Deep Convolutional Neural Models for Picture Quality Prediction

Jongyoo Kim, *Member, IEEE,* Hui Zeng, Deepti Ghadiyaram, *Student Member, IEEE,*
Sanghoon Lee, *Senior Member, IEEE,* Lei Zhang, *Senior Member, IEEE,*
and Alan C. Bovik, *Fellow, IEEE,*

Abstract

Convolutional neural networks (CNNs) have been shown to deliver standout performance on a wide variety of visual information processing applications. However, this rapidly developing technology has only recently been applied with systematic energy to the problem of picture quality prediction, primarily because of limitations imposed by a lack of adequate ground truth human subjective data. This situation has begun to change with the development of promising data-gathering methods that are driving new approaches to deep learning-based perceptual picture quality prediction. Here we assay progress in this rapidly evolving field, focussing in particular on new ways to collect large quantities of ground truth data, and on recent CNN-based picture quality prediction models that deliver excellent results on a large real-world picture quality database.

Index Terms

Convolutional neural network, deep learning, perceptual picture quality, image quality assessment, no-reference image quality assessment.

I. Introduction

Recent years have seen significant efforts applied to the development of successful models and algorithms that can automatically and accurately predict the perceptual quality of 2D and 3D digital images

J. Kim, and S. Lee (corresponding author: Sanghoon Lee) are with the Department of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea, 120-749 (e-mail: {jongky, slee}@yonsei.ac.kr).

H. Zeng and L. Zhang are with the Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong. e-mail: {cshzeng, cslzhang}@comp.polyu.edu.hk).

D. Ghadiyaram, and A. C. Bovik are with the Laboratory for Image and Video Engineering, Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712-1084 USA (e-mail: deepti@cs.utexas.edu, bovik@ece.utexas.edu).

and videos as reported by human viewers [1]. Concurrently, there has been a tremendous surge of work on exploiting large datasets of annotated image data as inputs to deep neural networks, towards solving such challenging problems as image classification and recognition [2]. These efforts have often produced dramatic improvement relative to the state-of-the-art. It is perhaps unsurprising that very deep models having universal representation capability, should produce excellent results when trained on massive datasets using fast graphical computing architectures. Nevertheless, the generalization capability of these models is remarkable.

Yet, until recently, there has been limited effort directed towards optimizing picture quality prediction models using deep networks, although in principal, this could also lead to greatly improved performance. The practical significance of the problem, and the relative ease of implementing algorithms learned on deep architectures make this a compelling topic. Indeed, the explosive consumption of visual media in recent years, owing to advances in digital camera technology, digital television, streaming video services, and social media applications, is driving a critical need for improved picture quality monitoring. The pipelines from picture content generation to consumption are fraught with numerous sources of distortions, including blur, noise, and artifacts arising from such processes as compression, scaling, format conversion, color modification, and so on. Multiple interacting distortions are often present, which greatly complicates the problem. Picture quality models that can accurately predict human quality judgments can be used to greatly improve consumer satisfaction, via automatic monitoring of the qualities of massively distributed pictures and videos, and to perceptually benchmark picture processing algorithms such as compression engines, denoising algorithms, and super-resolution systems that substantially affect viewed picture quality. While many successful picture quality models have been devised, the problem is hardly solved, and there remains significant scope for improvement [3]. Deep learning engines offer a potentially powerful framework for achieving sought-after gains in performance; however, as we shall explain, progress has been limited by a lack of adequate amounts of distorted picture data and ground truth subjective quality scores, which are much harder to acquire than other kinds of labeled image data. Further, typical data-augmentation strategies such as those used for machine vision are of little use on this problem.

A. Perceptual Picture Quality Prediction

Picture quality models are generally classified according to whether a pristine reference image is available for comparison. Full-reference and reduced-reference models assume that a reference is available; otherwise, the model is no-reference, or blind. Reference models are generally deployed when a process is applied to an original image, such as compression or enhancement. No-reference models are applied when the quality of an original image is suspect, as in a source inspection process, or when analyzing the output of a digital camera. Generally, no-reference prediction is a more difficult problem.

Both reference and no-reference picture quality models rely heavily on principles of computational visual neuroscience, and/or on highly regular models of natural picture statistics [1]. Heretofore, the most successful no-reference models have relied on powerful, but shallow regression engines to achieve results that approach the prediction accuracy of reference quality predictors.

B. Deep Learning and Convolutional Neural Networks

Deep learning has had a transformative impact on such difficult problems as speech recognition and image classification, achieving improvements in performance that are significantly superior to those obtained using conventional model-based methods optimized using shallower networks. In particular, most of the top-ranked image recognition and classification systems have been optimized using convolutional neural networks (CNNs). One of the principal advantages of deep learning models are the remarkable generalization capabilities that they can acquire when they are trained on large scale labeled datasets. Whereas models learned using conventional machine learning methods are heavily dependent on the determination of, and discrimination capability of sophisticated training features, by contrast, deep learning models employ multiple levels of linear and nonlinear transformations to generate highly general data representations, greatly decreasing dependence on the selection of features, which are often reduced simply to raw pixel values [2], [4]. In particular, deep CNNs optimized for image recognition and classification have greatly outperformed conventional methods. Open source frameworks such as TensorFlow [5] have also greatly increased the accessibility of deep learning models, and their application to diverse image processing and analysis problems has greatly expanded.

Unlike traditional neural networks, CNNs can be adapted to effectively process high-dimensional, raw image data such as RGB pixel values. Two key ideas underlie a convolutional layer: local connectivity and shared weights. Each output neuron of a convolutional layer is computed only on a locally connected subset of the input, called a local receptive field (drawing from vision science terminology). However, by stacking multiple convolutional layers, the effective receptive fields may enlarge to capture global picture characteristics. Usually, the parameters in a layer (i.e. filter weights) are shared across the entire visual field to limit their number. A common conception is that CNNs resemble precessing by neurons in visual cortex. This idea largely arises from the observation that, in deep convolutional networks deploying many layers of adaptation on images, early layers of processing often resemble the profiles of low-level cortical neurons in area V1, viz., directionally tuned Gabor filters [6], or neurons in visual area V2 implicated in

