determined by the requirements of the networks. However, the large size of the input images is important since the input sub-images should be at least 1/3 of the original images in order to capture more context information. This is a serious limitation of the patch sampling approach that samples very small, 32×32 patches from the original images. In our experiments, we sample 227×227 and 224×224 pixel images, depending on the network.

In testing, we randomly sample 30 sub-images from the original images according to the suggestion of [34], and pass all the sub-regions to the trained models. The average of all outputs of the sub-regions is the final score for each distorted image.

B. Evaluation protocols

Two traditional evaluation metrics are used to evaluate the performance of IQA algorithms: Linear Correlation Coefficient (LCC) and Spearman Rank Order Correlation Coefficient (SROCC). LCC is a measure of the linear correlation between the ground truth and the predicted quality scores. Given N distorted images, the ground truth of i-th image is denoted by x_i, and

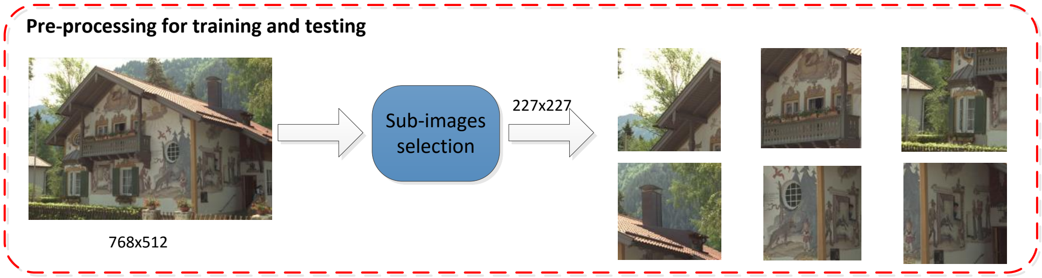


Fig. 10: The pre-process of the input images

the predicted score from the network is y_i. LCC is computed as:

$$LCC = \frac{\sum_{i=1}^{N}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sqrt{\sum_{i}^{N}\left(x_i - \overline{x}\right)^2}\sqrt{\sum_{i}^{N}\left(y_i - \overline{y}\right)^2}} \tag{9}$$

where $\overline{x}$ and $\overline{y}$ are the means of the ground truth and predicted quality scores, respectively.

SROCC is used to evaluate their correlation regarding to the monotonic relationship. Given N distorted images, rankings of i-th image can be converted from x_i and y_i to v_i and p_i, respectively, i.e., the numerical data of predicted image quality y_i are observed as 3.4, 5.1, 2.6, 7.3, the ranks of these data items would be p_i as 2, 3, 1 and 4 (similar from x_i to v_i). The SROCC is computed as:

$$SROCC \quad = 1 - \frac{6\sum_{i=1}^{N}\left(v_i - p_i\right)^2}{N\left(N^2 - 1\right)} \tag{10}$$

C. *Varying the loss function*

The objective of this experiment is to compare the contrastive loss function with the ranking loss function (see Eq. 2 and 3). We use 30,000 training and 3000 test samples randomly sampled from the JP2K and GB subsets to train and test AlexNet using the naive implementation of the Siamese network. The final classification results for each type of distortion are used to evaluate the performance for both loss functions.

The classification performance of contrastive and ranking loss functions is shown in Fig. 11. We plot a histogram of the values predicted by the network using the color of the ground truth distortion value (five in total). From this plot, we see that the network can separate different rankings clearly using ranking loss on GB distortions, while it is hard to separate the first two rankings using contrastive loss. With the increasing difficulty of distortion type, both methods suffer on the JP2K dataset. However, the ranking loss still performs acceptable in terms of classification, while contrastive loss fails to clearly separate ranks of JP2K distortion. Similar results we obtained on another two distortions, and suggest that ranking loss performs better than contrastive loss for the IQA problem. Therefore, from now on, the ranking loss are used in all following experiments.

D. *Comparison of naive Siamese network with fast Siamese network*

The objective of this experiment is to compare the performance of the naive Siamese network (see Fig. 6 and our proposed fast Siamese network (see Fig. 8). We adapt AlexNet to both Siamese architectures and compare their convergence rates on the training set of JPEG distortions.

The ranking loss as a function of training iterations is shown in Fig. 12. The mini-batch size for both networks is 18, which means 18 pairs of images for standard Siamese network and 18 ordered images for fast Siamese network. The dataset used in this experiment is JPEG dataset since it is the most challenging dataset in our task. It is clear that the fast Siamese network converges faster than the standard Siamese network and that training loss converges to a lower point. After the same number of training iterations, ranking performance is shown in Fig. 13 for both networks. From this we see that the standard Siamese network has trouble splitting the first two and the last two distortion levels. Our fast Siamese network still does not separate them perfectly, but the results are much better than the standard method.

E. *Varying network architecture for ranking*

In this experiment we show the qualitative performance of ranking networks when varying the architecture of networks (i.e. Shallow, AlexNet and VGG-16). We train the three networks on a mixed dataset which includes all four distortion types. Inputs of networks are the lists of four distortions at the same time.

