CNN-based Cross-dataset No-reference Image Quality Assessment

CNN-based Cross-dataset No-reference Image Quality Assessment

Dan Yang
Tampere University
Finland
dan.yang@tuni.fi

Veli-Tapani Peltoketo
Huawei Technologies Oy (Finland) Co. Ltd
Finland
veli.tapani.peltoketo@huawei.com

Joni-Kristian Kämäräinen
Tampere University
Finland
joni.kamarainen@tuni.fi

Abstract

Recent works on no-reference image quality assessment (NR-IQA) have reported good performance for various datasets. However, they suffer from significant performance drops in cross-dataset evaluations which indicates poor generalization power. We propose a Siamese architecture and training procedures for cross-dataset deep NR-IQA that achieves clearly better performance. Moreover, we show that the architecture can be further boosted by i) pre-training with a large aesthetics dataset and ii) adding low-level quality cues, sharpness, tone and colourfulness, as additional features.

1. Introduction

Image quality assessment (IQA) methods predict visual quality of images. Visual quality refers to the mean opinion score (MOS) averaged over a number of human subjects. Based on availability of a reference image, the IQA methods are classified to full-reference IQA (FR-IQA), reduced-reference IQA (RR-IQA) and no-reference IQA (NR-IQA). NR-IQA methods are further divided to distortion specific [43, 23, 5, 32] and general purpose methods [25, 24, 31, 46, 11, 48, 44, 12, 21, 37]. The most challenging and general setting is *general purpose no-reference image quality assessment*.

In our related work section (Sec. 2) we summarize the recent general purpose NR-IQA methods and datasets. We also report results from preliminary experiments with two recent methods for which the original code is publicly available. We make an important observation: despite of good performance in single-dataset experiments, there is a clear performance drop in the cross-dataset setting.

Motivated by the preliminary findings we propose a Siamese NR-IQA architecture (Fig. 1) and training procedures to improve performace in the cross-dataset setting. We experiment with two popular networks, VGGNet [35] and ILGNet [10], as the core network inside the Siamese structure. VGGNet is a *patch-based* and ILGNet is an *image-based* network. We also propose to add a pair-wise ranking loss [21, 14] to the standard Euclidean loss function. Moreover, we show that pre-training with aesthetics dataset and adding low level quality cues improves the performance. The proposed architecture achieves state-of-the-art cross-dataset accuracy on the challenging CID2013.

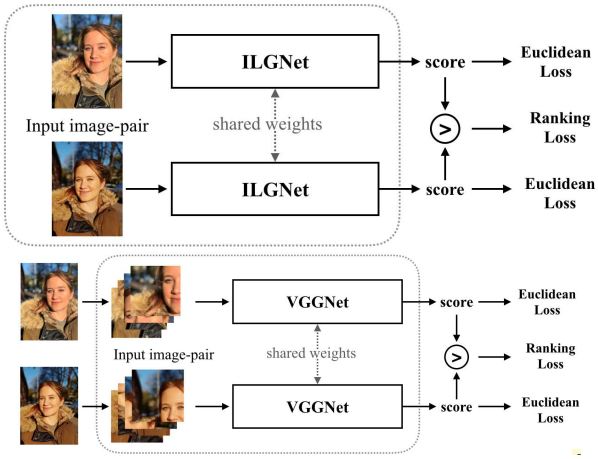

Figure 1. The proposed Siamese architecture for generic (cross-dataset) no-reference image quality assessment. The two different core networks compared in the experiments are the image-based ILGNet (top) and the patch-based VGGNet (bottom).

2. Background

We first provide a short survey to the recent works on no-reference imaqe quality assessment with particular emphasis on deep approaches. Then, we report results for two publicly available methods and multiple datasets in the cross-dataset setting.

2.1. Related work

NSS-based NR-IQA methods define the problem as a classification or a regression problem for features that represent *natural scene statistics* (NSS) or statistics learned from data [45]. NSS-based methods assume that there are statis-

Table 1. Summary of the recent related works on no-reference image quality assessment (SVR: Support Vector Regression; CNN: Convolutional Neural Network; *Cross*: Cross-dataset evaluation reported). Since LIVE IQA was used in all works, we also added their reported Pearson's linear correlation coefficient (PLCC) values.

Method	*Year*	*Feature*	*Regressor*	LIVE IQA [33]	TID2008 [28]	CSIQ [17]	TID2013 [27]	CID2013 [39]	LIVE WIQCD [6]	*Cross*
	→NSS based methods									
DIIVINE [25]	2011	Wavelet coef. stats.	SVR	✓(0.92)						
BLIINDS-II [31]	2012	DCT coef. stats.	SVR	✓(0.93)						
BRISQUE [24]	2012	Spatial norm. image stats.	SVR	✓(0.94)	✓					✓
CORNIA [46]	2012	Norm. image patches and pooling	SVR	✓(0.94)	✓					✓
SOM [48]	2015	Same as CORNIA	SVR	✓(0.96)	✓					✓
	→CNN based methods									
CNN [11]	2014	Norm. image patches	CNN	✓(0.95)	✓					✓
CNN-NR-d [20]	2016	Deep feats. on sub-images	CNN	✓(0.97)	✓					
BIECON [12]	2017	Deep feats.	CNN	✓(0.96)	✓					✓
RankIQA [21]	2017	Deep feats. (special training)	CNN	✓(0.98)			✓			
CNN+suml [2]	2017	Deep feats. on patches	CNN	✓(0.97)	✓					
deepIQA [4]	2018	Deep feats. on patches	CNN	✓(0.98)		✓	✓			✓
deepBIQ [3]	2018	Deep feats.	SVR	✓(0.98)	✓	✓	✓		✓	