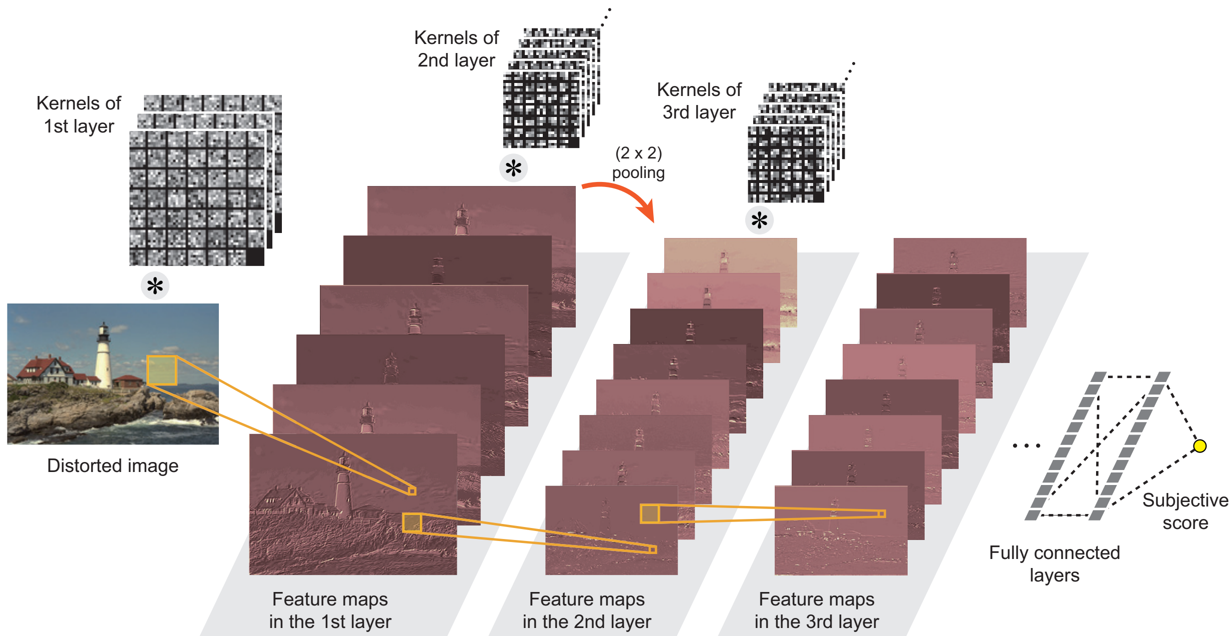

Fig. 1. Example of a CNN structure for no-reference picture quality prediction. The model consists of several convolutional layers followed by a few fully-connected layers. An activation function is applied at each output of the neural network processing flow.

assembling low-level representations of image structure [7]. At early layers of network abstraction, these "perceptual" attributes make them appealing tools for adaption to the picture quality prediction problem.

An example of a CNN structure similar to those studied here is shown in Fig. 1, which also illustrates the kernels that are learned and the feature maps that are obtained when the model is trained for the picture quality prediction task. Generally, a CNN model consists of several convolutional layers followed by fully-connected layers. Some convolutional layers may be followed by pooling layers which reduce the sizes of the feature maps. The fully-connected layers are essentially traditional neural networks, where all of the neurons in a previous layer are connected to every neuron in a current layer.

Motivated by the great success of CNNs on numerous image analysis applications, we comprehensively review and analyze the use of deep CNNs on the picture quality prediction problem. The rest of the paper is organized as follows. Section II assays progress on the picture quality prediction problem in regards to machine learning. In particular, the chief obstacle encountered (lack of data) when trying to apply deep CNNs to this problem is studied. Section III reviews previous successful picture quality prediction models which are based on conventional learning approaches. Section IV further studies recent successful no-reference and full-reference CNN-based picture quality predictors. The commonalities and differences of these models are also comprehensively compared and analyzed. Section V is devoted to evaluating several leading CNN-based picture quality predictors. We also study two general baseline models which are benchmarked on five public databases. Section VI concludes the paper with a discussion of possible future directions of CNN-based picture quality prediction research.

II. Overview of The Problem

Machine learning has played an important role in the development of modern picture quality models. Although these models have been largely driven by features drawn from meaningful quantitative perceptual models, mapping them against the wide variety of generally non-linear, often commingled, and poorly understood distortions that occur in practice is a formidable problem. Sophisticated, yet shallow mapping engines such as support vector regressors (SVR), have produced good prediction results (against human quality opinions), yet there remains substantial room for improvement [3], which greatly motivates the study of deep learners for this problem. Figure 2 shows conceptual flow diagrams of reference and no-reference learned picture quality predictors. A major difference of deep CNN models is the lack of a feature extraction stage, although preprocessing steps may still be put to effective use. In a deep CNN, features conducive to effective picture quality prediction are ostensibly learned by the network during the training process. The pre-processing stages may include, for example, color conversion, local debiasing, local (divisive) normalization, or a domain transformation to sparsify [8] or reduce redundancy in the data.

Most popular learned picture quality prediction models operate by regressing an extracted perceptual feature vector onto recorded subjective scores. Typically, shallow regressors such as SVRs, general regression neural networks, or random forests have been used [9]–[11]. A deep CNN model can instead alternate feature extraction and regression stages. High-dimensional input data (raw or pre-processed pixel values) can be fed into the CNN, and over many iterations or epochs of training on a large dataset, useful image representations are learned automatically. In the early layers of a deep CNN, low-level encoding or sparsifying features are learned, possibly followed by intermediate descriptors of feature correlations [7]. In the deeper layers, the learned features contain more abstract information that can capture relationships between image distortions and human perceptions of them. In a CNN, differentiable feature aggregation or pooling stages are interspersed with feature extraction and regression stages, enabling effective end-to-end optimization. However, despite significant successes on a wide array of other image analysis problems, the application of deep learning networks to the picture quality prediction problem has been complicated by a significant obstacle, which is a lack of an adequate amount perceptual training data, including accurate local ground truth scores.

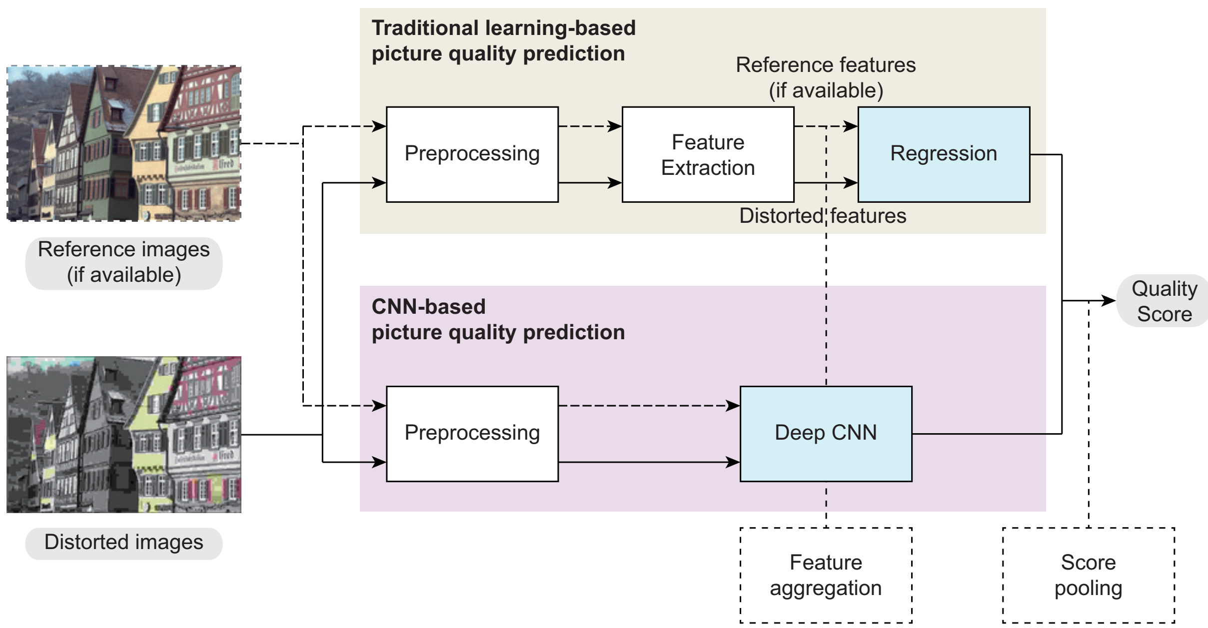

Fig. 2. Flow chart comparisons of traditional learning-based and CNN-based reference and no-reference picture quality models. Blue boxes indicate learning processes.