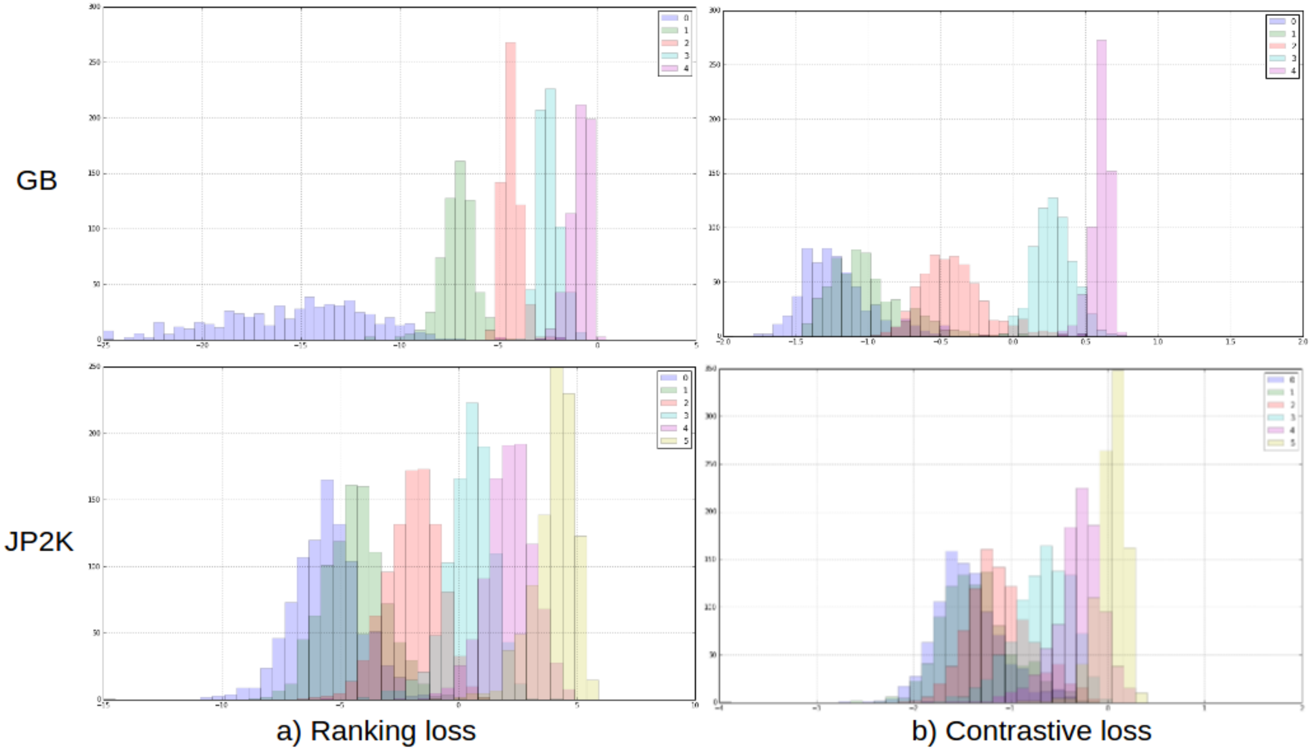


Fig. 11: The ranking results obtained using AlexNet with contrastive and ranking loss functions.

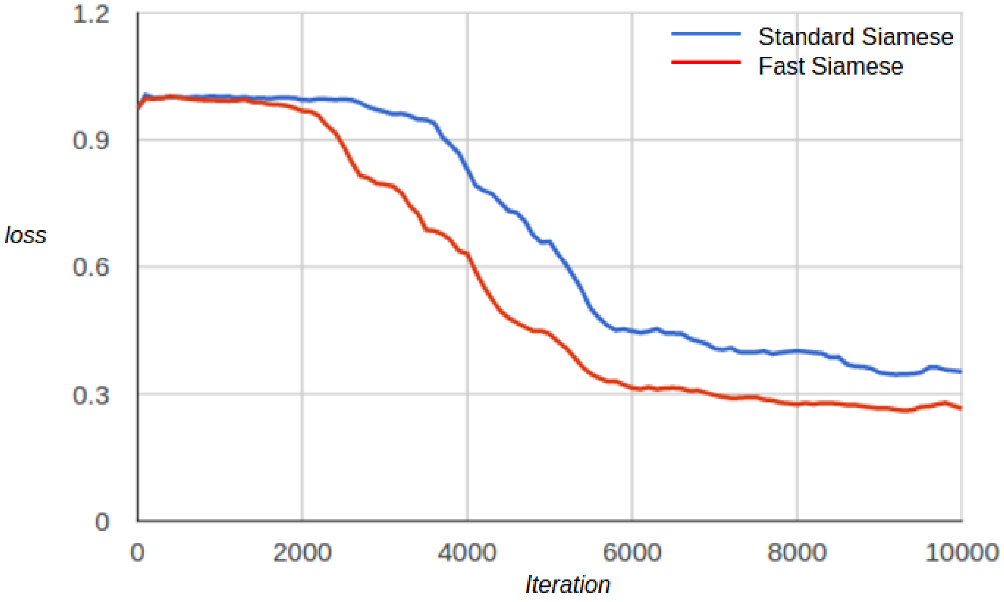
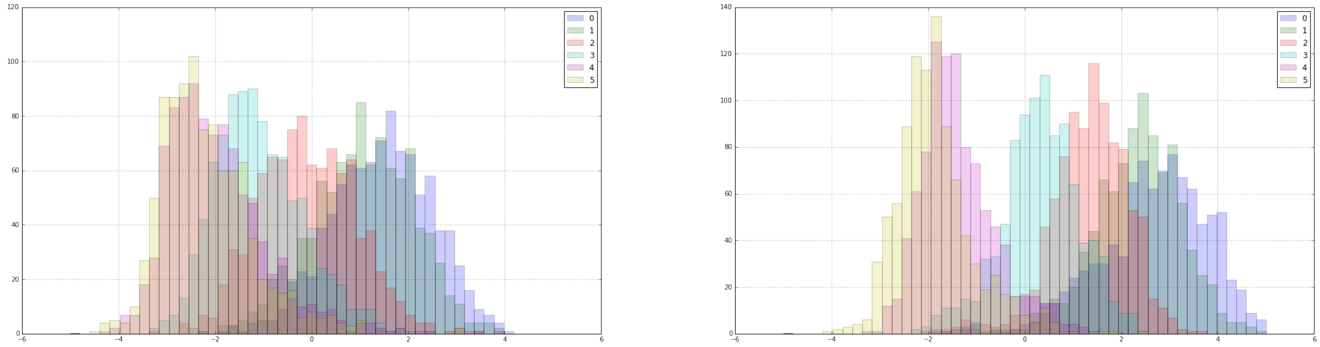


Fig. 12: The ranking loss as a function of training iterations for standard Siamese and fast Siamese networks on JPEG distortions dataset.

