Signal Processing: Image Communication 114 (2023) 116942

Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

Statistical hypothesis testing as a novel perspective of pooling for image quality assessment

Rui Zhu [a,b,*], Fei Zhou [c], Wenming Yang [d], Jing-Hao Xue [e]

[a] Faculty of Actuarial Science and Insurance, Bayes Business School, City, University of London, London EC1Y 8TZ, UK
[b] School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury CT2 7FS, UK
[c] College of Information Engineering, Shenzhen University, Shenzhen 518060, China
[d] Department of Electronic Engineering, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China
[e] Department of Statistical Science, University College London, London WC1E 6BT, UK

ARTICLE INFO

Keywords:
Image quality assessment
Pooling strategy
Hypothesis testing

ABSTRACT

Image quality assessment is usually achieved by pooling local quality scores. However, commonly used pooling strategies, based on simple sample statistics, are not always sensitive to distortions. In this short communication, we propose a novel perspective of pooling: reliable pooling through statistical hypothesis testing, which enables effective detection of subtle changes of population parameters when the underlying distribution of local quality scores is affected by distortions. To illustrate the significance of this novel perspective, we design a new pooling strategy utilising simple one-sided one-sample t-test. The experiments on benchmark databases show the reliability of hypothesis testing-based pooling, compared with state-of-the-art pooling strategies.

1. Introduction

Image quality assessment (IQA) has attracted considerable research interests recently [1]. IQA metrics are often obtained via two steps: (1) calculate local quality scores, and (2) pool local scores together to get an overall score for the image. Most studies focused on the first step, designing better local quality scores such as noise quality measure (NQM) [2], structural similarity index (SSIM) [3], multiscale structural similarity (MS-SSIM) [4], feature similarity index (FSIM) [5], gradient similarity (GSM) [6], and deep learning-based local scores [7–9].

However, fewer studies have paid attention to the pooling strategy. Due to its simplicity, mean pooling, which uses the mean of local scores as the overall score, is commonly used [3,7]. An issue with the mean pooling is that it treats all local scores equivalently, whereas some local scores should contribute more, because they are more associated with human visual fixation or visual region of interest. Hence several weighted-mean pooling strategies are then proposed to assign different weights to local scores based on various criteria, including information content weighting [10], visual saliency weighting [11], visual importance weighting [12], and the weights learned by neural networks [13]. However, the computation of weights is often costly, especially for deep models, which are further affected by their training data. In addition, each weighting scheme is usually designed for some specific IQA metrics and not suitable for pooling local scores of other IQA metrics. Besides mean, standard deviation has also been used as a pooling strategy, by assuming that the overall quality is associated with the variation of local quality scores [14]. It is worth noting that although they share similar names, the pooling strategy discussed in this paper is different from the pooling layer involved in deep models which aims to downsample the feature maps.

Sample mean and sample standard deviation are summary statistics that can only capture certain simple properties of the probability distribution of the local scores: mean measures the location of the distribution while standard deviation measures the spread of the distribution. Moreover, they are not always sensitive to distortions, in which case the inference based on them are not reliable.

Different from summary statistics, statistical hypothesis testing can make more reliable statistical inference from a set of observations [15,16]. For example, in statistical inference, a proper hypothesis test is usually hired to obtain a more reliable conclusion on which population mean is larger, instead of simply comparing the sample means directly.

Therefore in this short communication, we propose a novel perspective of pooling: statistical hypothesis testing-based pooling (HT Pooling). By using hypothesis tests, we can obtain more reliable ranks of the quality of the images, and thus the overall quality scores obtained from our HT pooling can achieve better prediction monotonicity with the mean opinion scores (MOS). Besides the reliable predictions, the HT pooling can have low computational cost and high generalisability to work with various IQA metrics.

* Corresponding author at: Faculty of Actuarial Science and Insurance, Bayes Business School, City, University of London, London EC1Y 8TZ, UK.
E-mail addresses: rui.zhu@city.ac.uk (R. Zhu), flying.zhou@163.com (F. Zhou), yangelwm@163.com (W. Yang), jinghao.xue@ucl.ac.uk (J.-H. Xue).

https://doi.org/10.1016/j.image.2023.116942
Received 21 April 2022; Received in revised form 7 January 2023; Accepted 31 January 2023
Available online 17 February 2023

To illustrate the feasibility and effectiveness of this new perspective, we develop the HT pooling counterpart of mean pooling by utilising one-sided one-sample t-test on the local quality scores of each image: we test the null hypothesis that the mean of local quality scores of an image is equal to a constant c, against the alternative hypothesis that the mean is larger than c. Then the one-sample t-test statistic score can be directly used as the overall quality score: the higher the t-test statistic score, the higher the quality of the image. Further, to achieve better prediction with MOS, we apply a log-transformation on the t-test statistic score and use the transformed score as final overall quality score. This simple HT pooling strategy can achieve superior overall results on benchmark databases, demonstrating its effectiveness. A further remark of advantage: this simple HT pooling strategy does not require information from reference images, and thus can be adopted for both full-reference and no-reference IQA [17–19].

The novelties and contributions in this short communication are two-fold.

1. First, to the best of our knowledge, we are the first to introduce statistical hypothesis testing as an IQA pooling strategy. The HT pooling is more reliable than prior pooling strategies based on simple summary statistics. In addition, different from most weighted average pooling strategies, the HT pooling is not restricted by the IQA metrics used to calculate local quality scores: it is applicable and generally reliable for various IQA metrics.
2. Second, the HT pooling opens a novel perspective for designing various new pooling strategies. This short communication shows one example by designing a new pooling strategy based on the t-test, which infers the mean of the underlying distribution. Various hypothesis tests in statistical inference can be further exploited to design other HT pooling strategies, such as the F-test to infer the variance.

2. Related work

The pooling strategy is one vital element for designing IQA metrics, which aims to decide the overall image quality score from a set of local image quality scores.

The mean pooling [3] is the most widely used pooling strategy, which provides the average of local scores as the overall quality score. However, researchers note that the mean pooling is not always a good strategy because different regions should receive different degrees of attention in pooling; for example, the region providing more information of human visual fixation should be weighted more. Thus a simple average that uses the same weight for all local scores can be improved. Several weighting schemes are developed via exploring visual information in different ways. Moorthy and Bovik [12] explore the concept of visual importance based on a visual fixation predictor and propose the visual importance weighting. Wang and Li [10] propose an information content weighting scheme which weights the local scores by measuring their local information content. Zhang et al. [11] utilise the visual saliency in the pooling strategy that measures the importance of the region to attract people's attention. Deep neural networks are also powerful to learn weights. Bosse et al. propose an end-to-end network with ten convolutional layers to learn local scores and five pooling layers to obtain local weights for weighted-mean pooling [7]. Jiang et al. develop a new metric for screen content images, following a similar approach to Bosse's method while considering the special characteristics of screen content images [9].

However, we note that finding appropriate weights can raise the computational cost substantially, especially for deep models. In addition, most weighting schemes are designed only for specific IQA metrics which cannot serve as a general pooling strategy for various IQA metrics. Moreover, the mean pooling or weighted-mean pooling only makes use of the mean statistic, which can only reflect the location information of the set of local scores.

The standard deviation pooling [14] uses the standard deviation of the local scores as the overall image quality score. This is based on the observation that the variation of local quality scores can reflect the overall quality degradation. However, the standard deviation pooling can only provide information about the spread information of the local scores.