The performance of deep learning models generally depends heavily on the size of the available training dataset(s). Currently available legacy public-domain subjective picture quality databases such as LIVE IQA [12] and TID2013 [13] are far too small to effectively train deep learning models. For example, the LIVE IQA and TID2013 databases each contain fewer than a microscopic 30 unique image contents, and no more than 24 different types of distortions per image, all of which are synthetic.[1] Even the recent LIVE "In the Wild" Challenge Database (hereafter "LIVE Challenge") [3], the largest available resource in these dimensions (with nearly 1,200 unique picture contents, each afflicted by a unique, unknown combination of highly diverse authentic distortions, and judged by more than 350,000 unique human subjects) is of insufficient size, although it provides an excellent challenge for any no-reference model. By comparison, image recognition datasets such as ImageNet [14] contain tens of millions of labeled images. Creating larger subjective quality datasets is a formidable problem. Controlled laboratory studies like [12], [13] are out of the question, and even the crowdsourced study in [3] exhausted the pool of high-quality human subjects available on Amazon Mechanical Turk.

[1]As applied to pristine images by a database designer. Algorithm-generated distortions such as Gaussian blur, noise, mean shifts (and so on) contained in these databases are poor models of picture impairments that actually arise in consumer digital photographs. Even JPEG/JPEG2000 coded images are created using much more liberal amounts and spreads of compression (to create perceptual separations) than is produced by real image capture devices.

Obtaining adequate quantities of reliable human subjective labels remains a very difficult problem. Unlike the binary (yes/no) confirmations of automatically-generated labels that are delivered by online human subjects, as used in the construction of object recognition datasets like ImageNet [2], each of which might be generated in a second or less, collecting human quality judgments is a complex, time-consuming psychometric task that is as much about assessing each subject's response, as it is of quality labeling the images. The human subjects determine an internal judgment of the overall quality of each image after holistically scrutinizing it, then record each of their judgments on a continuous, sliding subjective quality scale, while consciously discounting factors such as image content or photographic aesthetics. This highly engaged task requires dozens or even hundreds of human quality raters to spend 5-10 seconds on each image. Each subject's overall session is time-limited, to avoid reductions in attention and performance from visual fatigue.

Common strategies for attacking this labeled image paucity are data augmentation techniques, which seek to multiply the effective volume of image data via rotations, cropping, reflections, and so on. Unfortunately, with the likely exception of horizontal reflections, which we use later, applying these kinds of transformations to an image will generally significantly change its perceived quality. While generating large numbers of picture content is simple, ensuring adequate distortion diversity and realism is much harder.

In another common strategy, the images being trained on are divided into many small patches. However, this approach produces another problem, that distinct local ground truth subjective labels are not available for the patches. In every experimental scenario to date, human subjects supply a single scalar subjective score on each global image. Since images, distortions of images, and human perceptions of both are all highly non-stationary, the scores that subjects would apply to a local image patch will generally differ greatly from those applied to the entire image. Obtaining human judgments of local image patch quality is not practical, as it would greatly increase the overhead of acquiring human scores.

One way to try to overcome the lack of an adequate training dataset is to utilize unsupervised learning, e.g., by training a restricted Boltzmann machine or an autoencoder [4] with convolutional layers. With an unsupervised model, it is possible to train deep neural network models on very large datasets having no ground truth labels. However, picture quality prediction is a subtle problem that involves modeling detailed interactions between distortion and content. Conversely, unsupervised models that are designed to work well on tasks such as image recognition, may succeed in part by learning to promote gross shape-related features, while suppressing small variations. For example, a denoising autoencoder can be trained to reconstruct an original image from a noisy one by enforcing robustness against small corruptions of the input data, or by adding a regularization term to the objective function. By contrast, the representations learned by a picture quality predictor must be particularly sensitive to local and global degrees of distortion, as well as to perceived interactions between content and distortion. Successful, generalizable, deep unsupervised picture quality prediction models have not yet been reported.

The need for large-scale subjective picture quality data is underlined by the fact that the perception of picture distortions engages multiple complex processes along the visual pathway, including bandpass, multiscale, and directional decompositions [6], local nonlinearities, and normalization mechanisms. For example, contrast masking [15], whereby the spatially localized energy of image content can reduce or eliminate the visibility of distortions, is well-explained by a local cortical divisive normalization model [16]. Successful reference and no-reference picture quality models [9], [10], [15], [17] approximate these perceptual mechanisms by various models; however, errors in these approximations, along with a lack of information describing other relevant, perhaps higher-level processes, still limit their prediction efficacy [3]. Traces of such human response properties exist and are embedded in human subject data. This suggests that they might be unraveled by a deep network served by enough data.

III. Conventional Learning-Based Picture Quality Predictors

The most successful reference picture quality predictors, such as those deployed by the television industry,[2] are not learned models, but instead compute similarity or error measures modulated by perceptual criteria in some manner. Performance is high since a reference error, whether implicit or explicit, is available to be analyzed using perceptual models. No-reference models operate without the benefit of an implied error signal, so their design has relied heavily on machine learning. Broadly, these models deploy perceptually relevant, low-level feature extraction mechanisms based on simple, yet highly regular, parametric models of good-quality pictures. These “natural scene statistics” (NSS) models are predictably altered by the presence of distortions [18]. Simply stated, high-quality images subjected to bandpass filtering, followed by local energy normalization, become substantially decorrelated and gaussianized, while distorted images tend not to obey this model.[3] Picture quality prediction models of this type have been developed in the wavelet [18], discrete cosine transform, sparse [8] and spatial domains [9], and have also been applied to video signals using natural bandpass space-time video statistics models [19], [20]. The FRIQUEE model [21] achieved state-of-the-art performance on the LIVE Challenge database [3] by regressing on a “bag” of NSS features drawn from diverse color spaces and perceptually-motivated transform domains.

[2] Such as the Emmy-winning SSIM model [15], and the VIF index [18], which is a core element of the VMAF processing system that quality-controls all Netflix content encodes.