The ranking results obtained from different distortions using different architecture are shown in Fig. 14. It is obvious that deeper networks have more capacity to discriminate among different distortion levels. VGG-16 is much better than AlexNet and the Shallow network at distinguishing different levels of distortions for each type. While the Shallow network only shows good results on simple distortions like GN, and poor performances on more complicated ones. AlexNet performs well on all four distortions, but is still worse than the deeper VGG-16 network.

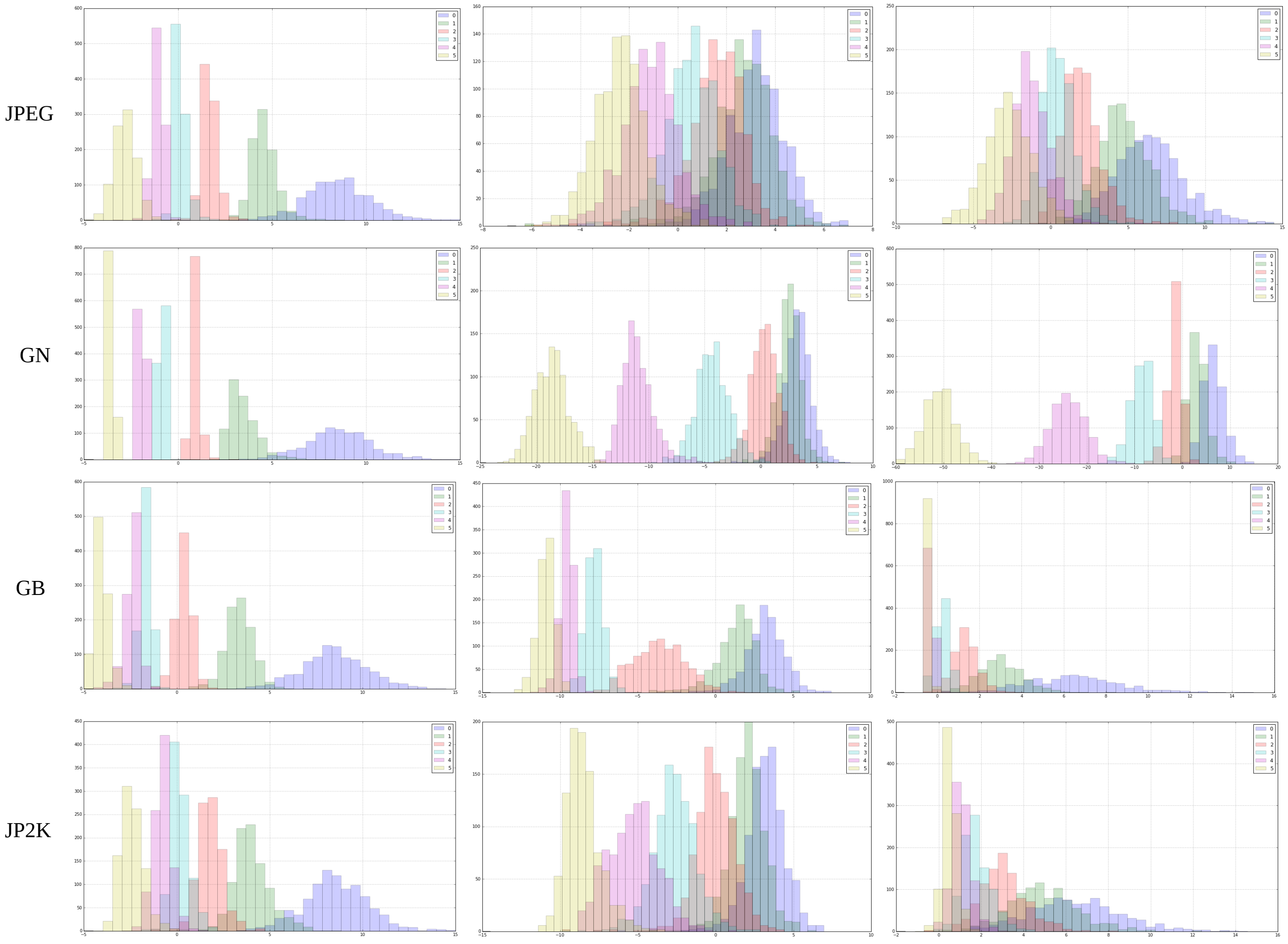
The conv1 filters of AlexNet network are shown in Fig. 15. We see that the properties of four distortions like noise, blocks,



(a) The performance of standard Siamese network

(b) The performance of fast Siamese network

Fig. 13: Ranking results of standard and fast Siamese networks. Both networks are trained for the same number of iterations.



(1) VGG-16 network

(2) Alex network

(3) Shallow network

Fig. 14: The ranking results on different distortions using VGG-16, AlexNet and Shallow networks.

blurs are learned from the data. These filters are significantly different from the filters learned by classification tasks [25].

F. Comparison of our approach with baseline

In this experiment, we use the best performing network VGG-16 as shown by the previous experiments. In Table III we compare the results of our approach with a baseline. In the baseline, it is not always true that deeper networks improve performance. AlexNet performance is worse than Shallow network since the LIVE dataset is not large enough to learn all the

