A Measurement for Distortion Induced Saliency Variation in Natural Images

Xiaohan Yang, Fan Li, *Member, IEEE,* and Hantao Liu

***Abstract*—How best to measure spatial saliency shift induced by image distortions is an open research question. Our previous study has shown that image distortions cause saliency to deviate from its original places in natural images, and the degree of such distortion-induced saliency variation (DSV) depends on image content as well as the properties of distortion. Being able to measure DSV benefits the development of saliency based image quality algorithms. In this paper, we first investigate the plausibility of using existing mathematical algorithms for measuring DSV and their potential limitations. We then develop a new algorithm for quantifying DSV, based on a deep neural network. In the algorithm, namely ST-DSV, we design a coarse-grained to fine-grained saliency similarity transformation approach to achieve DSV measurement. The experimental results show that the proposed ST-DSV algorithm significantly outperforms existing methods in predicting the ground truth DSV.**

***Index Terms*—Saliency, distortion, similarity measure, image quality assessment, deep neural network**

I. Introduction

WITH the rapid development of multimedia technology, a large amount of digital images are being generated, stored, processed and transmitted every day. These images are also widely shared across social media [1]-[4]. We are facing an unprecedented situation where end-users expect high quality visual experiences. Hence, image quality assessment (IQA) has become a popular research topic in both academia and industry [5]-[12]. Recently, a significant approach in IQA is to integrate saliency to its objective algorithms [13]-[18]. However, challenges to optimizing the use of saliency in IQA algorithms remain.

Recent research shows that image distortion causes gaze distraction, resulting in the shift of saliency from its original locations [19], [20]; and that being able to measure the distortion-induced saliency variation (DSV) significantly helps improve the accuracy of IQA algorithms [21]-[23]. How best to quantify DSV remains an open research question, which is the topic to be investigated in this paper.

Our previous research established a ground truth benchmark for the measurement of the distortion-induced saliency variation (DSV) [24]. First, the SIQ288 database was constructed, where human saliency maps of distorted images and their references were rigorously derived from 5760 eye movement trials recorded with 160 human observers. The database consists of 288 test images selected from the recognized LIVE database [25], including 18 pristine reference images and 270 distorted images with five different distortion types and three distortion levels. The distortion types include JPEG compression (JPEG), JP2K compression (JP2K), white noise (WN), Gaussian blur (GB), and fast-fading (FF). The same reference image is distorted by each distortion type with three quality degradation levels (i.e., High, Medium, Low). Table I illustrates the outline of the composition of the SIQ288 database, including the reference images (RI) and distorted images (DI). Note, each image has a saliency map derived from eye movement trials. The analysis of the saliency maps shows that, because of the distortion-induced saliency variation (DSV), the visual saliency of a distorted image is different from that of the pristine reference image. Then, we built a benchmark to measure DSV by using a subjective measurement method [26]. Sixteen experts in computer vision were requested to observe and score the similarity between the saliency map of a distorted image and that of the pristine reference image. Finally, the difference mean saliency variation score (DMSS) of each ”distorted” saliency map was obtained. DMSS indicates the perceived difference between the ”distorted” saliency map and the ”original” saliency map, which quantifies DSV. Fig. 1 shows the DSV of the FF distorted images in the benchmark. The higher the DMSS, the less similar the deviated saliency is from the “original” saliency.

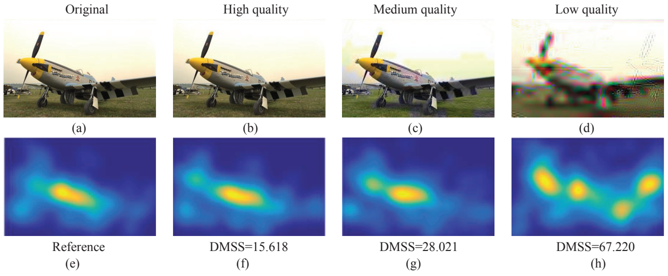


Fig. 1. Illustration of the distortion-induced saliency variation (DSV). DMSS (difference mean saliency variation score) represents the degree of similarity between the deviated saliency map and the reference saliency map. The higher the DMSS, the less similar the deviated saliency is from the “original” saliency.

In practice, saliency maps could be computed; however, measuring distortion-induced saliency variation (DSV) via subjective testing is impractical for real-world applications. A more realistic way to integrate DSV into IQA algorithms is to develop a metric for DSV, which can automatically predict the subjective DMSS. Unfortunately, so far there is no dedi-

Xiaohan Yang and Fan Li are with the Ministry of Education Key Laboratory for Intelligent Networks and Network Security, School of Information and Communications Engineering, Xi’an Jiaotong University, Xi’an, 710049, China. (e-mail: yangxiaohan@stu.xjtu.edu.cn; lifan@mail.xjtu.edu.cn).

Hantao Liu is with the School of Computer Science and Informatics, Cardiff University, Cardiff, CF243AA, U.K. (e-mail: LiuH35@cardiff.ac.uk).

This research work was supported in part by National Science Foundation of China (62071369)

TABLE I
The outline of the SIQ288 database, including the reference images (RI) and distorted images (DI). Each image has a saliency map derived from eye movement trials.

RI	DI
Bikes	5(distortion types)*3(quality levels)=15
Buildings	5*3=15
Caps	5*3=15
Cemetry	5*3=15
Lighthouse1	5*3=15
Lighthouse2	5*3=15
Manfishing	5*3=15
Monarch	5*3=15
Ocean	5*3=15
Paintedhouse	5*3=15
Parrots	5*3=15
Plane	5*3=15
Rapids	5*3=15
Sailing	5*3=15
Statue	5*3=15
Stream	5*3=15
Studentsculpture	5*3=15
Womanhat	5*3=15
Total number:18	Total number:270

cated DSV metric. There are few mathematical algorithms, as described by the MIT saliency benchmark [27], that calculate the similarity between saliency maps so could be potentially used to evaluate DSV [26]. However, these algorithms may not be adequate for DSV. In particular, these similarity measures focus on evaluating a computational saliency map's ability in predicting salient objects in natural images. These measures may not be able to capture the saliency variation inducted by images distortions, which is essential for DSV. In this paper, we analyze existing algorithms of saliency similarity and their capabilities and limitations of evaluating DSV.