3. Methodology

Suppose the set of local quality scores of an image is $\mathbf{x} = [x_1, x_2, \dots, x_N]^T \in \mathbb{R}^{N\times 1}$, where N is the total number of pixels. A pooling strategy aims to obtain an overall quality score for the image given $\mathbf{x}$. From a statistical point of view, we treat $\mathbf{x}$ as a set of observations from an underlying population probability distribution that describes the quality of the image. In this way, the aim of a pooling strategy can be regarded as to find a value to describe the underlying distribution based on the observations.

In Section 3.1, we first give an example in Fig. 1 when the mean pooling is not sensitive to distortions and the overall score is not reliable. This motivates us to propose a novel perspective to design new pooling strategies based on statistical hypothesis testing, which can provide more reliable objective overall scores than prior summary statistics-based pooling strategies. Then in Section 3.2, we develop an example of using one-sided one-sample t-test as the pooling strategy.

3.1. Motivation for the HT pooling

The mean pooling and standard deviation pooling strategies are not always sensitive to distortions. For example, an image that is severely visually-distorted by few pixels can have a high overall mean score: the few small local quality scores are compensated by a large number of high local quality scores. However, a good pooling strategy is expected to detect those small local quality scores and give a low overall quality score to this image. The mean pooling gives an incorrect high overall quality score because it is not sensitive to the few small local quality scores. In other words, the mean pooling cannot distinguish between the population mean of the distribution which describes the above low quality image and that of the distribution which describes a high quality image, because their sample means are similar.

Fig. 1 shows an example of the above situation in the TID2013 database. Fig. 1(b) is severely distorted by two local blocks and has a low MOS, while Fig. 1(c) is slightly distorted by the non eccentricity pattern noise and has a high MOS. The histograms of the local SSIM scores of the two images are shown in Fig. 1(d) and Fig. 1(e), respectively. It is obvious that most of the pixels have high local scores close to 1 in both images. However, different from Fig. 1(e), Fig. 1(d) has few small local scores around zero, due to the two local blocks. The mean pooling cannot detect those small local scores in Fig. 1(d), and it gives almost the same overall scores for the two images with distinct visual qualities. That is, the overall scores from the mean pooling are not consistent with MOS and thus not reliable.

This problem can be resolved by weighted mean pooling strategies. However, the weights have to be carefully engineered or learned from a costly algorithm to emphasise the importance of those distorted pixels. In contrast, here we show a much simpler solution from a paired t-test without heavy calculations, which can effectively detect the subtle difference between the means of the underlying distributions of the local SSIM scores of Fig. 1(b) and Fig. 1(c). The p-value of the paired t-test on the two sets of local SSIM scores is 3.95×10^{-11}, which indicates that we have very strong confidence to conclude that the two population means are significantly different. This is because the t-test statistic jointly considers both sample mean and sample standard deviation. Although the sample means are similar, the t-test can recognise the large difference between the standard deviations of the local scores in Fig. 1(d) and Fig. 1(e) and conclude that the two sets of local scores are significantly different.

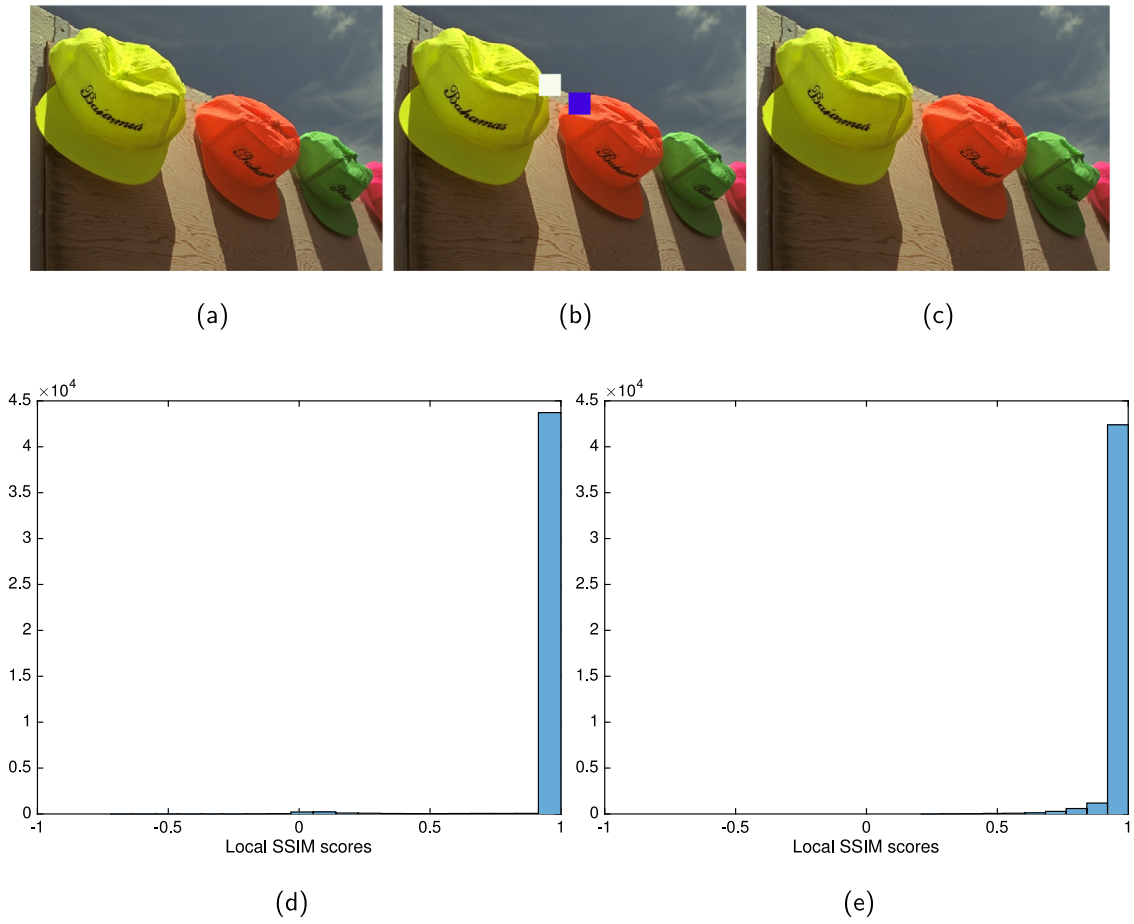


Fig. 1. An example in the TID2013 database when the mean pooling is not sensitive to distortions: (a) the original image; (b) the local block-wise distorted image with SSIM = 0.9825 and MOS = 4.80; (c) the non eccentricity pattern noise distorted image with SSIM = 0.9866 and MOS = 6; (d) the histogram of the local SSIM scores of the image in (b); and (e) the histogram of the local SSIM scores of the image in (c). Here the overall SSIM score is calculated by the mean pooling.

Similar examples can also be found when adopting standard deviation pooling. For instance, when the standard deviations of the local quality scores of the same image affected by the local block-wise distortion and spatially correlated noise are similar, the one with spatially correlated noise usually has a lower MOS. This is because there is a large amount of pixels affected by the spatially correlated noise, which results in a histogram with high frequency densities for lower scores, and such strong global noise can affect HVS more than a few mild local block distortions.

These observations inspire us to propose a new perspective of pooling based on statistical hypothesis testing (termed HT pooling), which can be more sensitive to distortions.

Under this perspective, the existing summary statistics-based pooling strategies can be replaced by new pooling strategies simply based on their corresponding hypothesis tests, to obtain more reliable results. Various hypothesis tests to compare different population parameters can be found in statistics literature, such as the t-test to compare means and the F-test to compare the variances [15], to name a few.