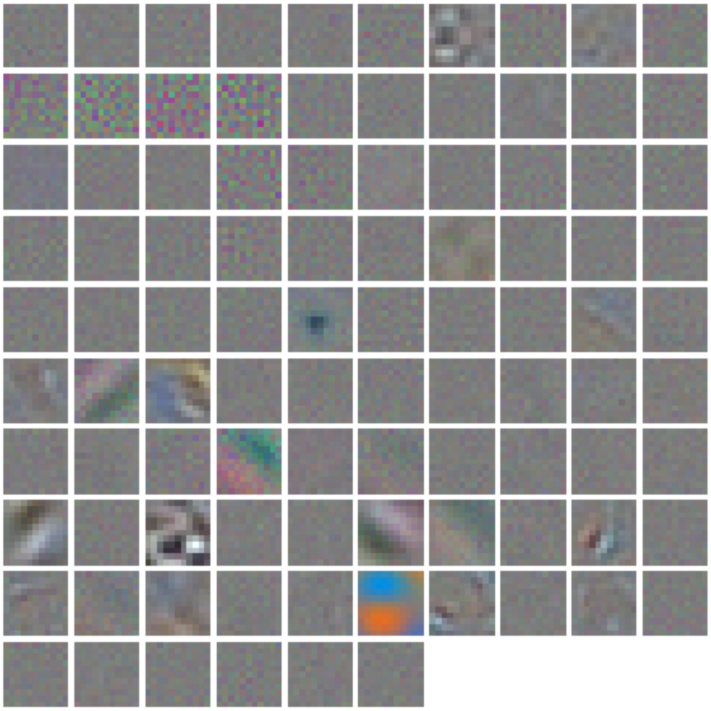


Fig. 15: Visualization results of the conv1 filters from AlexNet trained on four distortions.

TABLE III: SROCC and LCC results of baseline and our approach on the LIVE dataset. SC is the baseline and corresponds to results by models trained from scratch to directly estimate image quality. FT means the results from networks trained for raking using our approach, and then fine-tuned to estimate absolute image quality.

	Shallow	**AlexNet**	**VGG-16**
SC	0.858	0.808	0.926
FT	0.941	0.951	0.973

(a) SROCC evaluation

	Shallow	**AlexNet**	**VGG-16**
SC	0.831	0.801	0.915
FT	0.930	0.941	0.970

(b) LCC evaluation

parameters contained in AlexNet. However, VGG-16 is much deeper network than Alex and performs better than both Shallow and AlexNet, which indicates that the architecture of VGG-16 learns more efficiently.

By using the model fine-tuned from the ranking networks, it is interesting that AlexNet performs better than Shallow because of the good representation learned in the ranking model. The best result is obtained by VGG-16, which again shows that a deeper network with suitable initialization leads to the success of our approach.

More specifically, for SROCC, the Shallow network obtains 5% more than AlexNet but about 7% less than VGG-16 using the SC baseline method. It is notable that the Shallow network improves about 8% using fine-tuning after learning to rank, which is even 1.5% better than the VGG-16 trained from scratch. By taking advantage of the deeper representation, AlexNet network surpasses the Shallow network by 1%. Finally, the best results are obtained using VGG-16. Similar results are found when using LCC for evaluation.

The LCC changes with increasing number of iterations for the baseline and our approach are shown in Fig. 16. Our approach achieves a steep increase at the beginning of iterations and keeps going up and converges at a higher level. While the performance of the baseline is worse than our approach, the LCC increases slowly and finally stabilizes at relative lower value than our approach.

G. *Comparison of our approach with the state-of-the-art*

We compare the performance of our proposed method with start-of-the-art methods from the literature using VGG-16 network. We compare with FR-IQA methods including PSNR, SSIM [14] and PSIM [13], and with NR-IQA methods like DIVINE [49], BLIIDNS-II [17], BRISQUE [19], CORNIA [29], CNN [32] and SOM [31].

Evaluation on LIVE We follow the protocol of [32]. Firstly we randomly split the LIVE dataset into 80% training samples and 20% testing samples and repeat the process 10 times, and then compute the average of LCC and SROCC as the final

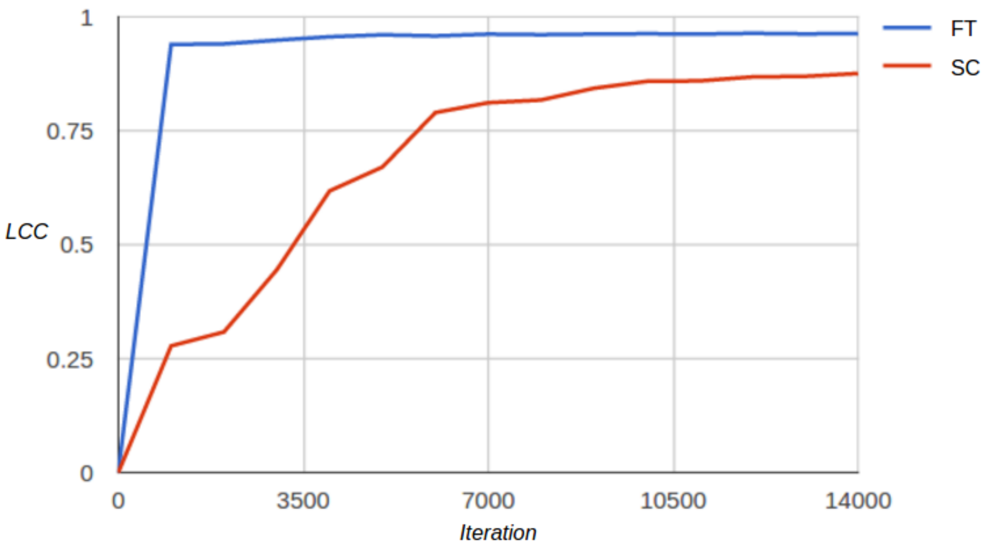


Fig. 16: The LCC changes with iteration increasing using VGG-16 for model trained from scratch as SC (the baseline) and model fine-tuned from ranking model as FT (our approach).

TABLE IV: LCC evaluation on the LIVE dataset.