We also propose a new algorithm to quantify DSV. The proposed algorithm, namely ST-DSV, uses a deep neural network (DNN) to realize a coarse-grained to fine-grained saliency similarity transformation. More specifically, first a coarse-grained saliency similarity DNN is constructed to capture the similarity level when comparing different deviated saliency maps relative to the same reference saliency map. Then, the coarse-grained DNN is transferred to a fine-grained saliency similarity DNN to learn the precise similarity score (i.e., DMSS) across different reference saliency maps. The proposed algorithm is proven effective for the measurement of DSV.

The remainder of this paper is organized as follows. Section II analyzes the ability of existing saliency similarity metrics for the evaluation of DSV. Section III describes the proposed ST-DSV algorithm in detail. Section IV demonstrates the experimental results. Section V gives discussions and Section VI concludes the paper.

II. Analysis of state-of-the-art algorithms for the evaluation of DSV

We hypothesize that existing saliency similarity measures cannot sufficiently capture saliency shift induced by image distortions, so cannot quantify DSV. Now, we verify this hypothesis below.

A. Saliency similarity metrics

We use similarity metrics to measure DSV between a reference saliency map (i.e., originated from a pristine reference image) and a deviated/distorted saliency map (i.e., originated from a corresponding distorted image of the reference). Popular metrics are the area under the receiver operating characteristic curve (AUC), including AUC-Borji [28] and s-ACU [29], Normalized Scanpath Saliency (NSS) [30], Information Gain (IG) [31], Similarity (SIM) [32], Pearson's Correlation Coefficient (CC) [33], Kullback-Leibler Divergence (KL) [34], and Earth Mover's Distance (EMD) [35]. The general use of these metrics is already described in more detail in [28]-[35], and we only briefly repeat their meaning in the context of DSV measurement as follows:

In the approach of the AUC-Borji [28] and s-AUC [29], the distorted map is interpreted as a classifier of whether pixels are fixated or not, according to the fixations of the reference. The perfect similarity of AUC-Borji and s-AUC corresponds to a score of 1 while a score of 0.5 indicates a chance level.

The NSS [30] measures the average of values of the distorted saliency map at fixation locations of the reference. When $NSS > 0$, the higher the value of the measure, the more similar the saliency maps are. $NSS <= 0$ indicates that the comparison is not meaningful.

The IG [31] measures to what extent the information gain of the saliency map is better than the centre prior baseline. The IG above zero indicates the measurement on the fixation locations of the distorted saliency map is better than the centre prior baseline.

The SIM [32] is to sum up the minimum saliency value at every pixel location between the distorted map S_d and the pristine reference saliency map S_p. The SIM ranges between 0 and 1. A larger SIM value indicates a higher similarity between S_d and S_p.

The CC [33] is the linear correlation coefficient, which assesses the linear relationship between S_d and S_p. The CC ranges between -1 to 1. When the correlation value is close to -1 or 1, there is a perfect similarity between S_d and S_p.

The KL [34] aims to use the entropy-based metric to measure the difference between S_d and S_p. The smaller the KL, the higher the similarity between S_d and S_p.

The EMD [35] aims to measure the spatial distance between S_d and S_p to solve the transportation problem from linear optimization. The goal is to describe how spatially far away S_d is from S_p. A larger EMD indicates a higher dissimilarity between S_d and S_p, while the EMD of zero indicates that the two saliency maps are the same.

It should be noted that the above-mentioned saliency similarity metrics exhibit different characteristics, so they capture different aspects of pattern "similarity", as already discussed in [34]. It is often not straightaway to compare the advantages and disadvantages between these metrics directly, as some metrics might be more suitable for certain applications (e.g., object detection), but less suitable for other applications (e.g., emotions).

The intention of the paper, however, is not to replace exciting saliency similarity metrics for general-purpose applications. Our goal is to find an effective solution to quantify

TABLE II
The performance (i.e.,SROCC, PLCC and KROCC) of existing saliency similarity metrics based on DSV benchmark database [26].

	AUC-Borji	s-AUC	NSS	IG	CC	SIM	KL	EMD
SROCC	0.509	0.536	0.612	0.299	0.734	0.708	0.302	0.573
PLCC	0.514	0.536	0.680	0.287	0.786	0.758	0.336	0.637
KROCC	0.372	0.354	0.433	0.203	0.544	0.516	0.209	0.402

TABLE III
Illustration of two categories of saliency similarity metrics.

Category	Location-based	Distribution-based
Metric	s-AUC,AUC-Borji,NSS, IG	SIM,CC,EMD,KL
Required format of reference	Fixation locations	Fixation density map
Required format of test stimulus	Fixation density map	Fixation density map

distortion-induced saliency variation (DSV); and analyzing existing metrics is an intuitive step to check if these metrics are suitable for such specific application of measuring DSV. The benchmark dataset gives ”ground truth” for DSV measurement [26]. In making the benchmark, expert human subjects assessed DSV in a fully-controlled perception experiment; the DSV quantification by means of DMSS was rendered from the subjective assessment. Since human visual system (HSV) is so far the most reliable assessor/metric of visual information [25], [39], [55], subjective DSV (i.e., DMSS) can be regarded as the ”ground truth” measurement. By analyzing existing metrics against the ”ground truth”, insights are provided as to what would be beneficial for developing a suitable metric.