3.2. The HT pooling by using t-test

In this section, we develop an example of the HT pooling: we design a pooling strategy based on a one-sided one-sample t-test. Specifically, we apply the one-sided one-sample t-test on the local quality scores and adopt a transformation of the t statistic as the final image quality metric. The null hypothesis H_0 is that the population mean $\mu = c$, and the alternative hypothesis H_1 is that $\mu > c$, where c is a constant. The test statistic t is

$$t = \frac{\bar{x} - c}{s/\sqrt{N}}, \tag{1}$$

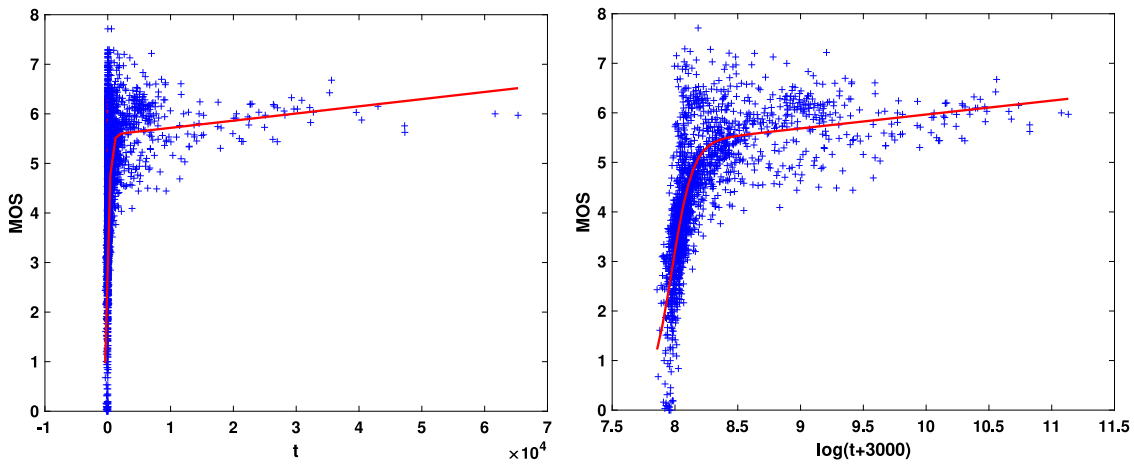
where $\bar{x}$ and s are the sample mean and the sample standard deviation of the local scores, respectively, and c is a pre-determined constant. If t is larger than the critical value, we reject the null hypothesis and conclude that the population mean is larger than c. Although the t-test assumes that the samples are from Gaussian populations, we can still use it for non-Gaussian samples with large sample size: the power of the test is still strong in this case [15]. As the number of pixels in an image is usually large, we can adopt the t-test without losing much testing power.

The value of the t statistic in (1) can be considered as the scaled difference between the sample mean and c. If the sample mean is much larger than c, we obtain very large t values. Alternatively, if the sample mean is much smaller than c, we obtain very small negative t values. Therefore, we can use the t values as the overall quality scores: the larger the t value, the higher the quality of the image.

It is clear in (1) that the t statistic contains information from sample mean $\bar{x}$ and sample standard deviation s. Therefore, the pooling strategy based on the t statistic can exploit the advantages of both the mean pooling and the standard deviation pooling and is expected to provide better performance than those two pooling strategies.

However, using the t value in (1) directly will result in poor prediction with MOS. This is because the t statistic can have extreme values when s is very small. To shrink the extreme t values and obtain good prediction accuracy, we propose to use the following monotonic log-transformation of t values as the final overall quality score:

$$\log(t + K), \tag{2}$$



(a) t (b) $\log(t + 3000)$

Fig. 2. The scatter plots of MOS against (a) t and (b) $\log(t + 3000)$ with fitting curves based on SSIM local scores on the TID2013 database.

Table 1
The ranks of the two images in Fig. 1 for MOS, the mean pooling and the HT pooling: the higher the rank, the higher the quality of the image.

	MOS	Mean pooling	HT pooling
Fig. 1(b)	880	2215	1131
Fig. 1(c)	1317	2261	1584

where K is a sufficiently large constant. We illustrate the effect of this log-transformation in Fig. 2, which depict the scatter plots of MOS against t in Fig. 2(a) and $\log(t + 3000)$ in Fig. 2(b) with fitting curves based on SSIM local scores on the TID2013 database. Clearly, the original t values have a wide range from zero to 7×10^4 with only few extremely large points, while the log-transformation can shrink the values to a reasonable range. Moreover, the log-transformation can improve the prediction accuracy: the Pearson linear correlation coefficient and the root mean squared error of using the original t value are 0.7865 and 0.8287, respectively, while those of using $\log(t + 3000)$ are 0.7881 and 0.8261. This can also be intuitively observed from Fig. 2: the fitting curve of $\log(t + 3000)$ is more linear than that of t.

In the HT pooling, we set $c = 0.8$ as the default value for c because most IQA metric provide local scores within the range $[-1, 1]$ and the local scores of images without severe distortions are usually large positive numbers close to 1. We set $K = 3000$ as the default value for K by observing that the t values calculated from the benchmark databases can be small negative numbers, so we need K to be sufficiently large to make $t + K$ nonnegative.

To illustrate the effectiveness of the HT pooling, we calculate the ranks of Fig. 1(b) and Fig. 1(c) and list them in Table 1, by using different strategies to pool SSIM local scores. We can clearly observe that the HT pooling can provide distinct ranks, which are more consistent with MOS, to well distinguish the qualities of Fig. 1(b) and Fig. 1(c), while the mean pooling provides similar ranks and fails to distinguish the two distortions.

4. Experiments

In the following experiments, we shall show the reliably superior performance of HT pooling for various full-reference and no-reference IQA metrics.

4.1. Experiment settings

4.1.1. For full-reference IQA

We compare the performances of five pooling strategies in the experiments: mean pooling (MP) [3], information content weighting pooling (IWP) [10], standard deviation pooling (SDP) [14], visual saliency pooling (VSP) [11] and hypothesis testing-based pooling (HTP). We use these pooling strategies to pool local scores obtained from five IQA metrics: SSIM [3], GSM [6], FSIM [5], GMSD [14] and VSI [11]. The local scores are calculated on four benchmark databases: LIVE [20], CSIQ [21], TID2008 [22] and TID2013 [23]. To be more specific, for each set of local scores calculated from an IQA metric, we apply five pooling strategies on it separately to calculate the overall scores. In the rest of this short communication, we use the notation 'IQA metric-pooling strategy' to denote an IQA model with the combination of an IQA metric (SSIM, GSM, FSIM, GMSD and VSI) and a pooling strategy (MP, IWP, SDP, VSP or HTP). For example, SSIM-MP denotes the model with mean pooling of the SSIM local scores. Note that FSIM-MP is calculated by using the weighted mean proposed in the original FSIM paper [5]. We use the default parameters for SSIM [3], GSM [6], FSIM [5], IWP [10], GMSD [14] and VSI [11], as stated in their works. For the HT pooling, we set $c = 0.8$ and $K = 3000$.

4.1.2. For no-reference IQA

We assess the performance of HTP on two no-reference IQA metrics, the perception-based image quality evaluator (PIQUE) [18] and blind/referenceless image spatial quality evaluator (BRISQUE) [17], on the LIVE database. PIQUE is an opinion-unaware metric, while BRISQUE is opinion-aware and relies on training a support vector regression model that regresses human ratings against the designed features.