tical properties (features) in natural images which are modified by distortions. Moorthy and Bovik [25] proposed a NSS-based framework, Distortion Identification-based Image Verity and INtegrity Evaluation (DIIVINE), where a large set of features are extracted from steerable pyramid wavelet transform coefficients. DIIVINE adopts two stages: at first, a support vector machine (SVM) classifier is employed to identify the distortion type and then a distortion-specific regressor is used to assess image quality. The obvious limitation of this approach is that it does not generalize well to distortion types beyond the ones in the training data or mixed distortions. Saad *et al*. [31] introduced another efficient model, BLind Image Integrity Notator, using DCT statistics (BLIINDS-II), which uses a probabilistic graphical model to directly map a small number of statistical features to a scalar quality without considering distortion types separately. A generalized Gaussian density function models block DCT coefficients of images and quality is predicted based on the parameters of the model. Another well-performing method was proposed by Mittal *et al*. [24] who proposed Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) which adopts an asymmetric generalized Gaussian distribution model of local normalized images in the spatial domain. The modeled image features are differences of spatially neighbored, mean subtracted and contrast normalized samples. A support vector machine regressor (SVR) is used to map from feature space to quality scores.

To avoid the limitations of handcrafted features, several methods that learn features from training data have been proposed [38, 18, 45]. These methods rely on a large number of training features that are not easy to interpret. Ye and Doermann [45] first introduced a codebook-based framework. The codebook is constructed from Gabor features extracted from patches in training images. Image quality is calculated by codeword histograms similar to the visual Bag-of-Words (BoW). The method suffers from a large codebook. CORNIA [46] extends the codebook method with unsupervised feature learning where the codebook is constructed by k-means clustering of normalized image patches. Features of a given image are extracted by max and min pooling of soft-encoded distances between normalized image patches and the codewords and the quality score computed by regression. SOM [46] is refined from CORNIA by adopting semantic obviousness of objects. Object-like regions are first detected and patches of detected regions are fed into CORNIA.

CNN-based NR-IQA methods have recently attracted significant attention since convolutional neural network (CNN) based architectures represent state-of-the-art in many computer vision and image processing applications [30, 7, 22]. CNN based methods have recently reached state-of-the-art performance also in NR-IQA [13].

Kang *et al*. [11] proposed a shallow CNN model of only 5 layers. Input images are subdivided into small patches and each patch is assigned the same subjective quality score during training. Division to multiple patches is used as the data augmentation method in many deep learning frameworks. Kim and Lee [12] designed a deep CNN framework, BIECON, which contains two steps: local image patches are regressed againts full-reference IQA metrics (Step-1), and then pooling of the local CNN scores is used to map the first step scores to subjective scores (Step-2). Bosse *et al*. [4] modified VGGNet [35] to learn a local weight for each image patch to measure the importance of its local quality and weighted average patch aggregation is adopted as the pooling method. Liang *et al*. [20] introduced a novel Dual-path deep Convolutional Neural Network (DCNN) for the both full-reference and no-reference IQA. In the training stage, distorted image and its relevant reference image go through weight-sharing paths so that the same kind of features are extracted. Then features from the both paths are concatenated into a feature vector and fed into regression step to predict quality score. By selecting only one of the two paths the same trained DCNN can be used for no-reference IQA. Inspired by the deep residual networks [9], Bare *et al*. [2] added two sum layers to a 9-layer CNN framework for no-reference IQA in order to achieve bet-

ter stability and performance. Liu *et al*. [21] proposed RankIQA approach to address the problem of limited IQA database size. Using an arbitrary set of images, a ranking dataset is generated by applying different levels of distortions for each image. The ranking dataset is used to train a Siamese network where only ranking information of input images is needed. Finally the core CNN from the trained Siamese network is extracted and is fine-tuned with NR-IQA data.