In order to evaluate the performance of the saliency similarity metrics on DSV, we can calculate the correlation between each metric (i.e., predicted scores) and the subjective scores of DSV (i.e., DMSS), using Spearman Rank-Order Correlation Coefficient (SROCC) [36], Kendall’s Rank Order Correlation Coefficient (KROCC) [37] and Pearson Linear Correlation Coefficient (PLCC) [38]. The SROCC and KROCC represent the prediction monotonicity and the PLCC measures the prediction accuracy [39]. Higher SROCC, KROCC and PLCC values indicate higher correlation between an objective metric and the ground truth of the DSV. The SROCC, KROCC and PLCC results on the DSV benchmark database are listed in Table II.

It can be seen from Table II that the saliency similarity metrics are not strongly correlated with the ground truth of the DSV. Among the eight metrics, the performance of CC and SIM is better than other metrics. Metrics, such as IG and KL show very poor correlation with subjective DSV. Therefore, it is worthwhile to further analyze the capabilities and limitations of these metrics in quantifying DSV.

B. The limitations of saliency similarity metrics

The saliency similarity metrics can be divided into two categories, including the location-based and the distribution-based metrics [34], as shown in Table III. The location-based metrics use the discrete fixation locations of the pristine reference and the fixation density (i.e., saliency) map of the distorted counterpart to measure saliency similarity, while the distribution-based metrics use the saliency maps of both pristine reference and distorted counterpart. Hence, s-AUC, AUC-Borji, NSS and IG metrics are classified as the location-based metrics and SIM, CC, EMD, KL metrics are the distribution-based metrics.

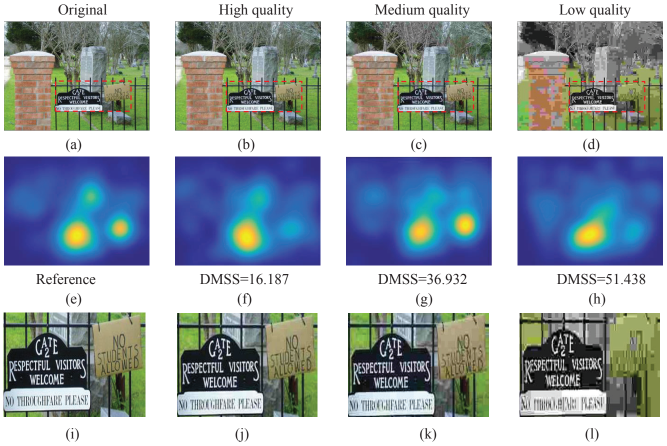


Fig. 2. The saliency maps of the pristine reference “Cemetry” image from the SIQ288 database and its high, medium and low quality images (distorted by JEPG). (a)-(d) are the images of reference, high, medium and low quality. (e)-(h) are their saliency maps. (i)-(l) are the image patches extracted from (a)-(d) to better visualize distortions (i.e., as indicated by the red boxes in (a)-(d).

TABLE IV
The comparison of DMSS and the location-based metrics for Fig. 2.

Metric	High quality	Medium quality	Low quality
DMSS	16.187	36.932	51.438
AUC-Borji	0.627	0.662	0.618
s-AUC	0.579	0.609	0.558
NSS	1.575	1.623	1.489
IG	0.547	0.612	0.544

1) The limitations of location-based metrics: As shown in Table II, the performance of location-based metrics is rather poor. The correlation of the IG metric is less than 0.3, which means IG is inconsistent with subjective DSV. Furthermore, the AUC-Borji , s-AUC and NSS metrics cannot quantify DSV accurately, because the average correlation is low and around 0.5-0.6.

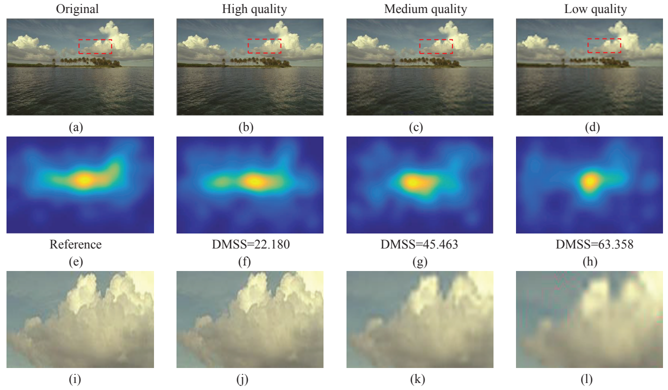


Fig. 3. The saliency maps of the pristine reference "Ocean" image from the SIQ288 database and its high, medium and low quality images (distorted by FF). (a)-(d) are the images of reference, high, medium and low quality. (e)-(h) are their saliency maps. (i)-(l) are the image patches extracted from (a)-(d) to better visualize distortions (i.e., as indicated by the red boxes in (a)-(d).

TABLE V
The comparison of DMSS and the location-based metrics for Fig. 3.

Metric	High quality	Medium quality	Low quality
DMSS	22.180	45.463	63.358
AUC-Borji	0.664	0.614	0.587
s-AUC	0.613	0.538	0.537
NSS	1.489	1.377	1.385
IG	0.253	0.267	0.214

Fig. 2 shows an examples of how DSV is measured subjectively by the ground truth DMSS scores; and that the higher the distortion (i.e., the lower the image quality), the larger the saliency variation relative to the original saliency. As shown in Table IV, the DMSS score increases with the increase of the distortion level. The table also shows the objective measures of DSV using the AUC-Borji, s-AUC, NSS and IG. It can be seen that these metrics fail in capturing the properties of DSV. They all fail the instance of ”Medium quality”, where metrics give the largest value of saliency variation. For a good metric, we should expect that the metric's values monotonically decrease as DMSS values increase. Note that the values of DMSS represent dissimilarity (i.e., the higher the value, the more dissimilar); the values of metrics represent similarity (i.e., the higher the value, the more similar.)