PIQUE adopts MP as its pooling strategy to pool the scores of local patches. Different from the full-reference IQA metrics discussed above, the lower the score of PIQUE, the higher the quality of the image. Thus we slightly adjust the t-test in Section 3.2 when applying it to the local scores of PIQUE. First, the t-test becomes left-sided with an alternative hypothesis H_1 of mean $\mu < c$, while the test statistic in (1) remains the same. Similarly to PIQUE, the lower the value of t, the better the quality of the image. Second, the default setting of $c = 0.8$ is not suitable now, because 0.8 indicates poor quality in PIQUE. The authors of PIQUE suggests that an image with a score of 0.5 can be considered as average quality; thus we follow their suggestion and set $c = 0.5$ in this experiment.

Different from PIQUE, BRISQUE does not provide a quality map: the final quality score is calculated by the regression model directly;

Table 2
Performance measurements on four benchmark databases. SROCC denotes the Spearman rank correlation coefficient. KROCC denotes the Kendall rank order correlation coefficient. PLCC denotes the Pearson linear correlation coefficient. RMSE denotes the root mean square error. The weighted average is the weighted mean of the results of all four databases. The weight for one database is the total number of images in that database divided by that in all four databases. The top two results for each IQA metric are in bold faces.

Database	Criterion	SSIM local scores					GSM local scores					FSIM local scores					GMSD local scores					VSI local scores				
		MP	IWP	SDP	VSP	HTP	MP	IWP	SDP	VSP	HTP	MP	IWP	SDP	VSP	HTP	MP	IWP	SDP	VSP	HTP	MP	IWP	SDP	VSP	HTP
LIVE	SROCC	**0.941**	**0.957**	0.918	0.493	0.940	**0.956**	0.949	0.946	**0.957**	0.946	**0.963**	0.947	**0.960**	0.147	0.958	**0.960**	0.947	**0.960**	**0.960**	0.958	0.951	**0.960**	**0.955**	0.952	0.952
	KROCC	**0.782**	**0.817**	0.740	0.348	**0.782**	**0.815**	0.809	0.797	**0.817**	0.797	**0.834**	0.797	**0.825**	0.101	0.821	0.824	0.797	**0.827**	**0.825**	0.821	0.803	**0.823**	**0.812**	0.806	0.806
	PLCC	0.926	**0.952**	0.908	0.564	**0.937**	**0.951**	0.938	0.941	**0.952**	0.881	**0.960**	0.942	**0.958**	0.177	0.954	0.956	0.942	**0.960**	**0.957**	0.954	0.947	**0.954**	**0.952**	0.948	0.822
	RMSE	10.334	**8.347**	11.428	22.556	**9.567**	**8.432**	9.434	9.256	**8.330**	21.514	**7.659**	9.199	**7.841**	26.894	8.233	8.049	9.199	**7.692**	**7.946**	8.198	8.811	**8.156**	**8.376**	8.682	15.572
CSIQ	SROCC	**0.870**	**0.922**	0.807	0.469	0.845	0.911	0.832	**0.931**	0.915	**0.930**	0.924	0.903	**0.957**	0.087	**0.944**	0.929	0.903	**0.957**	0.936	**0.941**	0.936	0.945	**0.957**	0.942	**0.953**
	KROCC	**0.685**	**0.753**	0.633	0.324	0.653	0.737	0.658	**0.767**	0.746	**0.764**	0.757	0.726	**0.809**	0.058	**0.786**	0.763	0.726	**0.813**	0.775	**0.783**	0.773	0.794	**0.814**	0.785	**0.806**
	PLCC	**0.857**	**0.914**	0.801	0.532	0.827	0.896	0.827	**0.920**	**0.901**	0.821	0.901	0.726	**0.953**	0.295	**0.929**	0.913	0.726	**0.953**	0.921	**0.924**	0.919	**0.928**	**0.952**	**0.928**	0.834
	RMSE	**0.135**	**0.106**	0.157	0.222	0.147	0.116	0.148	**0.103**	**0.114**	0.150	0.114	0.180	**0.080**	0.251	**0.097**	0.107	0.180	**0.080**	0.103	**0.100**	0.103	**0.098**	**0.080**	**0.098**	0.145
TID2008	SROCC	0.776	**0.856**	0.746	0.416	**0.795**	0.850	0.607	0.847	**0.879**	**0.852**	0.881	0.843	**0.892**	0.104	**0.894**	0.848	0.843	**0.891**	0.885	**0.896**	0.880	**0.893**	0.878	**0.898**	0.887
	KROCC	0.577	**0.663**	0.562	0.289	**0.601**	0.659	0.471	0.657	**0.696**	**0.663**	0.694	0.651	**0.706**	0.074	**0.707**	0.653	0.651	**0.709**	0.696	**0.712**	0.687	**0.710**	0.692	**0.712**	0.704
	PLCC	0.776	**0.858**	0.732	0.552	**0.788**	**0.842**	0.466	0.829	**0.867**	0.838	0.874	0.845	**0.880**	0.013	**0.879**	0.837	0.845	**0.880**	0.868	**0.879**	**0.860**	0.881	0.858	0.811	**0.869**
	RMSE	0.846	**0.689**	0.914	1.119	**0.826**	**0.724**	1.187	0.750	**0.669**	0.731	0.653	0.717	**0.638**	1.342	**0.640**	0.735	0.717	**0.637**	0.666	**0.640**	0.684	**0.634**	0.689	0.786	**0.663**
TID2013	SROCC	0.741	**0.778**	0.723	0.405	**0.753**	**0.795**	0.659	0.780	**0.811**	0.783	0.802	0.776	**0.808**	0.092	**0.813**	0.788	0.776	0.804	**0.813**	**0.810**	**0.886**	0.876	0.856	**0.897**	0.863
	KROCC	0.557	**0.598**	0.548	0.282	**0.575**	**0.625**	0.500	0.606	**0.647**	0.610	0.629	0.597	**0.636**	0.065	**0.640**	0.613	0.597	0.634	**0.642**	**0.639**	**0.701**	0.696	0.670	**0.718**	0.679
	PLCC	0.793	**0.832**	0.756	0.582	**0.811**	0.846	0.694	0.816	**0.863**	**0.847**	**0.859**	0.837	0.858	0.205	**0.882**	0.840	0.837	0.858	**0.862**	**0.878**	0.890	**0.891**	0.843	**0.900**	0.874
	RMSE	0.755	**0.688**	0.811	1.008	**0.724**	0.660	0.893	0.717	**0.627**	**0.658**	**0.635**	0.677	0.637	1.213	**0.584**	0.674	0.677	0.636	**0.629**	**0.593**	0.565	**0.564**	0.666	**0.540**	0.602
Weighted Average	SROCC	0.790	**0.838**	0.762	0.426	**0.797**	**0.843**	0.702	0.836	**0.859**	0.839	0.857	0.829	**0.866**	0.101	**0.868**	0.842	0.829	**0.865**	**0.865**	**0.866**	0.896	**0.898**	0.885	**0.907**	0.890
	KROCC	0.606	**0.661**	0.585	0.297	**0.616**	**0.671**	0.550	0.663	**0.692**	0.666	0.687	0.652	**0.700**	0.071	**0.698**	0.668	0.652	**0.700**	0.695	**0.698**	0.717	**0.726**	0.711	**0.734**	0.717
	PLCC	0.811	**0.862**	0.772	0.563	**0.820**	**0.862**	0.679	0.846	**0.877**	0.843	0.878	0.835	**0.886**	0.162	**0.894**	0.860	0.835	**0.887**	0.880	**0.891**	**0.890**	**0.898**	0.873	0.883	0.858

Table 3
The average performance measurements of the pooling strategies over four databases and five IQA metrics, calculated as the averages of the weighted average results in Table 2. SROCC denotes the Spearman rank correlation coefficient. KROCC denotes the Kendall rank order correlation coefficient. PLCC denotes the Pearson linear correlation coefficient. The best measurements are in bold faces.