LCC	JP2K	JPEG	GN	GB	FF	ALL
PSNR	0.873	0.876	0.926	0.779	0.87	0.856
SSIM	0.921	0.955	0.982	0.893	0.939	0.906
PSIM	0.91	0.985	0.976	0.978	0.912	0.96
DIVINE	0.922	0.921	0.988	0.923	0.888	0.917
BLIIDNS-II	0.935	0.968	0.98	0.938	0.896	0.93
BRISQUE	0.923	0.973	0.985	0.951	0.903	0.942
CORNIA	0.951	0.965	0.987	0.968	0.917	0.935
CNN	0.953	0.981	0.984	0.953	0.933	0.953
SOM	0.952	0.961	0.991	0.974	**0.954**	0.962
Ours	**0.976**	**0.987**	**0.995**	**0.988**	–	**0.970**

results, which are shown in Table IV and Table V, respectively. The best method for each dataset is highlighted in bold. We did not show the performance on the FF distortion since we only consider the other four distortions in this work. The column indicated with ALL means we combine all five distortions together on LIVE dataset to train and test the model. In our case, for the fair comparison with other start-of-the-art methods, we train our ranking model on four distortions except FF, but we fine-tune our model on all five distortions on LIVE dataset. Our approach achieves about 1% better than the best results reported before on ALL dataset, which indicates that our methods excellently outperforms existing work including the current start-of-the-art method SOM from CVPR 2015 [31].

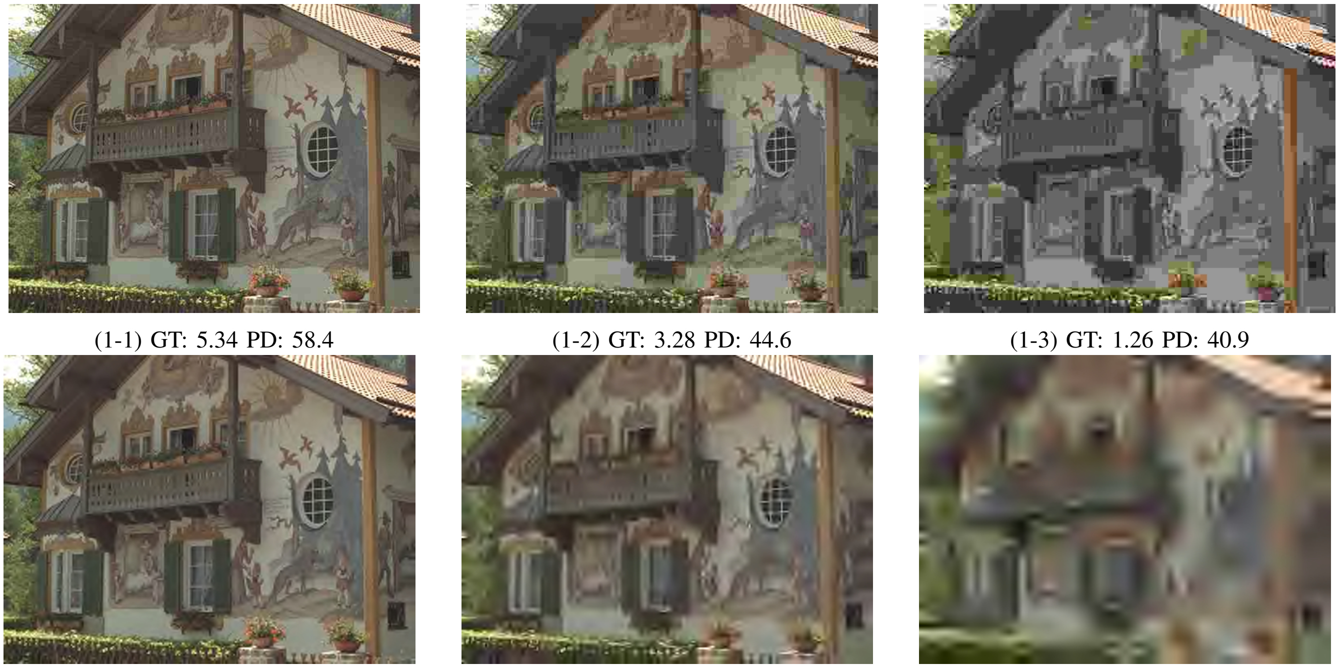
Our approach surpass all methods on all different distortions, with the exception of SROCC on JPEG distortions. More specifically, for LCC our approach boosts accuracy by about 2% on JP2K and GB distortions. It achieves 0.4% improvement

TABLE V: SROCC evaluation on LIVE dataset.

SROCC	JP2K	JPEG	GN	BLUR	FF	ALL
PSNR	0.87	0.885	0.942	0.763	0.874	0.866
SSIM	0.939	0.946	0.964	0.907	0.941	0.913
PSIM	0.97	0.981	0.967	0.972	0.949	0.964
DIVINE	0.913	0.91	0.984	0.921	0.863	0.916
BLIIDNS-II	0.929	0.942	0.969	0.923	0.889	0.931
BRISQUE	0.914	0.965	0.979	0.951	0.887	0.94
CORNIA	0.943	0.955	0.976	0.969	0.906	0.942
CNN	0.952	**0.977**	0.978	0.962	0.908	0.956
SOM	0.947	0.952	0.984	0.976	**0.937**	0.964
Ours	**0.963**	0.958	**0.990**	**0.985**	–	**0.973**

TABLE VI: The results of different methods on TID2008. The methods are trained on LIVE dataset

	BRISQUE	CORNIA	CNN	SOM	**Ours**
SROCC	0.882	0.892	0.920	**0.923**	0.907
LCC	0.892	0.880	0.903	0.899	**0.932**



(1-1) GT: 5.34 PD: 58.4

(1-2) GT: 3.28 PD: 44.6

(1-3) GT: 1.26 PD: 40.9

(2-1) GT: 4.20 PD: 55.2

(2-2) GT: 2.45 PD: 42.7

(2-3) GT: 0.16 PD: 38.2

Fig. 17: Comparison of ground truth (GT) and predicted score (PD) on JPEG and JP2K distortions at varying levels. The first row gives the JPEG distortion varying over three levels, and the second row gives JP2K at the same levels of distortion. Ground truth is in the range [0,9] and the predicted score is in range of [0,100] (the higher the score, the higher the quality of the image).

on GN, which is the most simple task in this problem. The improvement achieved on the JPEG distortion is 0.6%. Similar conclusions are obtained for SROCC, which is slightly lower than the LCC. To the best of our knowledge, it is the first time that NR-IQA methods surpass the performance of FR-IQA methods on all LIVE distortions using the LCC evaluation method.