[3] Although this is not always the case on authentically distorted pictures, as demonstrated in [3]

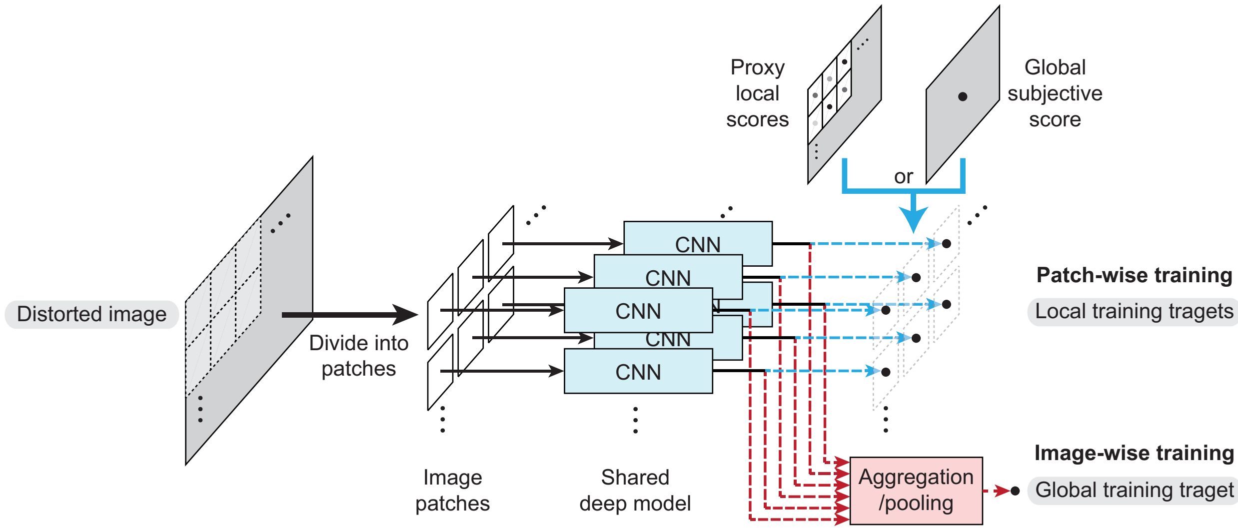

Fig. 3. Patch-wise and image-wise strategies used to train patch-based picture quality prediction models. First, an input image is partitioned into patches, then each is fed into the same CNN model. In patch-wise training, a proxy local score or global subjective score is used as a training target for each input patch. In image-wise training, extracted features or scores are aggregated, then regressed onto a single global subjective score.

There have also been recent attempts to apply other, earlier type of deep learning models to the no-reference picture quality prediction problem. For example, Hou *et al.* trained a deep belief network (DBN) on wavelet domain NSS features to classify distorted images into five discrete score categories [17], and Li *et al.* regressed shearlet NSS features onto subjective scores using a stacked auto-encoder [22]. These models generally used hand-crafted feature inputs, were not trained via end-to-end optimization, and achieved less impressive gains in performance.

IV. CNN-Based Picture Quality Prediction

A. CNN-Based No-Reference Picture Quality Models

As mentioned earlier, several CNN-based picture quality prediction models have attempted to use patch-based labeling to increase the set of informative (ground truth) training samples. Generally, two types of training approaches have been used: patch-wise and image-wise, as depicted in Fig. 3. In the former, each image patch is independently regressed onto its target. In the latter, the patch features or predicted scores are aggregated or pooled, then regressed onto a single ground truth subjective score.

The first application of a spatial CNN model to the picture quality prediction problem was reported in [23], wherein a high-dimensional input image was directly fed into a shallow CNN model without finding handcrafted features. To obtain more data, each input image was subdivided into small patches as a method of data augmentation, each being assigned the same subjective quality score during training. Following prior successful natural scene statistics based models [9], [18], this method applies a process of local divisive normalization on each input image, and uses both max and min pooling to reduce the feature maps. Patch-wise training was used, and during application, the predicted patch scores were averaged to obtain a single picture quality score.

Li *et al.* utilized a deep CNN model that was pre-trained on the ImageNet dataset [24]. A network-in-network (NiN) structure was used to enhance the abstraction ability of the model. The final layer of the pre-trained model was replaced by regression layers which mapped the learned features onto subjective scores. As in [23], image patches were regressed onto identical subjective quality scores during training.

The labeling of local patches with global subjective quality scores during training may be problematic. While the reported prediction accuracy of these model was competitive with that of "handcrafted" feature-based quality prediction models, it is not reasonable to expect local image quality to closely agree with global subjective scores, even when synthetic distortions are applied homogeneously. Picture quality is inevitably space-varying, because of the high degree of non-stationarity of picture contents and the complex perceptual interactions between content and distortions (such as masking). Towards solving this problem, image-wise training strategies have been studied.

Bosse *et al.* deployed a deeper, 12-layer CNN model fed only by raw RGB image patches to learn a no-reference picture quality model [25]. They proposed two training strategies: 'patch-wise training' (similar to [23]), and 'weighted average patch aggregation,' whereby the relative importance of each patch was weighted by training on a subnetwork. The overall loss function was optimized in an end-to-end manner. The authors reported state-of-the-art prediction accuracies on the major synthetic-distortion picture quality databases.

To overcome overfitting problems which can arise from a lack of adequate local ground truth scores, several authors have suggested training deep CNN models in two separate stages: a pre-training stage, using a large number of algorithm-generated proxy ground truth quality scores, followed by a stage of regression onto a smaller set of subjective scores. For example, [26] describes a two-stage CNN-based no-reference quality prediction model, whereby local quality scores generated by a full-reference algorithm are used as proxy patch labels, in a first stage of training. In the second stage, the feature vectors obtained from image patches are aggregated using statistical moments, then regressed onto subjective scores. In this instance, the first stage is patch-wise training, while the second stage is image-wise training. Since the local proxy scores reflect the non-stationary characteristics of perceived quality, they are reasonable local regression targets, and training of the CNN model is enabled by the abundant training samples. Following the second stage of training on human ground truth, their model attains highly competitive prediction accuracy on the legacy datasets.

The same authors later developed a two-stage training scheme for no-reference picture quality prediction called the Deep Image Quality Assessor (DIQA) [27]. The training process of that model was separated into an objective training stage, followed by a subjective training stage. Rather than using a sophisticated picture quality predictor to produce proxy scores, they instead efficiently computed PSNR (or MSE) maps. Using only convolutional layers, feature maps were obtained, which were then regressed onto objective error maps. The second stage aggregated the feature maps by weighted averaging, then regressed these global features onto ground truth subjective scores. The weighting maps were also learned during training. The reported prediction accuracy of these models is competitive with state-of-the-art models on the legacy databases.

B. CNN-Based Full-Reference Picture Quality Models

While CNNs were first used to model no-reference picture quality, more recently, they have been applied to the reference prediction problem as well.

Liang *et al.* [28] proposed a dual-path CNN-based full-reference quality prediction model. They generalized the problem by seeking to predict quality using a non-aligned image of a similar scene as reference. Locally normalized distorted and reference image patches are fed into a dual-path CNN model, each using the same parameter values. Then the concatenated learned feature vectors are regressed onto the subjective scores of source distorted images. They report state-of-the-art prediction accuracies in both aligned and non-aligned full-reference scenarios.

Gao *et al.* deployed a deep CNN model pre-trained on ImageNet. They used it to conduct full-reference picture quality prediction [29] by feeding pairs of reference and distorted pictures into the CNN, where each output layer is used as a feature map. Local similarities between the feature maps obtained from the reference and distorted images are then computed and pooled to arrive at global picture quality scores. The CNN model was not fine-tuned on any picture quality database.

The deep CNN-based full-reference quality prediction model in [30], called DeepQA, was trained to learn a visual sensitivity weight at each coordinate using measured local spatial characteristics of the distorted image. DeepQA accepts the distorted image and an objective error map (e.g. MSE) as inputs. The learned weight map is then used as a multiplier on the objective error map. The authors reported consistent state-of-the-art prediction accuracies as compared to other reference quality models, on the synthetic-distortion legacy picture quality databases.