	MP	IWP	SDP	VSP	HTP
SROCC	0.846	0.820	0.843	0.632	**0.852**
KROCC	0.670	0.648	0.672	0.499	**0.679**
PLCC	0.860	0.822	0.853	0.673	**0.861**

Table 4
Counts of the top two ranks in Table 2, excluding the weighted average results.

	MP	IWP	SDP	VSP	HTP
SSIM	6	16	0	0	11
GSM	8	0	4	14	6
FSIM	6	0	14	0	12
GMSD	1	0	12	8	12
VSI	3	11	8	8	4
Total	24	27	38	30	45

in other words, no pooling strategy is involved. Thus we make the following modification of getting the final BRISQUE score to assess the performance of HTP. For each test image, we divide it to non-overlapping patches of size 32×32, calculate the BRISQUE scores for all patches and pool them to obtain the final score. A lower BRISQUE score refers to a higher quality, thus we adopt the left-sided t-test following the same strategy for PIQUE. However, since the BRISQUE score is not between 0 and 1, we cannot use $c = 0.5$ as in PIQUE to represent average quality. To resolve this problem, we calculate the mean of the BRISQUE scores for the reference images in the LIVE database, which is just above 43, and thus we adopt 43 as the value for c.

4.2. Performance measurements

To compare the objective scores provided by the pooling strategies with the subjective scores provided by the databases, we first transform the original objective scores x using the following regression model to remove their nonlinearity [24]:

$$y = \beta_1 \left(\frac{1}{2} - \frac{1}{1 + \exp(\beta_2 (x - \beta_3))} \right) + \beta_4 x + \beta_5, \tag{3}$$

where x denotes the original objective score, y denotes the transformed objective score, and β_i $(i = 1, \ldots, 5)$ denote the parameters in the regression model which are estimated by minimising the mean squared error between y and the subjective score. In this short communication, we set the initial values of β_i $(i = 1, \ldots, 5)$ following the algorithm in [25], to avoid the non-convergence or the local optimal problems.

The transformed objective scores are compared with the MOS or DMOS provided by the databases in terms of two properties: prediction monotonicity and prediction accuracy. In this short communication, we measure the prediction monotonicity by the Spearman rank order correlation coefficient (SROCC) and the Kendall rank order correlation coefficient (KROCC). The prediction accuracy is measured by the Pearson linear correlation coefficient (PLCC) and root mean squared error (RMSE) [25]. That is, for full-reference IQA, in total we compare 100 combinations of pooling strategies, full-reference IQA metrics, and datasets; and for each combination, we compare their performances in terms of three measures.

4.3. Experiment results

In this section, we first show the results of full-reference IQA for the overall databases and the individual distortion types in Section 4.3.1 and those of no-reference IQA in Section 4.3.2. Then we discuss the sensitivity of the parameters, c and K, used in the HT pooling in Section 4.3.3.

Table 5
The p values of the one-sided paired t-test of HT pooling against other pooling strategies, with H_0: $\mu^m_{HTP} \leq \mu^m_s$ and H_1: $\mu^m_{HTP} > \mu^m_s$. The p values less than 10% are labelled by bold faces.

	HTP *v.s.*			
	MP	IWP	SDP	VSP
SROCC	**0.058**	**0.050**	**0.034**	**0.003**
KROCC	**0.060**	**0.048**	0.129	**0.003**
PLCC	0.853	0.142	0.769	**0.008**

4.3.1. Results of full-reference IQA

Results of the overall databases: The results in Table 2 demonstrate that HTP is a reliable pooling strategy to pool local scores for various IQA metrics in the case of overall databases.

First, on average, HTP ranks in the top two pooling strategies for three IQA metrics, SSIM, FSIM and GMSD. Other pooling strategies rank in the top two for less times: SDP, VSP and IWP rank in the top two for two IQA metrics while MP ranks in the top two for only one IQA metric.

Second, HTP can still provide competitive performance for GSM and VSI metrics, compared with their associated top two pooling strategies. However, other pooling strategies often perform poorly for some IQA metrics, especially the pooling strategies that are designed for specific IQA metrics. For example, IWP can provide superior performance for SSIM local scores. However, it provides the worst performance for GSM local scores. Similarly, VSP can provide superior performance for VSI local scores. However, it fails to provide reliable performances for SSIM and FSIM local scores.

Third, to make the above conclusions clearer, we show in Table 3 the average performance measurements of the pooling strategies over four databases and five IQA metrics, calculated as the averages of the weighted average results in Table 2. It is obvious that HTP has the best average performance for SROCC, KROCC and PLCC. Furthermore, we summarise the counts of ranking in the top two for each pooling strategy in Table 4. We can make two observations from this table. First, HTP totally ranks in the top two 45 times which is the largest among all pooling strategies. Second, the counts of HTP distribute roughly evenly over all IQA metrics, which demonstrates the general reliability of HTP. In contrast, the counts of other pooling strategies distribute unevenly with zero entries for some IQA metrics, which suggests that they are not suitable for all IQA metrics.

Finally, we formally compare the performances of HTP with other pooling strategies via the one-sided paired t-test for SROCC, KROCC and PLCC, respectively, at the significance level of 10%. To be specific, we test H_0: $\mu^m_{HTP} \leq \mu^m_s$ against H_1: $\mu^m_{HTP} > \mu^m_s$, where μ^m_{HTP} denotes the population mean of the performance metric $m \in \{\text{SROCC}, \text{KROCC}, \text{PLCC}\}$ for HT pooling while μ^m_s denotes that for the pooling strategy $s \in \{\text{MP}, \text{IWP}, \text{SDP}, \text{VSP}\}$. For each test, we aggregate the performance metric of each pooling strategy for all databases and IQA metrics in Table 2, excluding the weighted averages. That is, we conduct paired t-tests for two paired samples, each of size $4 \times 5 = 20$. The p values of all tests are reported in Table 5, with those less than 10% labelled by bold faces. Obviously, the two rank correlations of HTP, SROCC and KROCC, are statistically significantly better than other strategies at 10% significance level, except for HTP against SDP whose p value is slightly larger than 10%. This result formally shows the superior performance of HTP on prediction monotonicity.

Results of individual distortion types: To further demonstrate the effectiveness of HTP, we show the SROCCs for individual distortion types on three benchmark databases in Table 6. We do not present the results for TID2008 because TID2013 is an extension of TID2008 and

Table 6
SROCC for different distortion types on three benchmark databases. SROCC denotes the Spearman rank correlation coefficient. KROCC denotes the Kendall rank order correlation coefficient. PLCC denotes the Pearson linear correlation coefficient. RMSE denotes the root mean square error. The weighted average is the weighted mean of the results of all four databases. The weight for one database is the total number of images in that database divided by that in all four databases. The top two results for each IQA metric are in bold faces.