Fig. 3 shows an additional example of ground truth DSV measured by the DMSS scores. Table V shows the results of DSV measured by the location-based metrics (AUC-Borji, s-AUC, NSS and IG). As can be seen from Fig. 3 and Table V, DMSS scores increase as distortions increase. This means when image quality degrades, its saliency becomes more dissimilar to the reference saliency (of the pristine image). AUC-Borji and s-AUC metrics seem to capture the tendency of DSV in this example, but it should be noted that their values around 0.5 indicate the prediction is likely to be meaningless (i.e., around chance level) [23]. For the AUC based metrics, chance is at 0.5 and AUC scores larger than 0.5 indicate correspondence between maps above chance [28], [29], [34]. More critically, the MIT saliency benchmark [27] suggested the baseline for AUC-Borji is 0.66 and s-AUC is 0.63; and scores below the baseline indicate performance below chance level. Therefore, both metrics are not suitable for DSV measurement. The NSS metric fails the ”Medium quality” instance, where metric gives the lowest value of saliency variation; while the IG metric fails the ”Medium quality” instance, where metric gives the highest value of saliency variation.

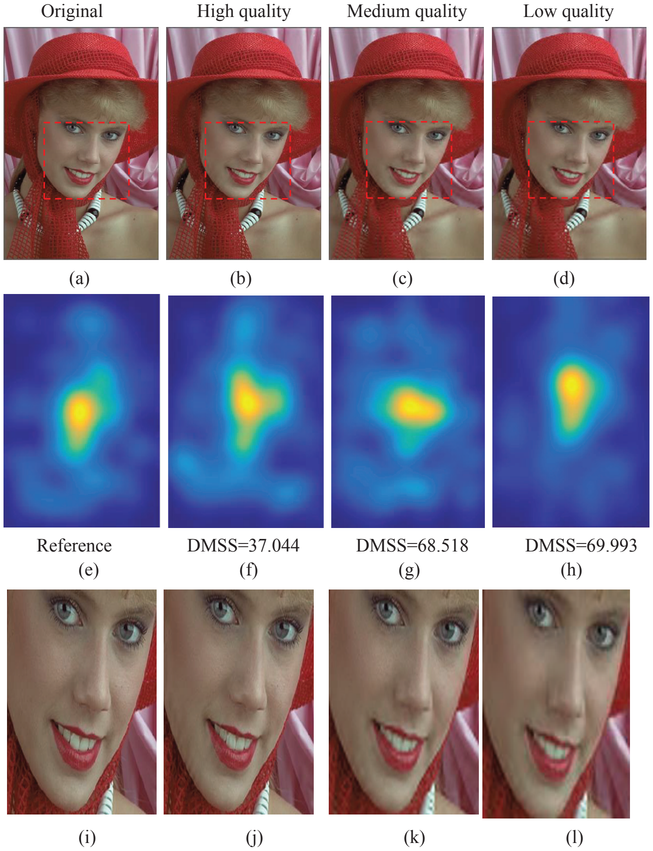


Fig. 4. The saliency maps of the pristine reference ”Womanhat” image from the SIQ288 database and its high, medium and low quality images (distorted by FF). (a)-(d) are the images of reference, high, medium and low quality. (e)-(h) are their saliency maps. (i)-(l) are the image patches extracted from (a)-(d) to better visualize distortions (i.e., as indicated by the red boxes in (a)-(d).

TABLE VI
The comparison of DMSS and the KL and EMD metrics for Fig. 4.

Metric	High quality	Medium quality	Low quality
DMSS	37.044	68.518	69.993
KL	0.144	0.143	0.139
EMD	0.725	0.727	0.613

In summary, these location-based metrics, when used for DSV measurement, are inconsistent with ”ground truth” DMSS scores, as shown in Table IV and V. The analyses indicate that s-AUC, AUC-Borji, NSS and IG metrics are not suitable candidates for DSV quantification. This is probably attributed to the fact that these metrics focus on locality (i.e., fixation locations) as their predominant determinant for the saliency similarity measurement. However, when DSV is measured between the reference and distorted saliency maps, both the locality and spatial distribution of saliency are important influencing factors [26]. This may explain why these location-based metrics give unsatisfactory results, as they cannot capture changes of spatial saliency distribution.

2) *The limitations of distribution-based metrics:* In general, distribution-based metrics are applicable for measuring the DSV, because their calculation is based on comparing the reference and distorted saliency maps. However, both KL and EMD give almost identical scores to the "High quality" and "Medium quality" instances. This indicates that KL and EMD fail in distinguishing the saliency variation induced by "high" and "medium" image quality. Also, both KL and EMD give an inconsistent measure for the "Low quality" instance, since its measure should be the largest for this instance. For KL and EMD metrics, Table II illustrates that they are inconsistent with the subjective DSV, as the correlation is low. These metrics are based on the measure of probability distribution of saliency values, which is, however, not sensitive to local structural changes in saliency. Fig. 4 and Table VI show the saliency maps of a reference image, and the corresponding distorted images from the SIQ288 database; and how the DSV is measured by the ground truth DMSS and by the objective metrics, KL and EMD, respectively.

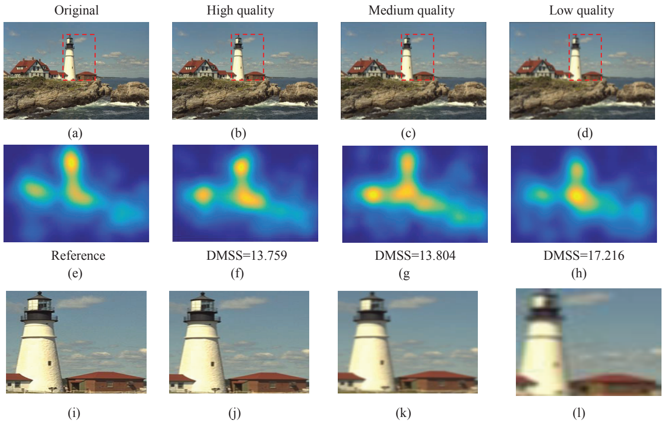


Fig. 5. The saliency maps of the pristine reference "Lighthouse2" image from the SIQ288 database and its high, medium and low quality images (distorted by GB). (a)-(d) are the images of reference, high, medium and low quality. (e)-(h) are their saliency maps. (i)-(l) are the image patches extracted from (a)-(d) to better visualize distortions (i.e., as indicated by the red boxes in (a)-(d).