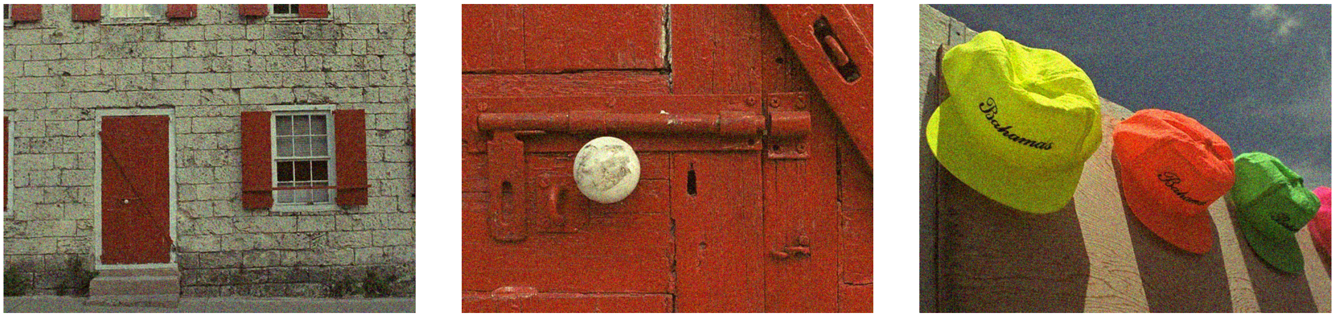
Cross Dataset Evaluation We performed another experiment to demonstrate the generalization of our approach. We trained our ranking network on four distortions and fine-tune the pre-trained models on the LIVE dataset using JPEG, JP2K, GB, and GN (the distortions shared with the TID2008 dataset). Then, the corresponding four distortions in TID2008 dataset are tested. We follow the protocol of [32], 80% of predictions are used to map the predicted scores to a certain range by a non-linear logistic function:

$$Q_p = \beta_1 \left(\frac{1}{2} - \frac{1}{1 + \exp(\beta_2 (Q - \beta_3))} \right) + \beta_4 Q + \beta_5, \tag{11}$$

where Q is the quality scores obtained by IQA methods and Q_p is the human subjects MOS. The remaining 20% are used to predict the scores using the predicted parameters in Eq. 11. We average the results over 100 random 80/20% splits to obtain the final results shown in Table VI. From this we see that we achieve 3% improvement for LCC evaluation and 1.5% lower than the best reported result on the TID2008 dataset.

To explore the reasons why we obtain the best result for LCC but lower result for SROCC compared to start-of-the-art, we compared the ground truth and predicted score on JPEG and JP2K distortions. Fig. 17 is an example of the results. We see that the predicted scores are well-correlated to the ground truth on both JPEG and JP2K. The order of predicted scores of images are also correct in terms of the ground truth. However, the difference among the same type of distortion is larger than that among different types of distortions, which means it is easier to make mistakes among different types of distortion. This is reasonable since when we train our ranking networks, only the pairs among same type of distortion are generated. The rankings among different types of distortion are learned automatically from the ranking networks and then fine-tuned on LIVE dataset. The results in Fig. 17 are quite good in light of this.

A comparison of the same distortion among different images is shown in Fig. 18. Notice that the ground truth of different images with same distortion is quite similar, which is difficult to estimate even for human beings. In this case, it is useless to



(1-1) GT: 4.31 PD: 60.7 (1-2) GT: 4.00 PD: 57.6 (1-3) GT: 4.20 PD: 59.5

Fig. 18: Comparison of ground truth (GT) and predicted score (PD) on GN distortions. The images are distorted by the same level of Gaussian noise. Ground truth is in the range [0,9] and the predicted score is in range of [0,100] (the higher the score, the higher the quality of the image).

force the IQA methods to give a 100% percent correct answer. If the predicted scores of images with same level distortion are in a very small range, the results are acceptable for our system. In conclusion, LCC is more sensitive to the different levels of distortion that are easy to predict using our approach, which explains why we obtain the best results for LCC. Even though the ranking between same distortion (but varying levels) is easy to predict, SROCC is more sensitive to the rankings among different images and different types of distortion with the same level of distortion, which is challenging for our approach. The SOM method achieves the best results for SROCC because semantic obviousness is considered and contributes more to SROCC than other approaches. We believe semantic obviousness method (SOM) can be complementary to our approach and be combined with our approach to improve the performance.

H. Generalization of our ranking networks

The final experimental objective is to verify that our framework is independent of the Waterloo dataset. This means we can achieve similar results using other datasets (like Places2) to train our ranking models with VGG-16 network. In this experiment, we conduct the previous experiment only changing the dataset from Waterloo to validation set of the Places2 dataset, which is not related to the IQA problem.

The ranking networks are trained on ranking sets generated using validation dataset of Places2 dataset used as classification dataset. To show the generalization of our ranking networks, we test the trained model on part of the Waterloo dataset. The ranking results are illustrated in Fig. 19. What is surprising is that the model can well-distinguish different types and levels of distortions on Waterloo dataset, since the collecting process and the scenes of the two datasets are totally different.

TABLE VII: SROCC and LCC results of models trained on Waterloo dataset as “Ours-W” and Places2 as “Ours-P”.