Database	Distortion	SSIM local scores					GSM local scores					FSIM local scores					GMSD local scores					VSI local scores				
		MP	IWP	SDP	VSP	HTP	MP	IWP	SDP	VSP	HTP	MP	IWP	SDP	VSP	HTP	MP	IWP	SDP	VSP	HTP	MP	IWP	SDP	VSP	HTP
LIVE	AWGN-c	**0.967**	**0.967**	0.788	0.736	**0.965**	0.977	**0.986**	**0.986**	0.982	0.985	**0.965**	0.963	**0.968**	0.735	0.961	0.967	0.963	**0.974**	**0.974**	**0.968**	0.980	**0.981**	**0.981**	**0.983**	**0.981**
	GB	0.935	**0.972**	0.943	0.392	**0.957**	0.952	**0.965**	0.958	0.951	**0.959**	**0.971**	**0.968**	0.958	0.405	0.960	**0.959**	**0.968**	0.957	0.957	0.958	0.954	**0.967**	**0.956**	0.953	**0.956**
	JPEG	0.974	**0.981**	0.974	0.460	**0.976**	**0.978**	**0.981**	0.975	**0.978**	0.975	**0.983**	0.978	0.978	0.309	**0.980**	**0.979**	0.978	0.978	**0.979**	**0.979**	0.974	**0.982**	0.975	**0.976**	0.974
	JP2K	0.958	**0.965**	0.955	0.363	**0.968**	0.970	**0.971**	**0.971**	0.970	**0.971**	**0.972**	0.965	0.970	0.030	**0.971**	0.968	0.965	**0.971**	0.969	**0.971**	0.959	**0.970**	**0.967**	0.960	0.966
	J2K	0.930	**0.944**	0.918	0.496	**0.931**	**0.940**	0.897	0.933	**0.939**	0.929	**0.950**	0.911	0.944	0.192	**0.943**	**0.943**	0.911	**0.942**	**0.942**	0.939	0.941	**0.953**	0.941	**0.943**	0.936
CSIQ	AWGN	**0.877**	**0.938**	0.826	0.322	0.858	0.944	0.949	**0.956**	0.953	**0.955**	0.926	0.948	**0.965**	0.113	**0.961**	0.962	0.948	**0.967**	**0.966**	**0.966**	0.963	0.962	**0.968**	0.963	**0.968**
	GB	0.964	**0.978**	**0.972**	0.489	0.959	0.959	**0.974**	**0.973**	0.967	0.970	**0.973**	**0.973**	0.971	0.465	0.966	0.958	**0.973**	**0.971**	0.965	0.962	0.961	**0.970**	**0.971**	0.968	0.967
	JPEG	0.952	**0.966**	**0.957**	0.437	0.954	0.963	**0.968**	0.965	**0.967**	0.965	**0.965**	**0.965**	0.964	0.066	0.963	0.959	**0.965**	**0.965**	0.961	0.962	0.960	**0.971**	**0.967**	0.962	0.965
	JP2K	0.962	**0.968**	**0.976**	0.522	0.961	0.965	0.973	**0.978**	0.971	**0.977**	0.969	0.969	**0.971**	0.355	**0.970**	0.957	**0.969**	**0.972**	0.965	0.964	0.961	**0.975**	**0.975**	0.969	0.972
	CC	**0.790**	**0.954**	0.296	0.486	0.767	0.935	**0.948**	0.934	**0.944**	0.938	**0.942**	**0.956**	0.903	0.395	0.913	0.933	**0.956**	0.904	**0.942**	0.913	0.944	**0.956**	0.917	**0.950**	0.925
	CCS	**0.891**	**0.906**	0.813	0.622	0.862	0.939	**0.951**	**0.948**	0.947	0.947	0.923	0.916	**0.946**	0.137	**0.942**	0.945	0.916	**0.950**	**0.952**	0.948	**0.963**	**0.963**	0.960	**0.964**	0.961
TID2013	AWGN	0.867	0.844	**0.938**	0.365	**0.921**	0.906	0.894	**0.934**	0.920	**0.933**	0.897	0.882	**0.945**	0.150	**0.938**	0.926	0.882	**0.946**	0.934	**0.945**	0.942	0.935	**0.950**	0.946	**0.950**
	AWGN-c	0.757	0.753	**0.847**	0.212	**0.844**	0.818	0.802	**0.843**	0.838	**0.843**	0.822	0.799	**0.868**	0.136	**0.868**	0.849	0.799	**0.869**	0.861	**0.869**	0.865	0.860	**0.869**	**0.871**	**0.869**
	SCN	0.854	0.817	**0.935**	0.574	**0.873**	0.916	0.906	**0.944**	0.933	**0.943**	0.875	0.851	**0.931**	0.385	**0.906**	0.908	0.851	**0.935**	**0.923**	0.919	0.934	0.931	**0.945**	0.937	**0.942**
	MN	**0.726**	**0.802**	0.683	0.021	0.679	**0.729**	**0.721**	0.688	0.693	0.688	**0.794**	**0.742**	0.723	0.034	0.723	**0.753**	**0.742**	0.707	0.724	0.708	**0.793**	**0.805**	0.754	0.769	0.754
	HFN	0.862	0.855	**0.906**	0.363	**0.896**	0.887	0.889	**0.899**	**0.900**	**0.899**	0.898	0.887	**0.915**	0.280	**0.914**	0.905	0.887	**0.916**	**0.915**	**0.915**	0.915	0.911	0.915	**0.920**	**0.916**
	IN	**0.755**	0.727	0.739	0.254	**0.742**	**0.796**	0.790	0.728	**0.813**	0.731	**0.807**	0.752	0.758	0.385	**0.780**	**0.816**	0.752	0.763	**0.835**	0.781	**0.856**	0.820	0.796	**0.874**	0.804
	QN	0.869	0.846	**0.885**	0.572	**0.894**	0.884	0.876	**0.914**	0.860	**0.913**	0.872	0.859	**0.901**	0.055	**0.897**	0.876	0.859	**0.905**	0.850	**0.902**	0.887	0.883	**0.898**	0.875	**0.898**
	GB	**0.964**	**0.970**	0.944	0.471	0.955	**0.969**	**0.967**	0.951	0.962	0.952	**0.955**	**0.968**	0.919	0.001	0.931	**0.960**	**0.968**	0.911	0.957	0.925	**0.966**	**0.964**	0.922	0.961	0.927
	ID	0.923	0.915	**0.945**	0.519	**0.944**	0.943	0.930	**0.948**	0.943	**0.948**	0.930	0.927	**0.950**	0.009	**0.948**	0.942	0.927	**0.952**	0.941	**0.952**	0.948	0.947	**0.956**	0.948	**0.956**
	JPEG	0.915	0.919	**0.931**	0.549	**0.928**	0.928	0.914	**0.936**	0.934	**0.935**	0.932	0.919	**0.951**	0.292	**0.948**	0.939	0.919	**0.951**	0.942	**0.948**	**0.958**	0.952	**0.958**	0.954	**0.958**
	JP2K	0.949	**0.951**	**0.958**	0.619	**0.951**	0.960	0.948	**0.967**	0.962	**0.966**	0.958	0.947	**0.966**	0.486	**0.963**	0.961	0.947	**0.966**	**0.964**	0.963	0.967	**0.970**	**0.970**	**0.971**	**0.970**
	JPEG-t	**0.852**	0.839	0.814	0.422	**0.845**	**0.851**	0.820	0.813	**0.862**	0.816	0.847	0.839	**0.848**	0.185	**0.853**	**0.855**	0.839	0.841	**0.864**	0.847	**0.899**	0.882	0.857	**0.922**	0.862
	JP2K-t	0.885	0.866	**0.902**	0.613	**0.886**	0.918	0.899	**0.922**	**0.923**	**0.922**	0.891	0.876	**0.916**	0.431	**0.912**	0.907	0.876	**0.914**	**0.919**	**0.914**	0.924	0.919	**0.926**	0.923	**0.928**
	NEPN	0.787	**0.801**	0.791	0.265	**0.793**	**0.813**	**0.830**	0.807	**0.813**	0.807	0.792	**0.825**	**0.812**	0.030	**0.812**	0.805	**0.825**	**0.814**	0.799	**0.814**	0.804	**0.808**	**0.806**	**0.806**	0.805
	LBD	**0.571**	0.372	**0.582**	0.092	**0.571**	**0.641**	0.236	0.540	0.525	**0.546**	0.549	0.417	**0.665**	0.254	**0.657**	0.645	0.417	**0.662**	0.551	**0.659**	0.334	0.359	**0.668**	0.172	**0.674**
	MS	**0.774**	**0.782**	0.696	0.256	0.719	**0.787**	0.751	0.785	**0.793**	0.785	**0.752**	0.572	**0.730**	0.157	**0.730**	0.726	0.572	**0.734**	0.692	**0.735**	0.771	0.749	**0.772**	0.769	**0.773**
	CC	**0.373**	**0.459**	0.004	0.002	0.194	**0.485**	0.459	0.446	**0.486**	0.447	**0.468**	**0.457**	0.307	0.282	0.330	**0.482**	0.457	0.323	**0.477**	0.346	**0.475**	**0.477**	0.393	**0.475**	0.399
	CCS	**0.415**	**0.419**	0.325	0.112	0.308	**0.357**	**0.335**	0.313	0.264	0.313	0.274	**0.386**	0.212	0.217	**0.335**	**0.369**	**0.386**	0.294	0.297	0.294	**0.814**	0.802	0.810	0.810	**0.813**
	MGN	0.783	0.773	**0.849**	0.343	**0.836**	0.835	0.847	**0.867**	**0.868**	0.866	0.847	0.820	**0.881**	0.207	**0.876**	0.859	0.820	**0.889**	0.884	**0.886**	0.894	0.892	**0.902**	**0.912**	**0.902**
	CN	0.862	0.876	**0.894**	0.558	**0.887**	0.912	0.911	**0.923**	**0.921**	**0.923**	0.912	0.904	**0.929**	0.085	**0.927**	0.912	0.904	**0.930**	0.921	**0.928**	0.917	0.922	**0.933**	0.924	**0.932**
	LCN	0.910	0.904	**0.958**	0.543	**0.942**	0.956	0.956	**0.967**	0.965	**0.967**	0.947	0.934	**0.964**	0.022	**0.961**	0.945	0.934	**0.963**	0.953	**0.962**	0.950	0.958	**0.964**	0.956	**0.963**
	ICQD	0.862	0.840	**0.900**	0.314	**0.899**	0.897	0.889	**0.912**	0.886	**0.912**	0.876	0.872	**0.911**	0.332	**0.912**	0.903	0.872	**0.910**	0.888	**0.911**	**0.912**	0.896	0.904	0.884	**0.906**
	CA	**0.864**	**0.868**	0.832	0.355	0.831	**0.882**	0.846	0.859	**0.875**	0.860	**0.871**	0.848	0.861	0.397	**0.863**	**0.880**	0.848	0.853	**0.873**	0.853	**0.899**	0.882	0.880	**0.890**	0.881
	SSR	0.949	0.947	**0.961**	0.686	**0.956**	0.967	0.964	**0.970**	0.966	**0.970**	0.956	0.957	**0.967**	0.278	**0.963**	0.961	0.957	**0.968**	0.960	**0.966**	0.962	0.966	**0.968**	0.963	**0.968**