TABLE VII
The comparison of DMSS and the SIM and CC metrics for Fig. 5.

Metric	High quality	Medium quality	Low quality
DMSS	13.759	13.804	17.216
SIM	0.825	0.840	0.817
CC	0.889	0.908	0.870

It can be seen from Fig. 4 that with the increase of distortion (see Fig. 4 (a)-(d)), the saliency variation becomes larger (see Fig. 4 (e)-(h)). As shown in Table VI, this trend is clearly reflected by the ground truth DMSS. However, both KL and EMD give almost identical scores to the "High quality" and "Medium quality" instances. This indicates that KL and EMD fail in distinguishing the saliency variation induced by "high" and "medium" image quality. Also, both KL and EMD give an inconsistent measure for the "Low quality" instance, where the predicted score should be the largest for either KL or EMD (note, for KL and EMD, the larger the value, the more dissimilar of the measure).

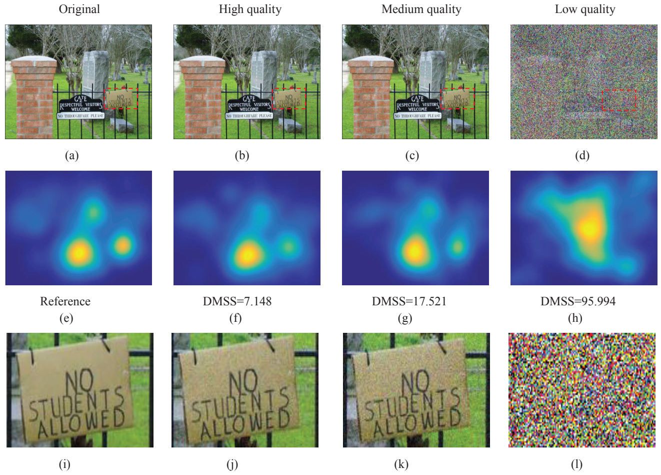


Fig. 6. The saliency maps of the pristine reference "Cemetry" image from the SIQ288 database and its high, medium and low quality images (distorted by WN). (a)-(d) are the images of reference, high, medium and low quality. (e)-(h) are their saliency maps. (i)-(l) are the image patches extracted from (a)-(d) to better visualize distortions (i.e., as indicated by the red boxes in (a)-(d).

TABLE VIII
The comparison of DMSS and the distribution-based metrics for Fig. 6.

Metric	High quality	Medium quality	Low quality
DMSS	7.148	17.521	95.994
EMD	1.226	0.733	1.718
SIM	0.883	0.887	0.630
CC	0.965	0.970	0.555

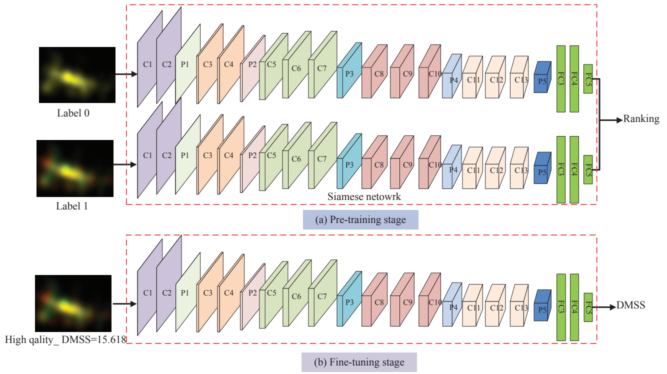
For CC and SIM metrics, as shown in Table II, their performance in measuring DSV is better than other metrics. CC and SIM metrics, to some extent, can reflect the degree of saliency deviation relative to the reference. However, these simple metrics exhibit some limitations in dealing with complex saliency patterns. For example, some distortion types, such as Gaussian blur (GB) and white noise (WN), represent evenly distributed distortions in an image. In the saliency map of the distorted image, gaze tends to be concentrated on (not deviated significantly from) the areas with highly salient features.

Fig. 5 shows the reference image and its GB distorted images (see Fig. 5 (a)-(d)), and their corresponding saliency maps (see Fig. 5 (e)-(h)). It can be seen from Fig. 5 that saliency variation is rather subtle for the “Medium quality” and “High quality” instances, i.e., the saliency map of Fig. 5 (g) appears to show only a slightly higher variation than that of Fig. 5 (f); and that saliency variation for the “Low quality” instance is more obvious, i.e., the saliency map of Fig. 5 (h) shows the highest variation amongst all instances. This is clearly reflected in the ground truth DSV measurement, as shown in Table VII that the DMSS score of the “Medium quality” instance is slightly larger than the score of the “High quality” instance, and that the DMSS score of the “Low quality” instance is much larger than the other two instances. CC and SIM metrics, on the other hand, give a higher score (i.e., larger similarity and smaller DSV) for the “Medium quality” instance than “High quality” instance, which is inconsistent with the ground truth.

Fig. 6 illustrates an additional example of ground truth DSV measured by the DMSS scores. Table VIII shows the results of DSV measured by EMD, SIM and CC. As can be seen from Fig. 6, DMSS scores increase as distortions increase, meaning when image quality degrades, its saliency becomes more dissimilar to the reference saliency (of the pristine image). So, for a good metric, we should expect that the values of EMD (i.e., measuring dissimilarity) monotonically increase as DMSS values (i.e., measuring dissimilarity) increase; and that the values of SIM and CC (i.e., measuring similarity) monotonically decrease as DMSS values increase. However, EMD, SIM and CC all fail the instance of ”Medium quality” (i.e., they all give the largest score), as show in Table VIII.