2.2. Cross-dataset performance

The original codes for two of the well-performing methods in the above section, BRISQUE [24] (NSS-based) and BIECON [12] (CNN-based), are publicly available and in the following we report their performance in the cross-dataset setting. For the experiments we selected the most popular IQA datasets in literature: LIVE IQA [33], CSIQ [17], TID2013 [27], CID2013 [39] and LIVE WIQCD [6]. LIVE IQA, CSIQ and TID2013 are generated datasets where the original undistorted images are available and therefore also full-reference IQA methods can be evaluated. On the other hand, LIVE WIQCD and CID2013 are general purpose no-reference IQA datasets where all quality distortions yield from cameras and their settings. The results are reported for the two standard performance indicators: *Spearman Rank Order Correlation Coefficient* (SROCC) and *Pearson Linear Correlation Coefficient* (PLCC), but for compactness we include only the PLCC values (SROCC provides the same interpretations). More details about the datasets, settings and performance indicators are in Section 4.1.

Results from the cross-dataset experiments are shown in Table 2 for one-vs-one comparison and in Table 3 for leave-one-dataset-out comparison.

Table 2. PLCC performance for BRISQUE and BIECON. The shaded diagonal values represent the single-dataset results and the non-diagonal values are one-on-one cross-dataset results.

	BRISQUE	*Testing*				
		LIVE IQA	CSIQ	TID2013	CID2013	LIVE WIQCD
Training	LIVE IQA	0.937	0.689	0.494	0.603	0.382
	CSIQ	0.889	0.787	0.528	0.588	0.340
	TID2013	0.798	0.692	0.624	0.501	0.381
	CID2013	0.671	0.432	0.400	0.777	0.391
	LIVE WIQCD	0.503	0.382	0.383	0.597	0.594
	BIECON					
Training	LIVE IQA	0.964	0.744	0.506	0.492	0.437
	CSIQ	0.761	0.790	0.482	0.602	0.440
	TID2013	0.859	0.660	0.602	0.616	0.361
	CID2013	0.281	0.514	0.222	0.801	0.467
	LIVE WIQCD	0.129	0.496	0.353	0.666	0.537

As the main experimental finding there are substantial performance drops for both BRISQUE and BIECON when moving from the single-dataset to cross-dataset evaluation. Poor performance on many combinations indicate severe overfitting to training data and poor generalization. The performance is moderate (≈ 0.8) between the datasets that share similar distortions (LIVE IQA $\leftarrow$ CSIQ and TID2013). The two clearly most difficult datasets are CID2013 and LIVE WIQCD that both represent general – "in the wild" – image quality without artificially generated distortions. TID2013 is clearly the most difficult of the three distortion-specific datasets.

Table 3. PLCC performance from the leave-one-dataset-out cross-dataset experiment.

	LIVE IQA	CSIQ	TID2013	CID2013	LIVE WIQCD
BRISQUE	**0.825**	**0.737**	**0.533**	**0.579**	0.404
BIECON	0.735	0.725	0.421	0.570	**0.460**
Best one-vs-one	0.889	0.744	0.528	0.666	0.467

3. Methodology

The main component of the proposed architecture is the "core CNN" which is replicated for training in Siamese style with shared weights (Figure 1). The core network is trained using a combination of the Euclidean loss for the MOS scores and a ranking loss for pair-wise comparison. Pair-wise training helps to exploit limited training data more effectively. While RankIQA [21] uses pair-wise training with synthetic data, the proposed network is trained with the original training images. We experiment two popular networks as the core networks: VGGNet [35] which was originally proposed for image classification and ILGNet [10] which was proposed for image aesthetics prediction. The main difference of these networks is that VGGNet is deeper and requires patch-based training (sub-windows) while ILGNet is trained using full image region.

3.1. Loss function

Using the Euclidean loss for MOS score is straightforward, but requires a large number of training examples which are laborious to obtain for image quality assessment. For this reason, we add a ranking loss term [14]:

$$loss_{rank} = \frac{1}{2N} \sum_{i,j} \max(0, \alpha - \delta(y_i \geqslant y_j)(y_i - y_j)) \quad (1)$$

where

$$\delta(y_i \geqslant y_j) \begin{cases} 1 & \text{if } y_i \geqslant y_j \\ -1 & \text{if } y_i < y_j \end{cases} \quad (2)$$

and α is a specific margin parameter. The ranking loss allows to train the core networks with image pairs where the target is to rank which of the two images has better quality. Pair-wise data augmentation augments the number of training samples from N images to $\binom{N}{2}$ image pairs. In the experiments, only image pairs capturing the same scene were mixed.

3.2. ILGNet

The first "core network" experimented in our Deep NR-IQA model is the ILGNet network that recently achieved state-of-the-art performance in image aesthetics classification [10]. ILGNet is built by stacking the Inception Modules introduced by Szegedy et al. [36]. The main change in our case is that that classification layer is replaced with a regression layer to output the MOS score. One of the reasons to select ILGNet for our work is that we also pre-train our model with the AVA [26] large scale image aesthetic dataset. Moreover, we use ILGNet inside the Siamese architecture which is trained with the two loss functions. For pair-wise training we use two images from the same scene but captured using two different cameras (Figure 1).