TABLE I
COMPARISON OF IMPLEMENTATIONS OF CNN-BASED PICTURE QUALITY PREDICTION MODELS. FR (NR) INDICATES FULL-REFERENCE (NO-REFERENCE) QUALITY PREDICTION MODELS. 'CONV' INDICATES CONVOLUTIONAL LAYERS, AND 'FC' INDICATES FULLY-CONNECTED LAYERS.

Models	Type	Layer depth	Preprocessing	Feature aggregation or score pooling
[23]	NR	2 Conv & 2 FC	Local normalization	Mean pooling (during testing)
[24]	NR	14 Conv (4 NiN blocks)	Local normalization	Mean pooling (during testing)
[25]	NR	10 Conv & 2 FC	Raw RGB image	Mean or weighted average pooling
[26]	NR	2 Conv & 6 FC	Local normalization	Mean and standard deviation aggregation
[27]	NR	8 Conv & 3 FC	Low-freq. subtraction	Mean or weighted average aggregation
[28]	FR	(2 Conv, 1 FC)$\times$2 & 2 FC	Local normalization	(Not mentioned)
[29]	FR	13 Conv & 3 FC	Raw RGB image	Mean aggregation and pooling
[30]	FR	(2 Conv)$\times$2, 6 Conv & 2 FC	Low-freq. subtraction	Weighted average aggregation

Models	Type	Training targets		Comments
		1st stage	2nd stage	(Comparison strategy for FR models)
[23]	NR	Subjective scores	N.A.	Patch-wise training
[24]	NR	Semantic label	Subjective scores	Fine-tuning of pre-trained CNN on ImageNet
[25]	NR	Subjective scores	N.A.	Weighted average patch aggregation
[26]	NR	Proxy scores	Subjective scores	Uses proxy patch labels
[27]	NR	Objective error map	Subjective scores	Uses proxy patch labels
[28]	FR	Subjective scores	N.A.	Concatenation of feature vectors
[29]	FR	Semantic label	N.A.	SSIM between feature maps of each layer
[30]	FR	Subjective scores	N.A.	Concatenation of feature maps

C. Summary of CNN-Based Picture Quality Models

Table I compares the implementations of reported CNN-based no-reference [23]–[27] and full-reference [28]–[30] picture quality models. For full-reference models, the strategies used to compare distorted and reference features are summarized in the last column. In [28] and [30], this merely amounts to supplying both to the network. Generally, the reviewed models were designed to overcome the lack of training data, which is the most important issue that needs to be resolved to employ deep CNN models successfully. Most of the models used some type of patch-based training to increase the training data volume. Several of the models used proxy ground truth scores generated by objective quality prediction models to augment the subjective scores, or alternately, to pre-train the network on a large amount of easily generated proxy data before fine-tuning on subjective scores. Since we have found no serious attempts to use unsupervised deep models, we make no comparisons of this type, although the success of the very simple model [31]

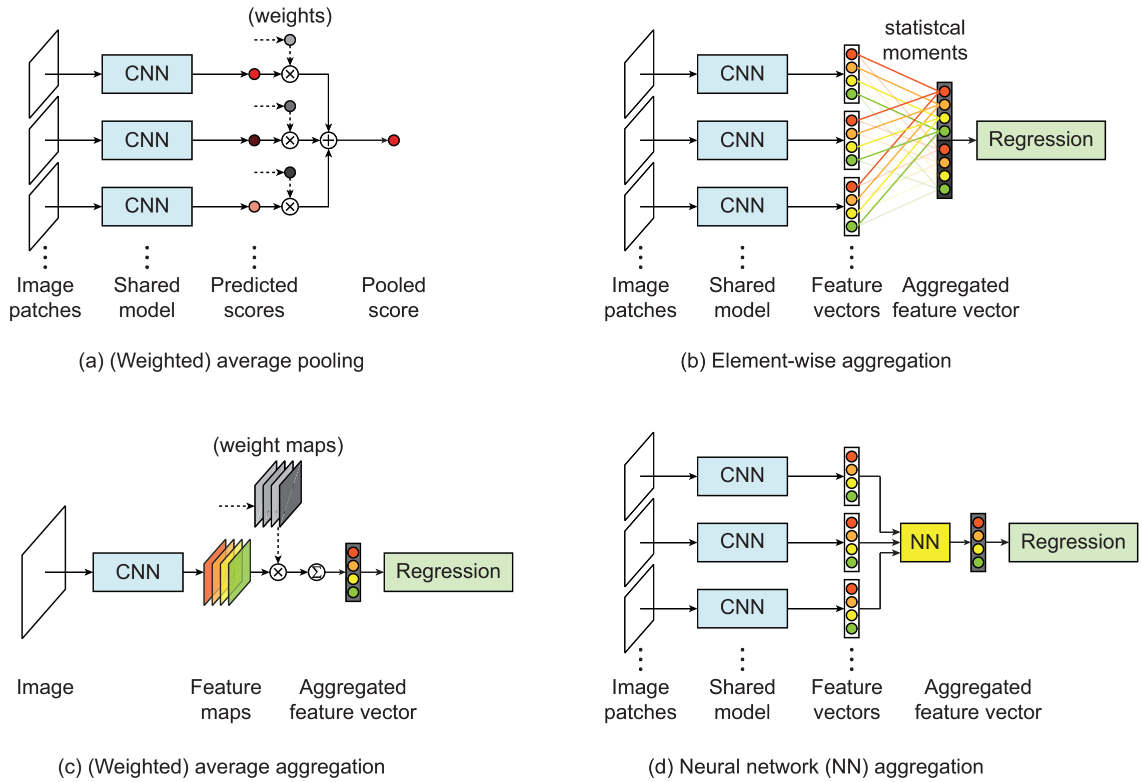

Fig. 4. Examples of aggregation and pooling strategies in CNN-based picture quality prediction models.

suggests this is an interesting research direction. Finding ways to embody models of perception into deep picture quality models is also an issue. While simpler models often use perceptually relevant bandpass processing and local divisive normalization [23], similar processes may be learned by the network at the early stages. However, it should be possible to impose perceptual weighting or pooling strategies on the network to account for aspects of visual sensitivity, which could accelerate the process of training on subjective scores.