In summary, these distribution-based metrics, when applied for DSV, are inconsistent with ”ground truth” DMSS scores, as shown in Table VI, VII and VIII. The analyses suggest that KL, EMD, SIM and CC metrics are not suitable candidates for DSV quantification. This is mainly because these metrics emphasize on the distribution of saliency intensity values but ignore local structural patterns in measuring saliency similarity. However, the measured DSV between the reference and distorted saliency maps is sensitive to the local structural changes in saliency patterns [26]. This may explain why these distribution-based metrics give unsatisfactory results, as they have limitations in dealing with local structural changes in saliency.

Existing similarity metrics mainly aim to use the linear accumulation approach to construct a pixel-based model to evaluate saliency similarity. These metrics, therefore, do not adequately reflect (both global and local) structural variation in saliency. This is the reason why these metrics show a poor correlation with subjective DSV benchmark. In light of the above analyses of the limitations of existing algorithms, we will now design a new algorithm which can effectively measure the distortion-induced saliency variation in natural images.



(a) Pre-training stage

(b) Fine-tuning stage

Fig. 7. The framework of the proposed ST-DSV method. Note that C indicates the convolution layer, and P indicates the pooling layer, and FC indicates the fully connected layer.

III. The proposed ST-DSV method

We propose a deep neural network based on a coarse-grained to fine-grained saliency similarity transformation to evaluate DSV (ST-DSV). Deep learning is a powerful technique to solve complicated problems, however, it requires large amounts of labeled data. For our problem of DSV, the current benchmark dataset [24] is limited in its size. It should be noted that the ground truth label of DSV (i.e., DMSS) is derived from a fully-controlled perception experiment (rather than simple image annotation); and acquiring a large DSV dataset is non-trivial (e.g., the size of dataset is increased often at the expense of the reliability of the perceptual label due to human limits [25],[46]. To leverage deep learning with limited data, we take the approach of using data augmentation in conjunction with transfer learning. More specifically, we leverage domain-relevant pre-trained models and then restructure and fine-tune them for the DSV problem. Fig. 7 shows the framework of the proposed ST-DSV method. First, to augment saliency data, we use a pairwise labeling strategy to construct an auxiliary domain, which represents a task highly related to the target task of quantifying DSV. Then, a Siamese network [9], [11] with twin tailored VGG is built to approximately discriminate the level of similarity between the distorted and reference saliency maps. Finally, a branch of the trained Siamese network (with tailored-VGG) is used for fine-tuning with the DSV benchmark database to predict the saliency similarity scores.

A. Pairwise labeling strategy

As mentioned above, transfer learning is exploited to build a DNN model by avoiding the small dataset problem. This approach is plausible for example the large-scale ImageNet database can be first used to learn a classification task [45], and the shared features of the deep network can then be transferred to learn a new but related target task using a small database.

For our DSV target task, the learning goal is to measure the saliency similarity score, which can intuitively show the degree of saliency shift between the distorted saliency map and the reference saliency map. Due to the essential difference between the DSV measurement task and the image classification task, direct transfer learning is difficult to maximize the utilization of shared features. Therefore, we propose a pairwise labeling strategy to construct an auxiliary domain, which can help bridge tasks and gradually approach our target task (e.g., DMSS prediction).

The pairwise labeling strategy aims to define a “similarity level” label to represent the relative level of similarity when comparing different saliency maps. Since there are four different forms of saliency maps in the DSV benchmark database, including three forms of distorted images (i.e., High-quality, Medium-quality and Low-quality) and one form of reference image, we design four levels of labeling (i.e., 0, 1, 2, 3) based on the ”similarity level” between the pairwise saliency maps. In this case, the similarity level is regarded as the degree of saliency shift to roughly quantify DSV. Fig. 8 shows the four levels of labeling for the pairwise saliency maps. The label 0 means the “perfect” similarity, which is obtained by comparing the two same reference saliency maps. Note that it is the best label among the four levels of labeling. The label 1 means the “good” similarity between the High-quality saliency map and the reference. It indicates that the similarity is lower than that of level 0. Similarly, the label 2 is used to represent the “poor” similarity between the Medium-quality saliency map and the reference. The similarity level is lower than that of label 1. The label 3 is used to represent the “bad” similarity between the Low-quality saliency map and the reference, which indicates that the similarity level is the lowest. Once the similarity level of saliency maps is defined, transfer learning is in place to transfer information from previously learned image classification task to the auxiliary domain of learning a saliency map classification task, and consequently to the target task of DMSS prediction. The auxiliary domain acts as an intermediate bridge to enhance the sharing of features, progressively. Thus, the relevance of multiple tasks is significantly enhanced.

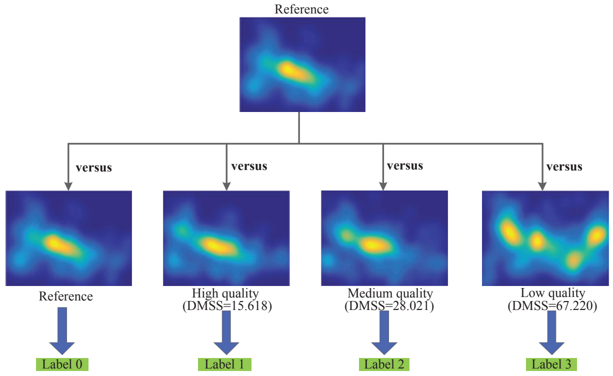

Reference
versus
versus
versus
versus
Reference
High quality (DMSS=15.618)
Medium quality (DMSS=28.021)
Low quality (DMSS=67.220)
Label 0
Label 1
Label 2
Label 3

Fig. 8. The four levels of labeling for the pairwise saliency maps.