3.3. VGGNet

The second core network tested in the proposed model is VGGNet [35] that was originally proposed for image classification, but has been used as a pre-trained network for other vision tasks such as color constancy [29] and semantic segmentation [34]. VGGNet is a deeper and more optimized version of the AlexNet [16]. To train the architecture with the VGGNet core network the training procedure of BIECON [12] was adopted. Each training image was randomly divided into patches of the same size and each patch was assigned the same quality score as the training image. For test images the patch scores are averaged.

3.4. Quality Attributes

In image aesthetics assessment [1, 42] the target is to estimate multiple *image aesthetics attributes*. Inspired by these works and the finding that certain low level cues performed well in the CID2013 cross-dataset experiments in [39] a number of low-level quality metrics were selected as image quality attributes:

- *Sharpness* - Sharpness has been found as an important cue for image quality and there are many proposed sharpness measures: S3 [40], FISH [41] and RISE [19]. For our experiments the spectral and spatial sharpness S3 was selected.

- *Tone* - Tone is another important cue for image quality and denotes the global lightness difference over entire image. The works on perceptual lightness indicate that the extreme values are more important than the mean luminance [15]. Therefore, we compute the 95th percentile and 5th percentile according to [1]. The top and bottom quantiles provide the two extrema, but are robust to a small number of isolated pixels with noisy values.

- *Colourfulness* - The third important cue that is also present in user studies is colourfulness [8] measured by the standard deviation and mean of the opponent color channels yellow-blue and red-green.

Computation of the above quality attributes does not require a reference image and therefore they could be used as additional features for CNN learning. However, we use them as extra outputs and train the core network as a multivariate regressor (Figure 2). The main benefit of this approach is that the low level cues act as regularization terms that enforce the core network to learn features that contribute to the both low level quality attributes and high level overall quality. Similar connections between visual tasks have been resently reported in Zamir et al. [47].

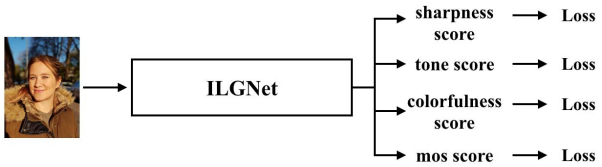

Figure 2. Example of the three quality attributes added to the core network (ILGNet) training stage. In addition to the MOS score the network also learns features that are relevant for the low level image quality cues. Euclidean loss is used for all outputs and in the ablation study the optimal weights for the MOS vs. attributes are experimentally investigated.

4. Experiments

In the experiments, the main focus is on *"quality in the wild"* in which quality distortions are not artificially generated but yield from the real capturing process. Results are reported only for the *cross-dataset setting* where training and test images are from different datasets. The results indicate generalization power of different methods since the contents and camera hardware are different. For single dataset numbers see Section 2.2.

4.1. Datasets and Settings

Datasets The two most recent image quality assessment benchmarks were selected (Figure 3):

- CID2013 [39]: CID2013 Camera Image Database consists of 480 images captured by 79 imaging devices of 8 scenes that represent typical contents taken by consumers. Quality difference yield from different camera-specific factors such as sensor types, optics and image signal processing pipelines. Subjective evaluation was conducted by 188 observers and Mean Opinion Scores (MOS) for each image are provided.

- LIVE WIQCD [6]: LIVE In the Wild Image Quality Challenge Database contains 1,162 authentic images captured with different and unknown mobile phone cameras. Those images are evaluated by over 350,000 crow-sourced observers and MOS values are provided.

We also collected our own dataset:

- HUAWEI: HUAWEI dataset contains 884 images taken by 4 high-end smartphone cameras and 32 content types (portrait, landscape, and different macro setups). MOS scores are generated from pair-wise comparisons where human subjects have ranked each image pair (winner gets 1 and loser gets 0). MOS values are computed from these preference scores and are consistent at least over each content type. HUAWEI dataset is similar to CID2013, but the cameras represent high-end smart phones and it contains more scenes.

In order to experimentally validate whether image aesthetic contributes to image quality, we selected the following large-scale aesthetics dataset to be used in cross-domain training (Figure 3):

- AVA [26]: The Aesthetic Visual Analysis (AVA) dataset contains $\sim$ 250k images collected from the Web and aesthetics of each image is voted by 78-549 crowd-sourced users with the scale from 1 to 10. The average scores are provided as the ground-truth aesthetic score for each image.

Performance Metrics Two standard performance metrics are used to report the results from the experiments: *Spearman Rank Order Correlation Coefficient* (SROCC) and *Pearson Linear Correlation Coefficient* (PLCC).

PLCC - Linear Correlation - is the standard measure for regression where $+1$ denotes perfect positive correlation and -1 perfect negative correlation. Values near zero denote poor correlation. In image quality assessment PLCC is used to measure the linear correlation between the true subjective and method predicted scores:

$$PLCC = \frac{\sum_{i=1}^{n}(s_i - \overline{s})(q_i - \overline{q})}{\sqrt{\sum_{i=1}^{n}(s_i - \overline{s})^2}\sqrt{\sum_{i=1}^{n}(q_i - \overline{q})^2}} \quad (3)$$

where s_i is the ground truth subjective score (MOS/DMOS) and q_i is the predicted score for the i-th image. $\overline{s}$ and $\overline{q}$ are the mean values computed over the ground truth and predicted scores, respectively.