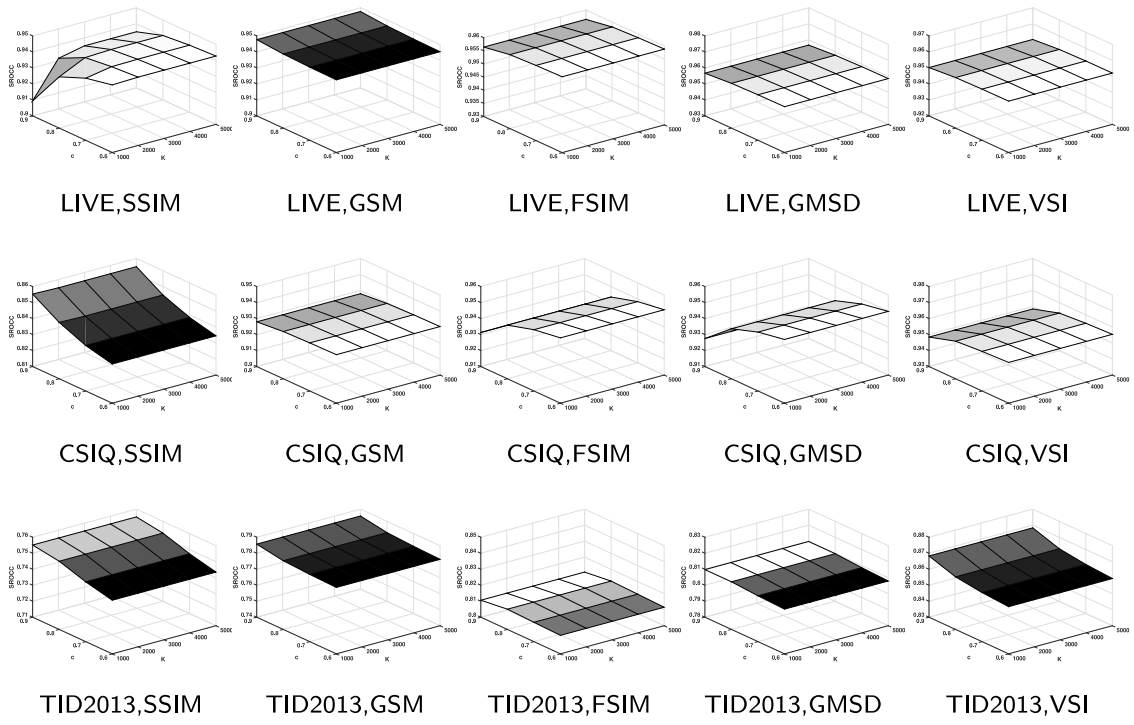


Fig. 3. The sensitivities of SROCC of SSIM-HTP, GSM-HTP, FSIM-HTP, GMSD-HTP and VSI-HTP to parameters c and K for three databases.

Table 7
Counts of the top two ranks in Table 6.

	MP	IWP	SDP	VSP	HTP
SSIM	13	19	18	0	23
GSM	12	12	20	14	18
FSIM	14	9	22	0	27
GMSD	10	9	25	14	19
VSI	10	15	22	14	20
Total	59	64	107	42	107

Table 8
Performance comparison between MP and HTP on the LIVE database based on two no-reference IQA metrics, PIQUE and BRISQUE. Better performances are in bold faces.

	PIQUE local scores		BRISQUE local scores	
	MP	HTP	MP	HTP
SROCC	0.840	**0.856**	0.825	**0.835**
KROCC	0.638	**0.657**	**0.622**	**0.622**
PLCC	0.836	**0.850**	0.821	**0.833**

covers all the distortion types in TID2008. The top two SROCCs for each IQA metric are in bold faces.

Similarly to that for overall databases, we summarise the counts of ranking in the top two for each pooling strategy in Table 7. Paired t-tests are not performed here, because for a lot of distortions in TID2013 we only have five observations for each pooling strategy, which makes the tests less reliable. The following two conclusions can be drawn from Table 7.

First, HTP has the largest total counts and distributes evenly over different IQA metrics. This observation demonstrates the superior performance of HTP for individual distortion types.

Second, SDP also has superior performance for individual distortions. However, it shows worse performance in Table 4 for overall databases.