B. Preprocessing

In the DSV benchmark database, each saliency map represents a single-channel (i.e., gray-scale) image. However, the input of a DNN should be a three-channel (i.e., R, G and B) image. We need to perform data preprocessing to generate a suitable format of input for our ST-DSV method. The idea is to mimic the way DSV is subjectively measured, reflecting the situation of comparing the distorted saliency map against the reference. Fig. 9 shows the data preprocessing and label assignment, where the input format of the DNN is

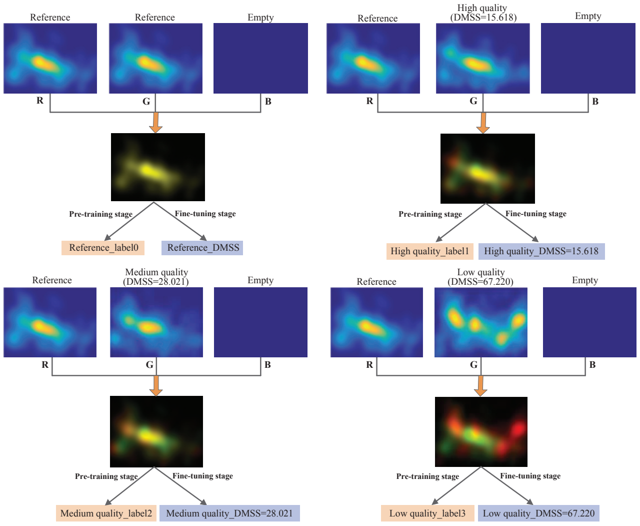


Fig. 9. The “pseudo color” image and label assignment in the ST-DSV method. (a) The reference saliency map with label 0 and DMSS=0. (b)The High-quality saliency map with label 1 and DMSS=15.618. (c) The Medium-quality saliency map with label 2 and DMSS=28.021.(d) The Low-quality saliency map with label 3 and DMSS=67.220.

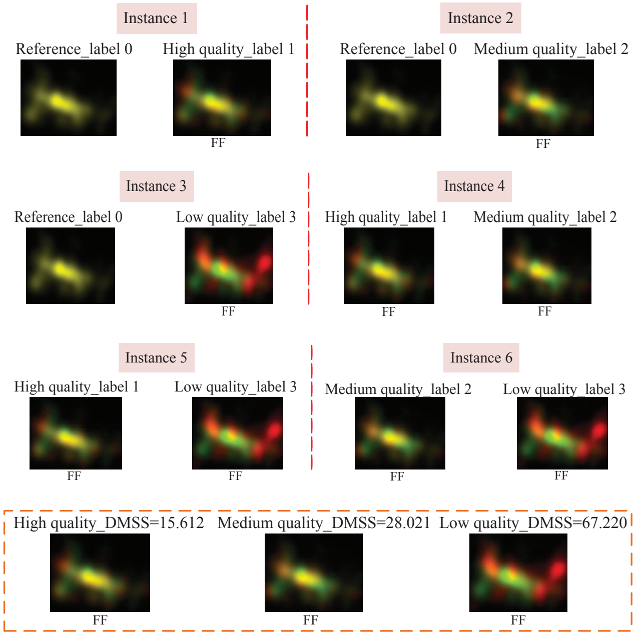


Fig. 10. The input instances of saliency maps of image named “Plane” with FF distortion in the pre-training and fine-tuning stages. (a) The six possible input instances of the Siamese network in the pre-training stage. (b) The input of the tailored VGG in the fine-tuning stage.

shown both for the pre-training stage and fine-tuning stage. We convert the single-channel saliency map into a three-channel “pseudo color” image to meet the required format of the DNN input. Because the subjective DSV (e.g., DMSS) is rendered by comparing the similarity between the distorted saliency map and the reference, the “pseudo color” image is contructed as follows: R component represents the reference, G component represents the distorted saliency map, and B component represents an “empty” map with all values set to be 0. By doing so, in the pre-training stage, the input label of the DNN is the similarity level (label 0, label 1, label 2 or label3), while in the fine-tuning stage, the input label of the DNN becomes DMSS, which represents a regression score.

Once the “pseudo color” images and the corresponding labels are constructed, the inputs are clearly defined for the pre-training and fine-tuning stages, respectively. Note that the input represents a pairwise “pseudo color” images with pairwise label levels for the pre-training stage and the input represents a single “pseudo color” image with DMSS for the fine-tuning stage. In order to clearly illustrate the DNN input, we take the saliency maps of the image named “Plane” and their labels as an example to show possible instances of the DNN input. Fig. 10 (a) shows six possible inputs for the Siamese network in the pre-training stage. Fig. 10 (b) shows the input of the tailored VGG (a branch of trained Siamese network) in the fine-tuning stage.

C. The network architecture of ST-DSV

We first train a Siamese network to rank saliency maps in terms of similarity levels. Data augmentation is achieved as the paired saliency maps with four levels actually expand the training samples by 9 times as to the original data. As shown in Table IX, with the same available saliency data, the training samples in the pre-training stage are 9 times the training samples in the fine-tuning stage. Note that DSM indicates the “distorted saliency map” and RSM indicates the ”reference saliency map”. More specifically, the input of the pre-training stage is a pair of saliency maps and their corresponding ”similarity level” labels. This stage represents a classification task of different input options. The input of the fine-tuning stage is a distorted saliency map and its corresponding ”DMSS score” label. This stage represents a regression task that follows a one-on-one relationship. Then, we use fine-tuning to transfer the coarse-grained similarity level represented in the trained Siamese network to the fine-grained similarity score (i.e., DMSS).

In the pre-training stage (Fig. 7(a)), the Siamese network [47] consists of twin tailored VGG network branches. The twin tailored VGG is used with the aim of sharing weights of previously learned models. More specifically, the architecture of the tailored VGG network is constructed by removing the softmax layer of the original VGG network [50] (i.e., used for classification task) and replacing it with an output FC layer. By doing so, the parameters/weights of the pre-trained VGG network can be transferred to tailored VGG network. An example of the input instances of the Siamese network is illustrated in Fig. 10(a), where pairs of “pseudo color” images and level labels represent six input instances for the saliency maps originated a reference image and images distorted with a specific distortion type. Once pairs of saliency maps and level labels are fed into the Siamese network, the network can learn to rank the similarity level of the saliency maps. The tailored VGG is to learn the properties of DSV by comparing the pairwise saliency maps at different similarity levels. Since DMSS essentially reflects the changes of complex structural information, the advantage is that the DSV features representing the higher-level semantics of structural information could be learned. At the same time, the Siamese network with the twin tailored VGG is designed to compare different similarity levels of saliency maps to capture the relative changes of saliency patterns. It makes matched maps in a pair are pulled closer and unmatched maps are pushed further away.