The PLCC measure is suitable for scores with monotonic linear relationship, i.e. for the cases where the linear regression also performs well. However, this is not always the case in image quality assessment and therefore SROCC performance metric is used in parallel with PLCC. SROCC is also suitable for the cases of monotonic but non-linear relationship since it uses rank-order statistics. SROCC is defined as:

$$SROCC = \frac{1 - 6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \quad (4)$$

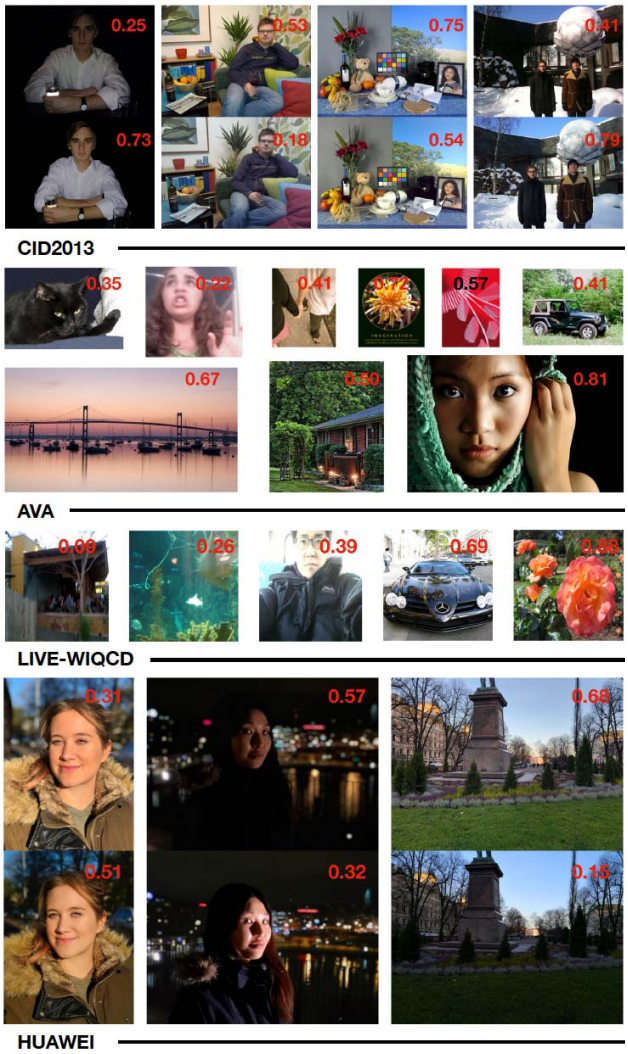

Figure 3. Examples from the three image quality and one aesthetics datasets used in the experiments.

where d_i is the rank-order difference between the i-th image indeces in the sorted lists of the subjective ground truth and predicted scores. SROCC values are interpreted similar to PLCC values.

4.2. One-on-one cross-dataset results

The purpose of the first experiment is to study the effect of the core network (VGGNet vs. ILGNet) and training data (AVA vs. LIVE WIQCD vs. CID2013 vs. HUAWEI) to the model performance. As a side study also the effect of pair-wise vs. direct training was tested. Note that the main difference between the two core networks is that the VGGNet based NR-IQA model was trained patch-wise (details in Section 3.3) and the ILGNet based model image-wise (Section 3.2). The results are shown in Table 4.

The main findings from the first experiment are:

Table 4. Cross-dataset performance for the CID2013 and LIVE WIQCD test datasets using various (single) datasets for training and using two different CNNs, VGGNet and ILGNet, as the core networks in the proposed deep NR-IQA model. With the HUAWEI dataset training was conducted both direct and pair-wise training.

Core CNN w/ Tr. Data		CID2013		LIVE WIQCD	
	pair-wise	SRCC	PLCC	SRCC	PLCC
VGGNet w/ AVA		0.483	0.558	0.193	0.186
VGGNet w/ LIVE WIQCD		**0.642**	**0.684**	-	-
VGGNet w/ CID2013		-	-	**0.366**	**0.360**
VGGNet w/ HUAWEI		0.392	0.486	0.095	0.075
VGGNet w/ HUAWEI	✓	0.462	0.515	0.225	0.206
ILGNet w/ AVA		**0.575**	**0.647**	**0.363**	**0.348**
ILGNet w/ LIVE WIQCD		0.412	0.477	-	-
ILGNet w/ CID2013		-	-	0.329	0.315
ILGNet w/ HUAWEI		0.506	0.545	0.348	0.335
ILGNet w/ HUAWEI	✓	0.507	0.604	0.333	0.320

- Image aesthetics is closely connected to image quality as the models trained with only the AVA aesthetics dataset and using the aesthetic scores directly as quality scores (MOS) performed moderately well with the both VGGNet and ILGNet based models. Rather strikingly, ILGNet model achieved better performance with AVA training data (PLCC 0.647), than with any of the image quality datasets.
- The selected core network has significant effect to the performance and the best results for the both tested datasets strongly depends on the combination of the core network and training data. This clearly indicates that either the datasets or the core networks or both have complementary properties. The best combination for CID2013 is VGGNet w/ LIVE WIQCD and for LIVE WIQCD VGGNet w/ CID2013.
- With VGGNet pair-wise training has clear positive effect and with ILGNet there is no strong effect.
- The best results for CID2013 are moderate (PLCC 0.684) and for LIVE WIQCD poor (0.360) indicating that the LIVE WIQCD dataset is very challenging.