In CNN-based schemes, the process of feature aggregation or score pooling determines the form of a loss function. Examples of aggregation and pooling strategies are shown in Fig. 4. The patch-based algorithms described in [23], [24], did not use aggregation or pooling during training. Instead, each image patch was independently regressed onto the global subjective quality score. The loss function used is

$$\mathcal{L} = \frac{1}{N}\sum_{i}^{N} \|f(p_i) - S\|, \tag{1}$$

where p_i refers to the i-th patch obtained, N is the number of patches, S is the ground truth score, and $f(\cdot)$ is a neural network process. The models were trained via a patch-wise optimization, and during testing, the outputs of multiple patches composing an input image were averaged to obtain a final predicted

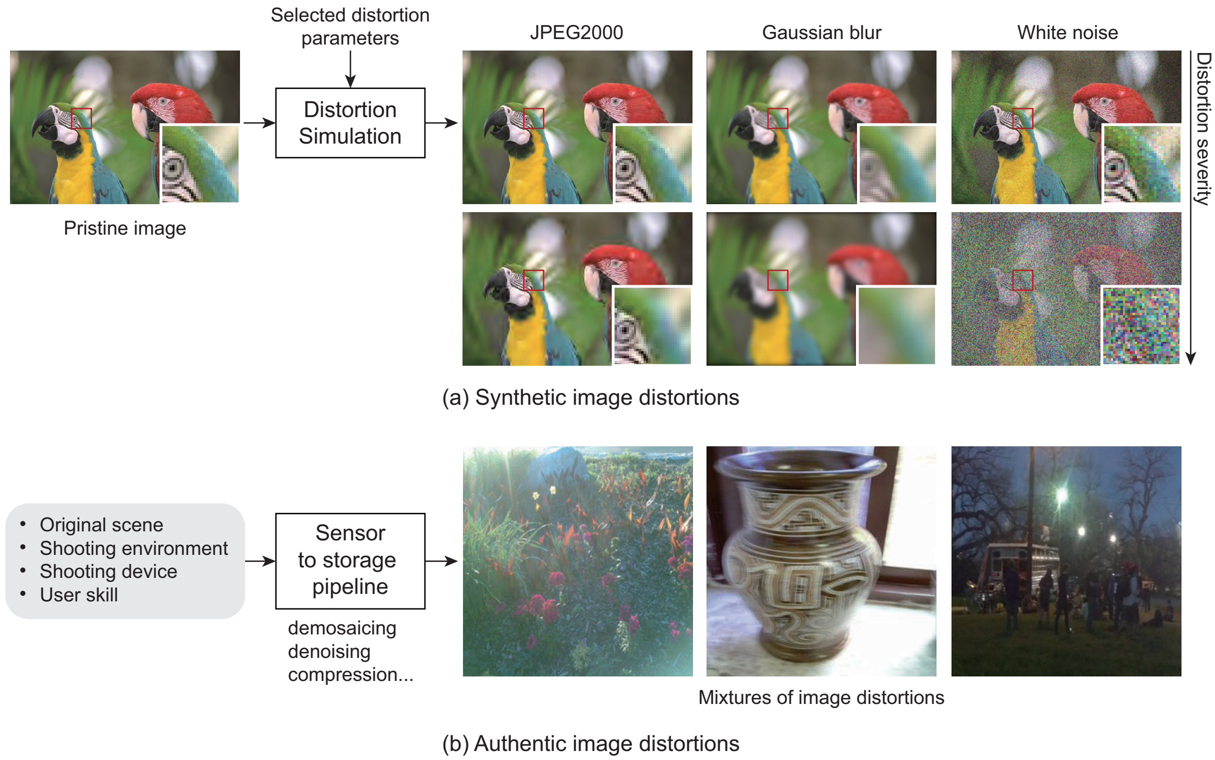

Fig. 5. Synthetic and authentic image distortions found in picture quality databases.

subjective score. Conversely, image-wise approaches use aggregation or pooling during training. For example, weighted average pooling methods [25] may be used, where the loss function looks like

$$\mathcal{L}' = \|pool(f(p_1), ...f(p_N)) - S\| \tag{2}$$

where $pool(\cdot)$ refers an unspecified pooling method (Fig. 4(a)). In [26] and [27] (Figs. 4(b) and (c)), simple feature aggregation was used. A more complicated model, such as a multilayer perceptron or recurrent neural network [4], could also be used for aggregation (Fig. 4(d)). Here, the loss function becomes

$$\mathcal{L}'' = \|g(aggr(f(p_1), ...f(p_N))) - S\| \tag{3}$$

where $aggr(\cdot)$ refers to a feature aggregation process, and $g(\cdot)$ is a regression neural network. The forms (2) and (3) have the advantage that the model can be trained under the same conditions as the actual testing conditions, where the image-wise scores are predicted.

D. Description of Picture Quality Databases

The choice and consideration of database for training is important for deep learning-based models, since their performance depends highly on the size of the training set. In most picture quality databases,

TABLE II
COMPARISON OF IQA DATABASES IN TERMS OF NUMBERS OF REFERENCE (REF.) IMAGES, DISTORTED (DIST.) IMAGES, DISTORTION TYPES, AUTHENTICITY OF DISTORTIONS, TYPE OF SUBJECTIVE SCORES, WHETHER DISTORTIONS ARE MIXED, AND PUBLISHED DATE.

Database	Number of Ref. images	Number of Dist. images	Number of Dist. types	Authenticity of distortions	Subjective score type	Mixtures of distortions	Published date
LIVE IQA [12]	29	779	5	Synthetic	DMOS	N.A.	2003
CSIQ [32]	30	866	6	Synthetic	DMOS	N.A.	2010
TID2013 [13]	25	3,000	24	Synthetic	MOS	N.A.	2015
LIVE MD [33]	15	405	2	Synthetic	DMOS	✓	2012
LIVE Challenge [3]	N.A.	1,162	Numerous	Authentic	MOS	✓	2016

the distorted images are afflicted by only a single type of synthetically introduced distortion, such as JPEG compression, simulated sensor noise, or simulated blur, as exemplified in Fig. 5(a). Since they have played important roles in the development of perceptual picture quality studies, we briefly describe several popular legacy databases.

The LIVE IQA database [12], which was the first successful public-domain picture quality database, and is still the most widely-used, contains 29 reference images and 982 images, each impaired by one of five types of synthetic distortions: JPEG and JPEG2000 (JP2K) compression, white Gaussian noise (WN), Gaussian blur (GB), and Rayleigh fast-fading channel distortion (FF). The differential Mean Opinion Score (DMOS) of each distorted image is provided. The CSIQ database [32] includes 30 reference images and 866 synthetically distorted images of six types: JPEG, JP2K, WN, GB, pink Gaussian noise (PGN), and global contrast decrements (CTD). The DMOS of the distorted images is also provided. TID2013 [13] contains the largest number of distorted images. It consists of 25 reference images and 3,000 synthetically distorted images with 24 different distortions at five levels of degradation. The database also provides the Mean Opinion Scores (MOS). The LIVE MD database [33], was the first to include multiply (synthetically) distorted images. Images in it are distorted by two types of distortion in two combinations: simulated Gaussian blur followed by JPEG compression, and Gaussian blur followed by additive white Gaussian noise. It contains 15 reference and 405 distorted images, and the DMOS of each distorted image is provided.

Lastly, the LIVE Challenge Database [3] contains nearly 1,200 unique image contents, captured by a wide variety of mobile camera devices under highly diverse conditions. As such, the images were subjected to numerous types of authentic distortions during the capture process, often in complex combinations of multiple interacting impairments, as shown in Fig. 5(b). The distortions include, for example, low-light blur and noise, motion blur, camera shake, overexposure, underexposure, a variety of color errors, compression errors and many combinations of these and other impairments. There are no reference images in the LIVE Challenge database, since the distorted images are originals, captured by dozens of ordinary photographers. The LIVE Challenge pictures were judged by more than 8,100 human subjects in a tightly monitored crowdsourced study, yielding more than 350,000 human judgments that exhibit excellent internal consistency [3].

A summary of the attributes of these five databases is shown in Table II.