SROCC	JP2K	JPEG	GN	GB	ALL
Ours-W	0.963	0.958	0.990	0.985	0.973
Ours-P	0.963	0.961	0.990	0.986	0.971

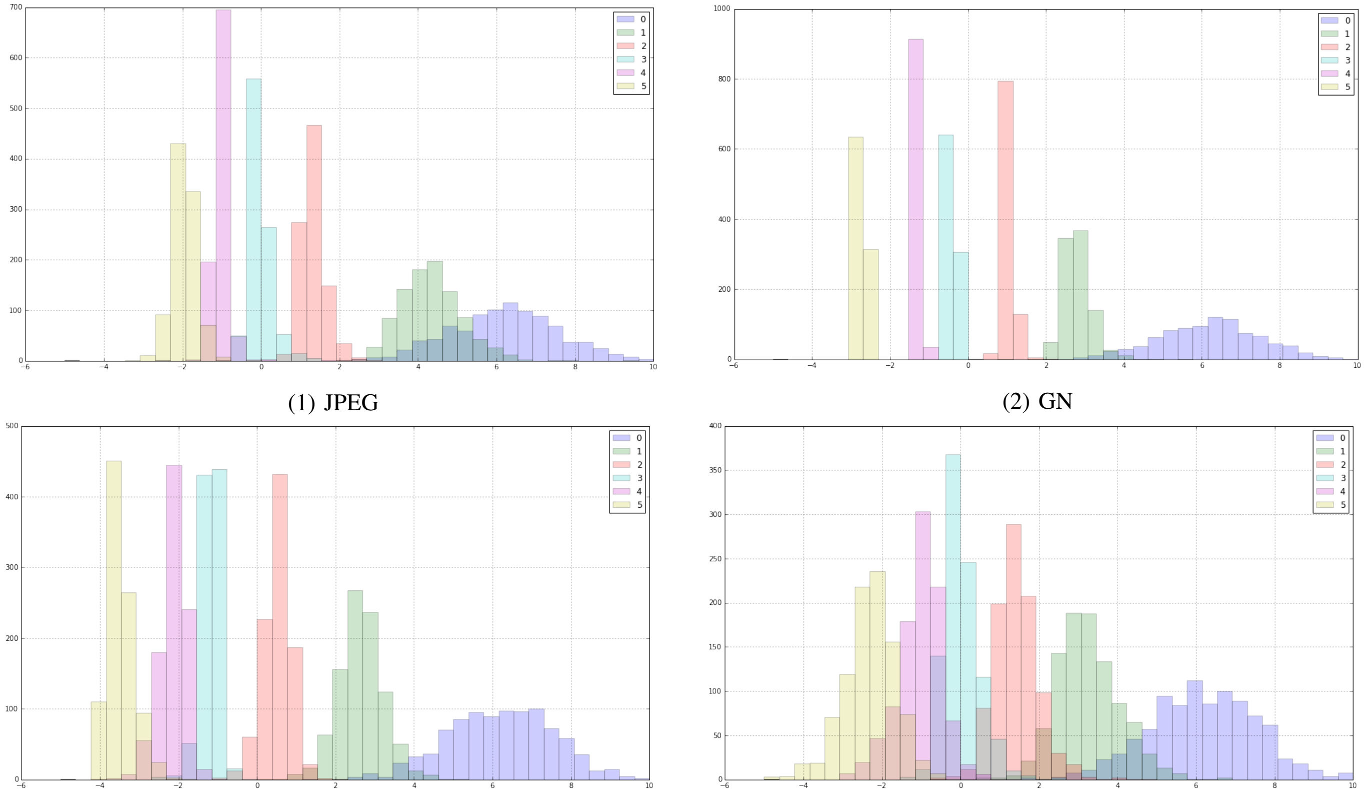
LCC	JP2K	JPEG	GN	GB	ALL
Ours-W	0.976	0.987	0.995	0.988	0.970
Ours-P	0.978	0.983	0.996	0.990	0.975

The final image quality scores are predicted by fine-tuning this network on the LIVE dataset. The performance of this model is compared with the results trained on Waterloo in Table VII. The SROCC and LCC values are very similar, demonstrating the generalization of our approach.

V. Conclusions and future work

In this thesis we developed several techniques to improve the start-of-the-art in image quality assessment. To address the absence of large data sets for image quality assessment, Siamese Networks are trained to rank images in term of image quality by using ranked image sets which can be automatically generated without the use of laborious human labelling and whose relative image quality is known. We then use fine-tuning to transfer the knowledge represented by the trained Siamese Network to a traditional CNN that is able to estimate absolute image quality from single images. To solve the difficulty of pair selection for Siamese network training, we demonstrate how our approach can be made significantly more efficient than traditional Siamese Networks by forward propagating a batch of images through a single network and backpropagating gradients derived from all pairs of images in the batch.

In the experiments we show that the main contribution – the idea to learn from automatically generated rankings – provides considerable performance gains (see Table III). Our second contribution, which is our proposed efficient fast Siamese backpropagation method, improves over the traditional Siamese network both in convergence speed and optimal solution



(1) JPEG

(2) GN

(3) GB

(4) JP2K

Fig. 19: Ranking performance on Waterloo. The models were trained on the validation dataset of Places2. The ranking results obtained are on Waterloo dataset.

(see Fig. 12). In a comparison with state-of-the-art on the LIVE dataset our approach is demonstrated to be superior to the existing NR-IQA techniques (see Table IV and Table V). Furthermore, we are the first NR-IQA method to surpass the state-of-the-art full-reference IQA (FR-IQA) methods. Moreover, we performed a cross dataset experiment, training on the LIVE dataset and testing on the on TID2008 dataset (see Table VI). These experimental results show the generalization ability of our IQA method. Apart from that, we train our ranking model on validation dataset of Places2 to show our approach performs well even using datasets which are not designed specificly for IQA problem (see Table VII).

For future work, we will study the behaviour of taking advantage of new techniques such as batch normalization [50] to accelerate our training process and boost the performance of the VGG-16 network. Moreover, We only train our network for 4 different types of distortions in this thesis. It is easy to extend to more different types of distortions for our approach, which would make it useful for a wider range of applications. Since the real-world distortions are more challenging and complicated, exploring an approach to solve this problem will be difficult but significant. Moreover, we will make our code and model publicly available to support further progress on IQA problem.

Acknowledgments

I would like to thank my supervisors Joost Van de Weijer and Andrew D. Bagdanov for their valuable ideas, suggestions and help during the thesis from the details of experimental design to rules and technique of writing. I would also like to thank our group meetings held every week and our group members for discussing and sharing knowledge. I would like to give a special thank to my team members of master courses Akhil, Enric and Oguz. I would like to thank my girlfriend Lu for supporting me as well as taking care of my life.