4.3. Many-vs-one cross-dataset results

Table 5. CID2013 performance using multiple training datasets (pre-training and fine-tuning) and an attribute layer.

Core CNN w/ pre + fine			CID2013	
	pair-wise	attributes	SRCC	PLCC
VGGNet w/ LIVE WIQCD			**0.642**	**0.684**
VGGNet w/ LIVE WIQCD		✓	0.519	0.610
VGGNet w/ AVA + LIVE WIQCD			0.561	0.632
VGGNet w/ HUAWEI + LIVE WIQCD	✓		0.501	0.584
VGGNet w/ LIVE WIQCD + HUAWEI	✓		0.509	0.561
ILGNet w/ AVA			0.575	0.647
ILGNet w/ AVA + HUAWEI	✓		0.606	0.672
ILGNet w/ AVA + HUAWEI	✓	✓	**0.666**	**0.710**

In this experiment, the complementary properties of datasets and networks are further investigated, and multiple training datasets are used. The experiments are conducted only for CID2013 since the LIVE WIQCD results are always far from sufficient (PLCC ≥ 0.8 is needed for practical applications). Moreover, we investigate the effect of adding semantic attributes available during training (Section 3.4). For all possible combinations, we carefully tune the training parameters via cross-validation and the best performing combinations are presented in Table 5.

The starting point for these experiments is to take the best combination from Table 4 and fine-tune it further with additional data and attributes. The main findings from the second experiment are:

- The model using VGGNet core network cannot expoit additional value of more data or attributes or pair-wise training but the performance actually degrades in all cases as compared to the initial setup of training only with the LIVE WIQCD dataset.
- Interestingly, different to VGGNet the ILGNet core network benefits from all additional data - fine-tuning with HUAWEI data, pair-wise training and adding the attributes to the network. The best PLCC performance is 0.710 which outperforms VGGNet and is the best reported result for the CID2013 dataset in the cross-dataset setting to the authors' best knowledge.

4.4. Ablation study

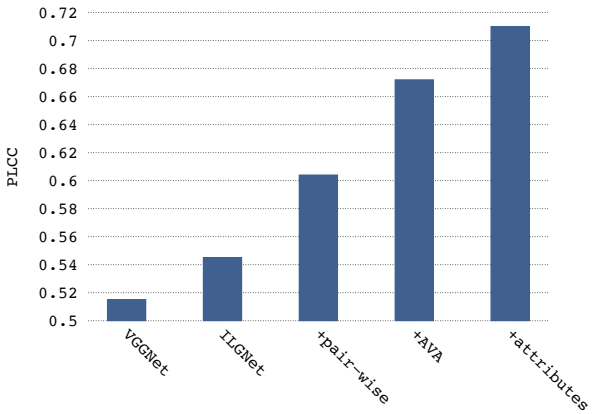

Figure 4. Ablation study of the effect of each component in the proposed model.

The best result in the previous experiment was achieved with the following settings: 1) ILGNet, 2) combining the Euclidean loss and the pair-wise ranking loss, 3) pre-training with AVA, 4) using additional loss terms for the low-level quality attributes. The effect of the each design choice is shown in the ablation study graph in Figure 4.

We want to further study the optimal balance between the MOS error and the attribute error term. The results are shown in Table 6. Weighting the MOS error twice more than the attribute estimation provides the best performance.

Table 6. Ablation study of the balance between the Euclidean loss (λ_{MOS}) and attribute estimation loss (λ_{attr}). The best results were achieved at $2/1$ where the MOS error is given twice more weight.

weights settings $\lambda_{MOS}/\lambda_{attr}$	CID2013 SRCC	CID2013 PLCC
1/1	0.616	0.678
2/1	0.666	0.710
3/1	0.632	0.686

4.5. More detailed analysis on CID2013

The original authors of CID2013 provide labels for each image that represent a "sub-clusters" of similar content. According to the original paper, the sub-cluster accuracies were also computed and are shown in Table 7 for the best combinations in the previous experiment (only the PLCC numbers are included for compactness). Both VGGNet and ILGNet based models work well for certain type of images (Cluster-1: portraits in dim light and Cluster-3: small groups of people in dim light) and fail for another type (Cluster-5: small groups in sunny/cloudy outdoor and Cluster-7: zoomed groups in outdoors) (see Figure 5 for examples of these clusters).