V. Performances of CNN Picture Quality Models

Since only a few CNN-based picture quality models have been released, we provide the prediction accuracies of baseline models on the five databases as performance references to be compared against. We selected the well-known very deep CNN models, AlexNet [2], and ResNet50 [34], as the architectures of the baseline models, where each was pre-trained on the ImageNet classification task. Both of these pre-trained models are available for download. The specific network configurations can be found in the original papers. For each pre-trained architecture, two types of back-end training strategies were tested: using an SVR to regress the extracted features from the CNN model onto subjective scores, and fine-tuning the pre-trained networks to conduct picture quality prediction. We did not test direct training of these models on any of the picture quality databases, since they are not large enough. Very deep networks easily overfit on insufficient training samples, causing significant decreases in testing accuracy (AlexNet has 62 million and ResNet50 has 26 million parameters). Instead, we tested a smaller CNN network as a baseline model of direct training.

In the first approach, the output of the 6^{th} fully connected layer (4096 dimensions) from AlexNet and averaged-pooled features (2048 dimensions) from ResNet50 were used as the input feature vectors to the SVR. From each input image, 25 randomly cropped image patches (the patch size is pre-defined by the pre-trained models: 227 x 227 for AlexNet, and 224 x 224 for ResNet50) were used for training and testing. The obtained feature vectors from these 25 image patches were averaged to obtain a single global feature vector.

In the second approach, we randomly cropped 100 image patches from each training image to be used for training (except on the TID2013 database, where 30 cropped patches were used, due to the large number of distorted images in the database). The image patches inherited the quality scores from the source distorted images, which were first normalized to the range [0, 1]. This preprocessing enabled us to use the same parameter settings on all databases. The basic regression loss (1) was used. To alleviate

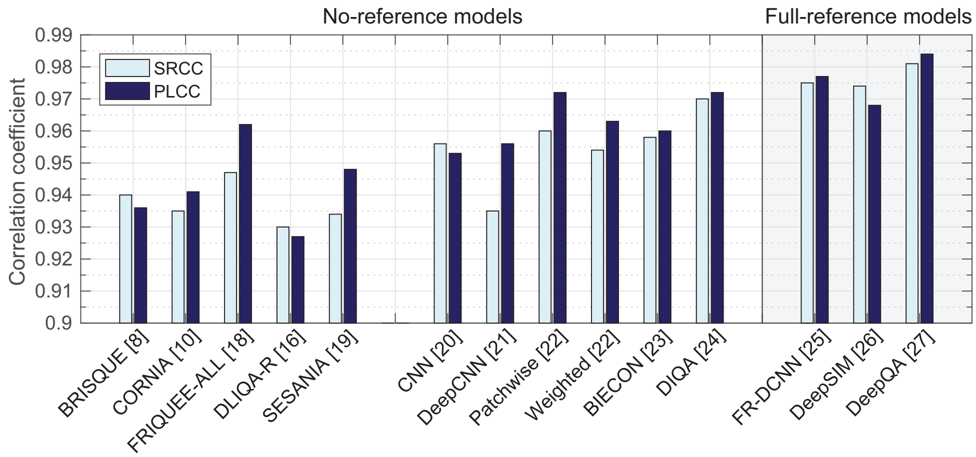

Fig. 6. Comparison of SRCC and PLCC of learned picture quality models on the legacy LIVE IQA database.

overfitting, one dropout layer with dropout rate 0.5 was added before the last fully-connected layer. The learning rate was set to 10^{-3}, and the fine-tuning process iterated for 8 and 6 epochs on AlexNet and ResNet50, respectively. The batch size was fixed at 48 for both models. In the testing stage, the pre-trained models were used to predict quality scores on each of 25 random image crops. These were average pooled to produce the final picture quality scores.

For the direct training approach, we used the following CNN architecture: Conv-48, Conv-48 with stride 2, Conv-64, Conv-64 with stride 2, Conv-64, Conv-64, Conv-128, Conv-128, FC-128, FC-128, and FC-1. Here, “Conv” refers to convolutional layers, “FC” refers to fully-connected layers, and the trailing numbers indicate the number of feature maps (of Conv) or output nodes (of FC). The model accepts 112×112 images as inputs. All of the convolutional layers were configured to use 3×3 filters, using zero-padding to preserve the spatial size. Each layer used a rectified linear unit as the activation function. Following the convolutional layers, each 28×28 feature map (assuming two convolutional layers with a stride of 2) was averaged yielding an 128-dimensional feature vector, which is then fed into the fully-connected layers. The number of parameters in this model is about 0.4 million, which is much lower than AlexNet or ResNet50. This baseline model was trained using the image-wise L_2 loss in (3). Each input image was partitioned into 112×112 patches when training on the LIVE IQA database, while full-sized images were used on the other databases. On the LIVE IQA database, non-overlapping patches were used so that overlapped regions could not be accessed multiple times by the CNN model during training and/or testing. The data was augmented by supplementing the training set with horizontally flipped replicas of each image. Each minibatch contained patches extracted from five images. The training was iterated over 80 epochs.

TABLE III

SRCC and PLCC comparison on five public-domain subjective picture quality databases. FR (NR) indicates full-reference (no-reference) quality prediction models, and italics indicate CNN-based methods. Boldface entries indicate the top three performers on each database for each performance metric.

		LIVE IQA		CSIQ		TID2013		LIVE MD		LIVE Challenge	
Type	Methods	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
FR	PSNR	0.876	0.872	0.806	0.800	0.636	0.706	0.725	0.815	N.A.	N.A.
	SSIM [15]	0.948	0.945	0.876	0.861	0.775	0.691	0.845	0.882	N.A.	N.A.
	FSIMc [35]	0.963	0.960	0.931	0.919	0.851	0.877	0.863	0.818	N.A.	N.A.
	DeepQA [30]	0.981	0.982	0.961	0.965	0.939	0.947	0.938	0.942	N.A.	N.A.
NR	BRISQUE [9]	0.939	0.942	0.756	0.797	0.572	0.651	0.897	0.921	0.607	0.585
	CORNIA [11]	0.942	0.943	0.714	0.781	0.549	0.613	0.900	0.915	0.618	0.662
	FRIQUEE-ALL [21]	0.948	0.962	0.839	0.863	0.669	0.704	**0.925**	**0.940**	0.720	0.720
	BIECON [26]	0.958	0.960	0.815	0.823	0.717	0.762	0.909	**0.933**	0.595	0.613
	DIQA [27]	**0.970**	**0.972**	**0.844**	**0.880**	**0.843**	**0.868**	**0.920**	**0.933**	0.687	0.701
	AlexNet + SVR	0.901	0.908	0.712	0.736	0.263	0.365	0.760	0.803	**0.769**	**0.790**
	ResNet50 + SVR	0.925	0.935	0.654	0.700	0.435	0.495	0.797	0.833	**0.806**	**0.825**
	AlexNet + fine-tuning	0.947	0.952	0.817	0.840	0.615	0.668	0.899	0.914	0.748	0.779
	ResNet50 + fine-tuning	0.950	0.954	**0.876**	**0.905**	0.712	0.756	0.909	0.920	**0.819**	**0.849**
	Image-wise CNN	**0.963**	**0.964**	0.812	0.791	**0.800**	**0.802**	0.914	0.929	0.663	0.705

Two performance metrics were used to benchmark the models: Spearmans rank order correlation coefficient (SRCC), and Pearsons linear correlation coefficient (PLCC). To evaluate the baseline models, we randomly divided each database into two subsets of non-overlapping content (distorted or otherwise), 80% for training and 20% for testing. Of course, all of the LIVE Challenge pictures contain different contents. The SRCC and PLCC were averaged after 10 repetitions of this random process.