To sum up, based on the results in this section, HTP has a high general reliability: it can provide superior overall quality scores for both overall databases and individual distortions and also for different IQA metrics. However, other pooling strategies discussed in this paper do not exhibit such general reliability.

4.3.2. Results of no-reference IQA

The performances of MP and HTP to pool PIQUE and BRISQUE local scores are presented in Table 8. Clearly, HTP performs better than MP in terms of all three measures, which shows the superior ability of HTP to pool the local scores of these two no-reference IQA metrics. Note that here the performances of BRISQUE are not as good as in Anish et al. [17]. This is because BRISQUE is originally calculated based on the features extracted from the whole image while our setting obtains the BRISQUE local scores on small patches which affects the final performance. Nevertheless, the purpose of this experiment is to demonstrate the performance of HTP strategy rather than providing the best IQA metric.

4.3.3. Sensitivity to parameters

The sensitivities of SROCC of SSIM-HTP, GSM-HTP, FSIM-HTP, GMSD-HTP and VSI-HTP to parameters c and K are shown in Fig. 3 for three databases (LIVE, CSIQ, TID2013). The parameter c is tested on four values: 0.6, 0.7, 0.8 and 0.9; and the parameter K is tested on five values: 1000, 2000, 3000, 4000 and 5000. It is clear that SROCC is stable with the value of K over different methods and different databases. Although SROCC is slightly sensitive to c, the conclusions in Table 2 still hold, even with the worst values of SROCC with respect to c. This also demonstrates the effectiveness of HT pooling.

5. Conclusion and future work

In this short communication, for the first time, we introduce statistical hypothesis testing to pooling for IQA. The HT pooling can provide more reliable scores than the summary statistics-based pooling. For illustrative purposes, we design a new HT pooling strategy based on

the one-sample one-sided t-test. This new strategy shows reliable performance on local scores calculated by five full-reference IQA metrics (SSIM, GSM, FSIM, GMSD and VSI) and two no-reference IQA metric (PIQUE, BRISQUE).

The current version of HT pooling does not consider the spatial information provided by the local neighbourhood which is vital in image data analysis; thus we could involve such information to enhance HT pooling. The HT pooling can also be applied to recent deep learning-based no-reference IQA metrics. Another future direction is to explore the incorporation of the HT pooling in the training procedures of deep networks for no-reference IQA.

CRediT authorship contribution statement

Rui Zhu: Conceptualization, Methodology, Formal analysis, Writing. **Fei Zhou:** Methodology, Writing. **Wenming Yang:** Methodology, Writing. **Jing-Hao Xue:** Methodology, Writing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data are publicly available.

Acknowledgements

The authors would like to thank the anonymous reviewers and the editor for their constructive comments. This work was partly supported by the Overseas Cooperation Foundation of Tsinghua University, China and the Special Foundation for the Development of Strategic Emerging Industries of Shenzhen, China (No. JCYJ20200109143010272).

References

[1] F. Zhou, W. Yang, X. Gao, H. Liu, R. Zhu, J.-H. Xue, Special issue on advances in statistical methods-based visual quality assessment, Signal Process., Image Commun. 83 (2020) 115695.

[2] N. Damera-Venkata, T.D. Kite, W.S. Geisler, B.L. Evans, A.C. Bovik, Image quality assessment based on a degradation model, IEEE Trans. Image Process. 9 (4) (2000) 636–650.

[3] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.

[4] Z. Wang, E.P. Simoncelli, A.C. Bovik, Multiscale structural similarity for image quality assessment, in: Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on, Vol. 2, IEEE, 2003, pp. 1398–1402.

[5] L. Zhang, L. Zhang, X. Mou, D. Zhang, FSIM: A feature similarity index for image quality assessment, IEEE Trans. Image Process. 20 (8) (2011) 2378–2386.

[6] A. Liu, W. Lin, M. Narwaria, Image quality assessment based on gradient similarity, IEEE Trans. Image Process. 21 (4) (2012) 1500–1512.

[7] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, W. Samek, Deep neural networks for no-reference and full-reference image quality assessment, IEEE Trans. Image Process. 27 (1) (2017) 206–219.

[8] M.-l. Zhu, D.-y. Ge, Image quality assessment based on deep learning with FPGA implementation, Signal Process., Image Commun. 83 (2020) 115780.

[9] X. Jiang, L. Shen, G. Feng, L. Yu, P. An, An optimized CNN-based quality assessment model for screen content image, Signal Process., Image Commun. 94 (2021) 116181.

[10] Z. Wang, Q. Li, Information content weighting for perceptual image quality assessment, IEEE Trans. Image Process. 20 (5) (2011) 1185–1198.

[11] L. Zhang, Y. Shen, H. Li, VSI: A visual saliency-induced index for perceptual image quality assessment, IEEE Trans. Image Process. 23 (10) (2014) 4270–4281.

[12] A.K. Moorthy, A.C. Bovik, Visual importance pooling for image quality assessment, IEEE J. Sel. Top. Sign. Proces. 3 (2) (2009) 193–201.

[13] J. Gu, G. Meng, S. Xiang, C. Pan, Blind image quality assessment via learnable attention-based pooling, Pattern Recognit. 91 (2019) 332–344.

[14] W. Xue, L. Zhang, X. Mou, A.C. Bovik, Gradient magnitude similarity deviation: A highly efficient perceptual image quality index, IEEE Trans. Image Process. 23 (2) (2014) 684–695.

[15] G.E. Box, J.S. Hunter, W.G. Hunter, Statistics for Experimenters: Design, Innovation, and Discovery, Vol. 2, Wiley-Interscience New York, 2005.

[16] R. Zhu, F. Zhou, W. Yang, J.-H. Xue, On hypothesis testing for comparing image quality assessment metrics [Tips & Tricks], IEEE Signal Process. Mag. 35 (4) (2018) 133–136.

[17] A. Mittal, A.K. Moorthy, A.C. Bovik, No-reference image quality assessment in the spatial domain, IEEE Trans. Image Process. 21 (12) (2012) 4695–4708.

[18] N. Venkatanath, D. Praneeth, M.C. Bh, S.S. Channappayya, S.S. Medasani, Blind image quality evaluation using perception based features, in: 2015 Twenty First National Conference on Communications, NCC, IEEE, 2015, pp. 1–6.

[19] R.A. Manap, L. Shao, A.F. Frangi, PATCH-IQ: a patch based learning framework for blind image quality assessment, Inform. Sci. 420 (2017) 329–344.

[20] H.R. Sheikh, Z. Wang, L. Cormack, A.C. Bovik, LIVE Image Quality Assessment Database Release 2, URL http://live.ece.utexas.edu/research/quality.

[21] E.C. Larson, D.M. Chandler, Most apparent distortion: full-reference image quality assessment and the role of strategy, J. Electron. Imaging 19 (1) (2010) 011006.

[22] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, F. Battisti, TID2008-a database for evaluation of full-reference visual quality assessment metrics, Adv. Mod. Radioelectron. 10 (4) (2009) 30–45.

[23] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, et al., Image database TID2013: Peculiarities, results and perspectives, Signal Process., Image Commun. 30 (2015) 57–77.

[24] H.R. Sheikh, M.F. Sabir, A.C. Bovik, A statistical evaluation of recent full reference image quality assessment algorithms, IEEE Trans. Image Process. 15 (11) (2006) 3440–3451.

[25] W. Sun, F. Zhou, Q. Liao, MDID: A multiply distorted image database for image quality assessment, Pattern Recognit. 61 (2017) 153–168.