5. Conclusion

We investigated general purpose no-reference image quality assessment in the challenging cross-dataset evaluation setting. The results for the two most popular datasets, LIVE IQA and CSIQ, are already saturing as their single dataset performance are very good and even the cross-dataset results are moderately good (assuming suitable training data): best PLCC 0.889 for LIVE IQA and 0.744 for CSIQ. On the other hand, TID2013, CID2013 and LIVE WIQCD are still challenging and the tested methods' performance collapsed in the cross-dataset setting.

To improve generalization power of deep NR-IQA we proposed a deep architecture that exploits various generalization tricks proposed in literature. In the final architecture, image-wise trained ILGNet won patch-wise trained VGGNet. Moreover, ILGNet benefits from pre-training with large image aesthetics data (AVA), pair-wise training with ranking loss and low-level quality attributes (sharpness, tone and colourfulness). For the CID2013 dataset the network achieved state-of-the-art cross-dataset performance (PLCC 0.710). Our code will be made publicly available.

References

[1] T.O. Aydın, A. Smolic, and M. Gross. Automated aesthetic analysis of photographic images. *IEEE Trans. on Visualization and Computer Graphics*, 21(1), 2015. 4

[2] B. Bare, Ke Li, and Bo Yan. An accurate deep convolutional neural networks model for no-reference image quality assessment. In *IEEE Int. Conf. on Multimedia and Expo (ICME)*, 2017. 2

[3] S. Bianco, L. Celona, P. Napoletano, and R. Schettini. On the use of deep learning for blind image quality assessment. *Signal, Image and Video Processing*, 12, 2018. 2

[4] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Trans. on Image Processing*, 27(1), 2018. 2

[5] R. Ferzli and L.J. Karam. A no-reference objective image sharpness metric based on the notion of just noticeable blur (jnb). *IEEE Trans. on Image Processing*, 18(4), 2009. 1

[6] D. Ghadiyaram and A.C. Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Trans. on Image Processing*, 25(1), 2016. 2, 3, 4

[7] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015. 2

[8] D. Hasler and S.E. Suesstrunk. Measuring colorfulness in natural images. In *Human vision and electronic imaging VIII*, volume 5007, pages 87–96, 2003. 4

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016. 2

[10] X. Jin, J. Chi, S. Peng, Y. Tian, C. Ye, and X. Li. Deep image aesthetics classification using inception modules and fine-tuning connected layer. In *8th Int. Conf. on Wireless Communications & Signal Processing (WCSP)*, 2016. 1, 3, 4

[11] Le Kang, P. Ye, Yi Li, and D. Doermann. Convolutional neural networks for no-reference image quality assessment. In *CVPR*, 2014. 1, 2

[12] Jongyoo Kim and Sanghoon Lee. Fully deep blind image quality predictor. *IEEE Journal of Selected Topics in Signal Processing*, 11(1):206–220, 2017. 1, 2, 3, 4

[13] Jongyoo Kim, Hui Zeng, Deepti Ghadiyaram, Sanghoon Lee, Lei Zhang, and Alan C Bovik. Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment. *IEEE Signal Processing Magazine*, 34(6):130–141, 2017. 2

[14] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *ECCV*, 2016. 1, 3

[15] Grzegorz Krawczyk, Karol Myszkowski, and Hans-Peter Seidel. Lightness perception in tone reproduction for high dynamic range images. In *Computer Graphics Forum*, volume 24, pages 635–645. Wiley Online Library, 2005. 4

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 4

[17] Eric C Larson and Damon M Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):011006–011006, 2010. 2, 3

Table 7. More detailed results with the CID2013 dataset (PLCC). Clusters represent the different content types as defined in the original paper. Note that the *All* result is computed as a weighted average over the per-cluster means.

# of imgs	cluster-1 79	cluster-2 79	cluster-3 79	cluster-4 53	cluster-5 65	cluster-6 79	cluster-7 14	cluster-8 26	All
FISH-bb [39]	0.69	**0.55**	0.31	0.12	**0.48**	**0.68**	**0.79**	0.30	0.49
S3 [39]	0.65	0.43	0.27	0.05	0.44	0.61	0.73	0.19	0.48
FISH [39]	0.67	0.45	0.25	0.17	0.41	0.61	**0.79**	0.13	0.46
VGGNet w/ LIVE WIQCD	**0.773**	0.234	**0.781**	**0.324**	0.108	0.405	-0.082	0.192	0.683
ILGNet w/ AVA + n/a	0.746	0.316	0.765	0.252	0.150	0.401	-0.241	**0.396**	0.647
ILGNet w/ AVA + HUAWEI	0.760	0.332	0.751	0.257	0.133	0.384	-0.272	0.278	0.672
ILGNet w/ AVA + HUAWEI (+attributes)	0.771	0.353	0.757	0.299	0.111	0.358	-0.483	0.310	**0.710**

Figure 5. Successful examples are from cluster-1: close-up in dark lighting conditions and cluster-3: small group in dim lighting conditions; Failed examples are from cluster-5: small group in cloudy/sunny conditions and cluster-7: same as 5 but with 3x zoom.