The performances of all of the exemplar picture quality prediction models on the LIVE IQA database are shown in Fig. 6. The first five (from left) are no-reference learning-based models, where the last two of these used deep learning. The next seven are CNN-based no-reference quality prediction models, and the last three are CNN-based full-reference models. The reported SRCC and PLCC scores of the listed models were taken from the original papers. Overall, the CNN-based full-reference models, followed by the CNN-based no-reference models achieved higher prediction accuracies relative to conventional learning-based models on the legacy databases.

Table III compares the performance of the various picture quality prediction models on all of the reviewed databases. The last five rows show results for the baseline models. The three best performing no-reference picture quality models in each column are boldfaced. Generally, the existing CNN-based models were able to achieve remarkable prediction accuracies on the legacy databases. However, it is much harder to obtain successful results on the LIVE Challenge database. For example, the model proposed in [27], DIQA, achieved SRCC of 0.687, which is lower than the results attained by a recent successful SVR-based method, FRIQUEE-ALL [21], which achieved an SRCC of 0.72.

However, the baseline models that were pre-trained on the ImageNet databases achieved standout accuracies on the LIVE Challenge database. This is likely because the real-world ImageNet pictures are not synthetically distorted. Instead, like the LIVE Challenge pictures, any distortions occurred as a natural consequence of photography, without intervention by the database creator. This further suggests that the pre-trained CNNs are, to some degree, already "quality-aware," meaning that their learned internal features assist the performance of the task (recognition) by adapting to the presence of authentic distortions.

The baseline models using the first approach achieved very low accuracies on the legacy databases, since they were not exposed to any synthetic distortions during training, and hence the learned features were not very useful to the SVR for quality prediction. Fine-tuning the pre-trained baseline deep models significantly improved performance on the legacy synthetic databases, but not enough to make them competitive, since there was not enough data to train them adequately. The exception was the directly trained shallow CNN baseline model, which achieved competitive performance on the legacy databases, but lower accuracies on the LIVE Challenge database.

A possible explanation for these results is that the pre-trained deep models adapted easily to the authentic distortions in LIVE Challenge as a consequence of having learned image recognition tasks on real-world pictures. Applying them to databases with synthetic distortions, however, like LIVE IQA and TID2013, likely failed to exploit what was learned regarding authentic distortions, hence significant retraining would be needed to deal with the synthetic distortions. This may help explain the excellent generalization power of pre-trained models when applied to other real world image tasks: their ability to handle authentic distortions, by representing them to improve task performance.

VI. Envisioning The Future

The sizes of the training sets used is critical to the success of deep neural network models. Current public-domain databases have insufficient size as compared to widely-used image recognition databases. However, constructing large-scale perceptual quality databases is a much more difficult problem than image recognition databases. Creating databases for picture quality assessment requires time-consuming and expensive subjective studies, which must be conducted under controlled laboratory conditions. Even if the number of reference images is small, the required number of subjective tests quickly becomes excessive. Conducting subjective tests using online crowdsourcing is one possible solution (like the LIVE Challenge database), yet even online tests are (probably) prohibitively difficult to scale up to the necessary size, especially while ensuring the aggregate quality of the collected human data. Another possibility would be if a large social media company were to engage their customers to provide picture quality scores, similar to the Netflix DVD ratings-by-email of a decade ago. Generally, understanding how to successfully create reliable, very large-scale, and authentic picture quality databases remains an open question.

References

[1] A. C. Bovik, "Automatic prediction of perceptual image and video quality," *Proc. IEEE*, vol. 101, no. 9, pp. 2008–2024, 2013.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 1097–1105.

[3] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 372–387, 2016.

[4] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013.

[5] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015.

[6] M. Clark and A. C. Bovik, "Experiments in segmenting texton patterns using localized spatial filters," *Pattern Recognition*, vol. 22, no. 6, pp. 707–717, 1989.

[7] H. Lee, C. Ekanadham, and A. Y. Ng, "Sparse deep belief net model for visual area V2," in *Advances in Neural Information Processing Systems (NIPS)*, 2008, pp. 873–880.

[8] Y. Yuan, Q. Guo, and X. Lu, "Image quality assessment: A sparse learning way," *Neurocomputing*, vol. 159, pp. 227–241, 2015.

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

[10] H. Tang, N. Joshi, and A. Kapoor, "Learning a blind measure of perceptual image quality," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 305–312.

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

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

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

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

[15] Z. Wang, A. C. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[16] D. J. Heeger, "Normalization of cell responses in cat striate cortex," *Vis. Neurosci.*, vol. 9, no. 2, pp. 181–197, 1992.

[17] W. Hou, X. Gao, D. Tao, and X. Li, "Blind image quality assessment via deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. PP, no. 99, pp. 1–1, 2014.

[18] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, 2006.

[19] X. Li, Q. Guo, and X. Lu, "Spatiotemporal statistics for video quality assessment," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3329–3342, 2016.

[20] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind prediction of natural video quality," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1352–1365, 2014.

[21] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *J Vision*, vol. 17, no. 1, 2017.

[22] Y. Li, L.-M. Po, X. Xu, L. Feng, F. Yuan, C.-H. Cheung, and K.-W. Cheung, "No-reference image quality assessment with Shearlet transform and deep neural networks," *Neurocomputing*, vol. 154, pp. 94–109, 2015.

[23] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1733–1740.

[24] Y. Li, L. M. Po, L. Feng, and F. Yuan, "No-reference image quality assessment with deep convolutional neural networks," in *IEEE International Conference on Digital Signal Processing (DSP)*, 2016, pp. 685–689.

[25] S. Bosse, D. Maniry, T. Wiegand, and W. Samek, "A deep neural network for image quality assessment," in *IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 3773–3777.

[26] J. Kim and S. Lee, "Fully deep blind image quality predictor," *IEEE J. Sel. Top. Signal Process.*, vol. 11, no. 1, pp. 206–220, 2017.

[27] ——, "Deep CNN-based blind image quality predictor," *IEEE Trans. Neural Netw. Learn. Syst.*, submitted.

[28] Y. Liang, J. Wang, X. Wan, Y. Gong, and N. Zheng, "Image quality assessment using similar scene as reference," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 3–18.

[29] F. Gao, Y. Wang, P. Li, M. Tan, J. Yu, and Y. Zhu, "DeepSim: Deep similarity for image quality assessment," *Neurocomputing*, 2017.

[30] J. Kim and S. Lee, "Deep Learning of Human Visual Sensitivity in FR-IQA Framework," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[31] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, 2013.

[32] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imaging*, vol. 19, no. 1, pp. 011 006–011006–21, 2010.

[33] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik, "Objective quality assessment of multiply distorted images," in *Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, 2012, pp. 1693–1697.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[35] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, 2011.