References

[1] Z. Wang, A. C. Bovik, and L. Lu, "Why is image quality assessment so difficult?" in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 4. IEEE, 2002, pp. IV–3313.

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

[3] 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*, vol. 10, no. 4, pp. 30–45, 2009.

[4] A. K. Katsaggelos, *Digital image restoration*. Springer Publishing Company, Incorporated, 2012.

[5] J. Van Ouwerkerk, "Image super-resolution survey," *Image and Vision Computing*, vol. 24, no. 10, pp. 1039–1052, 2006.

[6] J. Yan, S. Lin, S. B. Kang, and X. Tang, "A learning-to-rank approach for image color enhancement," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 2987–2994.

[7] J. Chen, Y. Deng, G. Bai, and G. Su, "Face image quality assessment based on learning to rank," *Image and Vision Computing*, vol. 22, no. 1, pp. 90–94, 2015.

[8] S. Aja-Fernandez, R. San-José-Estépar, C. Alberola-Lopez, and C.-F. Westin, "Image quality assessment based on local variance," in *Proc. 28th Annu. IEEE Int. Conf. Engineering in Medicine and Biology Society (EMBS 2006)*, 2006, pp. 4815–4818.

[9] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: assessing the quality of resulting images," *Photogrammetric engineering and remote sensing*, vol. 63, no. 6, pp. 691–699, 1997.

[10] M. Katsura, J. Sato, M. Akahane, I. Matsuda, M. Ishida, K. Yasaka, A. Kunimatsu, and K. Ohtomo, "Comparison of pure and hybrid iterative reconstruction techniques with conventional filtered back projection: image quality assessment in the cervicothoracic region," *European journal of radiology*, vol. 82, no. 2, pp. 356–360, 2013.

[11] Z. Wang and A. C. Bovik, "Modern image quality assessment," *Synthesis Lectures on Image, Video, and Multimedia Processing*, vol. 2, no. 1, pp. 1–156, 2006.

[12] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "A comprehensive evaluation of full reference image quality assessment algorithms," in *Image Processing (ICIP), 2012 19th IEEE International Conference on*. IEEE, 2012, pp. 1477–1480.

[13] ——, "Fsim: a feature similarity index for image quality assessment," *IEEE transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[14] 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*, vol. 13, no. 4, pp. 600–612, 2004.

[15] A. Rehman and Z. Wang, "Reduced-reference image quality assessment by structural similarity estimation," *Image Processing, IEEE Transactions on*, vol. 21, no. 8, pp. 3378–3389, 2012.

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

[17] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the dct domain," *Image Processing, IEEE Transactions on*, vol. 21, no. 8, pp. 3339–3352, 2012.

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

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

[20] Q. Yan, Y. Xu, and X. Yang, "No-reference image blur assessment based on gradient profile sharpness," in *Broadband Multimedia Systems and Broadcasting (BMSB), 2013 IEEE International Symposium on*. IEEE, 2013, pp. 1–4.

[21] S. A. Golestaneh and D. M. Chandler, "No-reference quality assessment of jpeg images via a quality relevance map," *Signal Processing Letters, IEEE*, vol. 21, no. 2, pp. 155–158, 2014.

[22] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, "Live image quality assessment database." [Online]. Available: http://live.ece.utexas.edu/research/quality

[23] E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, pp. 011 006–011 006, 2010.

[24] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[27] K. Sharifi and A. Leon-Garcia, "Estimation of shape parameter for generalized gaussian distributions in subband decompositions of video," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 5, no. 1, pp. 52–56, 1995.

[28] A. Chetouani, A. Beghdadi, S. Chen, and G. Mostafaoui, "A novel free reference image quality metric using neural network approach," in *Proc. Int. Workshop Video Process. Qual. Metrics Cons. Electrn*, 2010, pp. 1–4.

[29] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1098–1105.

[30] P. Ye and D. Doermann, "No-reference image quality assessment using visual codebooks," *Image Processing, IEEE Transactions on*, vol. 21, no. 7, pp. 3129–3138, 2012.

[31] 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*, 2015, pp. 2394–2402.

[32] 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*, 2014, pp. 1733–1740.

[33] ——, "Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2791–2795.

[34] S. Bianco, L. Celona, P. Napoletano, and R. Schettini, "On the use of deep learning for blind image quality assessment," *arXiv preprint arXiv:1602.05531*, 2016.

[35] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1717–1724.

[36] W. Chen, T.-Y. Liu, Y. Lan, Z.-M. Ma, and H. Li, "Ranking measures and loss functions in learning to rank," in *Advances in Neural Information Processing Systems*, 2009, pp. 315–323.

[37] D. Cossock and T. Zhang, "Statistical analysis of bayes optimal subset ranking," *Information Theory, IEEE Transactions on*, vol. 54, no. 11, pp. 5140–5154, 2008.

[38] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005, pp. 89–96.

[39] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: from pairwise approach to listwise approach," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 129–136.

[40] D. Sculley, "Large scale learning to rank," in *NIPS Workshop on Advances in Ranking*, 2009, pp. 1–6.

[41] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 539–546.

[42] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4353–4361.

[43] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Similarity-Based Pattern Recognition*. Springer, 2015, pp. 84–92.

[44] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1386–1393.

[45] F. Gao, D. Tao, X. Gao, and X. Li, “Learning to rank for blind image quality assessment,” *Neural Networks and Learning Systems, IEEE Transactions on*, vol. 26, no. 10, pp. 2275–2290, 2015.

[46] K. Ma, Q. Wu, Z. Wang, Z. Duanmu, H. Yong, H. Li, and L. Zhang, “Group mad competition- a new methodology to compare objective image quality models.”

[47] B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva, “Places: An image database for deep scene understanding,” *arXiv preprint*, 2016.

[48] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[49] A. K. Moorthy and A. C. Bovik, “Blind image quality assessment: From natural scene statistics to perceptual quality,” *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, 2011.

[50] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *arXiv preprint arXiv:1502.03167*, 2015.