[18] Chaofeng Li, Alan Conrad Bovik, and Xiaojun Wu. Blind image quality assessment using a general regression neural network. *IEEE Transactions on Neural Networks*, 22(5):793–799, 2011. 2

[19] L. Li, W. Xia, W. Lin, Y. Fang, and S. Wang. No-reference and robust image sharpness evaluation based on multiscale spatial and spectral features. *IEEE Trans. on Multimedia*, 19(5), 2017. 4

[20] Yudong Liang, Jinjun Wang, Xingyu Wan, Yihong Gong, and Nanning Zheng. Image quality assessment using similar scene as reference. In *European Conference on Computer Vision (ECCV)*, pages 3–18. Springer, 2016. 2

[21] X. Liu, J. van de Weijer, and A.D. Bagdanov. RankIQA: Learning from rankings for no-reference image quality assessment. In *ICCV*, 2017. 1, 2, 3

[22] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015. 2

[23] Pina Marziliano, Frederic Dufaux, Stefan Winkler, and Touradj Ebrahimi. Perceptual blur and ringing metrics: application to jpeg2000. *Signal processing: Image communication*, 19(2):163–172, 2004. 1

[24] A. Mittal, A. Krishna Moorthy, and A.C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Trans. on Image Processing*, 21(12), 2012. 1, 2, 3

[25] Anush Krishna Moorthy and Alan Conrad Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE transactions on Image Processing*, 20(12):3350–3364, 2011. 1, 2

[26] Naila Murray, Luca Marchesotti, and Florent Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *CVPR*, 2012. 4, 5

[27] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. Image database tid2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30:57–77, 2015. 2, 3

[28] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti. TID2008 - a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, 10, 2009. 2

[29] Y. Qian, K. Chen, J. Nikkanen, J.-K. Kämäräinen, and J. Matas. Recurrent color constancy. In *Int. Conf. on Computer Vision (ICCV2017)*, 2017. 4

[30] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 2

[31] Michele A Saad, Alan C Bovik, and Christophe Charrier. Blind image quality assessment: A natural scene statistics approach in the dct domain. *IEEE transactions on Image Processing*, 21(8):3339–3352, 2012. 1, 2

[32] H.R Sheikh, A.C Bovik, and L. Cormack. No-reference quality assessment using natural scene statistics: JPEG2000. *IEEE Trans. on Image Processing*, 14(11), 2005. 1

[33] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. A statistical evaluation of recent full reference image quality

assessment algorithms. *IEEE Transactions on image processing*, 15(11):3440–3451, 2006. 2, 3

[34] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4), 2017. 4

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 2, 3, 4

[36] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 4

[37] H. Talebi and P. Milanfar. NIMA: Neural image assessment. *IEEE Trans. on Image Processing*, 27(8), 2018. 1

[38] Huixuan Tang, Neel Joshi, and Ashish Kapoor. Learning a blind measure of perceptual image quality. In *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 305–312. IEEE, 2011. 2

[39] T. Virtanen, M. Nuutinen, M. Vaahteranoksa, P. Oittinen, and J. Häkkinen. CID2013: A database for evaluating no-reference image quality assessment algorithms. *IEEE Trans. on Image Processing*, 24(1), 2015. 2, 3, 4, 8

[40] C. T Vu, T.D. Phan, and D.M. Chandler. S_3: A spectral and spatial measure of local perceived sharpness in natural images. *IEEE Trans. on Image Processing*, 21(3), 2012. 4

[41] P.V. Vu and D.M. Chandler. A fast wavelet-based algorithm for global and local image sharpness estimation. *IEEE Signal Processing Letters*, 19(7), 2012. 4

[42] Z. Wang, D. Liu, S. Chang, F. Dolcos, D. Beck, and T. Huang. Image aesthetics assessment using deep chatterjee's machine. In *Int. Joint Conf. on Neural Networks (IJCNN)*, 2017. 4

[43] Zhou Wang, Hamid R Sheikh, and Alan C Bovik. No-reference perceptual quality assessment of JPEG compressed images. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 1, pages I–I. IEEE, 2002. 1

[44] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann. Blind image quality assessment based on high order statistics aggregation. *IEEE Trans. on Image Processing*, 25(9), 2016. 1

[45] Peng Ye and David Doermann. No-reference image quality assessment using visual codebooks. *IEEE Transactions on Image Processing*, 21(7):3129–3138, 2012. 1, 2

[46] P. Ye, J. Kumar, Le Kang, and D. Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *CVPR*, 2012. 1, 2

[47] Amir R. Zamir, A. Sax, W. Shen, L. Guibas, J. Malik, and S. Saverese. Taskonomy: Disentangling task transfer learning. In *CVPR*, 2018. 4

[48] Peng Zhang, Wengang Zhou, Lei Wu, and Houqiang Li. SOM: Semantic obviousness metric for image quality assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2394–2402, 2015. 1, 2