TABLE IX
Illustration of input in pre-training (PT) and fine-tuning (FT) stages.

Option	Input(PT)	Input(FT)
1	RSM with label 0 RSM with label 0	DMS with DMSS
2	RSM with label 0 DSM with label 1	
3	RSM with label 0 DSM with label 2	
4	RSM with label 0 DSM with label 3	
5	DSM with label 1 DSM with label 1	
6	DSM with label 1 DSM with label 2	
7	DSM with label 1 DSM with label 3	
8	DSM with label 2 DSM with label 2	
9	DSM with label 2 DSM with label 3	

In the fine-tuning stage (Fig. 7(b)), the “pseudo color” images with the DMSS labels illustrated in Fig. 10(b) are fed into a learned tailored VGG branch to predict DMSS. Since the tailored VGG branch is able to discriminate the saliency similarity by using the coarse-grained levels, the process of fine-tuning, therefore, transfers the coarse-grained similarity levels to the fine-grained similarity scores to learn the regression task of DSV measurement.

D. The loss function of ST-DSV

When we pre-train the Siamese network, the loss function is adopted to discriminate the relative similarity levels. We consider the difference of deep features between the pairwise saliency maps under different similarity levels and design a trade-off threshold to control the optimal level of the Siamese network output. The loss function L_1 is defined as:

$$L_1(S^m;S^n;\theta)=max(0,f(S^m;\theta)-f(S^n;\theta)+\zeta) \tag{1}$$

where S^m,S^n denote the two paired saliency maps (i.e., originated from two distorted images, one distorted and one reference images, or two reference images). $f(S^m;\theta)$,$f(S^n;\theta)$ respectively denote the output feature representation in the last layer of the two branches of the Siamese network. In our Siamese network, the output of the final layer is a single scalar, which aims to be indicative of DSV similarity. θ represents the parameters of Siamese network. ζ denotes the margin.

Here we assume, without loss of generality, that the similarity level ranking of S^m is higher than S^n. Since our goal is to rank similarity level, the gradient of the loss L_1 (Equation 1) is given by:

$$\begin{cases} \nabla_\theta L = 0 & case1 \\ \nabla_\theta L = \nabla_\theta f(S^m;\theta) - \nabla_\theta f(S^n;\theta) & case2 \end{cases} \tag{2}$$

Note that case 1 means $f(S^m;\theta) - f(S^n;\theta) + \zeta \leq 0$ and case 2 means otherwise condition. When the outcome of the Siamese network is in accordance with the pre-defined ranking, the gradient should be zero. Otherwise, the network needs to adjust parameter θ to meet the requirement that the similarity level ranking of S^m is higher than S^n. Given the gradient of the loss L_1 with respect to model parameters θ, we can train the Siamese network to rank similarity level. The margin parameter ζ is set to 10.

After training the Siamese network, we extract the single branch for fine-tuning and use the Euclidean distance as the loss function L_2

$$L_2(y_i;\hat{y}_i) = \frac{1}{M}\sum_{i=1}^{M}(y_i - \hat{y}_i)^2 \tag{3}$$

where M is the number of saliency maps in mini-batch. y_i is the ground truth score of DSV (i.e., DMSS) of the i-th saliency map. $\hat{y}_i$ is the predicted score of DSV.

TABLE X
The sizes of saliency maps in the DSV benchmark database.

Size(pixel)	Number of saliency maps
768 × 512	192
632 × 505	16
634 × 438	16
627 × 482	16
480 × 720	48

E. The training strategy of ST-DSV

Since the sizes of the saliency maps in the DSV benchmark database are not the same, as shown in Table X, image cropping is applied. In implementation of the DNN, we randomly crop the image path of 430×430 pixels to generated a "sub-image", and this process is done repeatedly by the DNN to cover the entire image content. If the cropped size is too small, it is likely that the patch dose not contain representative structure of the DSV. However, if the cropped patch is too large, the input saliency map needs zero padding by adding a large number of zero values, This may lead to the change of saliency patterns e.g., the proportion of salient and non-salient areas might be changed.

We use the Caffe [48] framework and train it using the mini-batch Stochastic Gradient Descent [49] with an initial learning rate of 1e-4 for efficient Siamese network training and 1e-6 for fine-tuning. Training rates are decreased by a factor of 0.1 every 10k iterations for a total of 50K iterations. During training we sample a single "sub-image" from each training saliency map per epoch. The trade-off threshold is set to be 10.

IV. Experiments and results

A. Database and evaluation metrics

The performance of our proposed ST-DSV method is validated against the DSV benchmark database [24], [26]. The three commonly used metrics, SROCC, PLCC, KROCC as already mentioned in Section II, are used for the performance evaluation. These metrics measure the correlation between a set of predicted scores of DSV and a set of human subjective scores of DSV. The subjective scores are DMSS values, which represent the ground truth DSV measurement. A DMSS score measures the perceived difference between the distorted saliency map and the reference saliency map. For the three performance metrics, a value close to 1 indicates high performance of an objective DSV measure.

B. Performance on the DSV benchmark database

We compare the performance of our proposed ST-DSV method to the state-of-the-art saliency similarity methods, including AUC-Borij [29], s-AUC [30], NSS [31], CC [32], SIM [32], IG [33], KL [34], EMD [35]. In addition, we also compare the performance of our method to some ”structural similarity/visual fidelity” methods, including the traditional methods (SSIM [40], MS-SSIM [41], VIF [42] and FSIM [43]) and the deep learning methods (DIQaM-FR [44] and GraphIQA [54]).

Because our ST-DSV method is based on deep learning, we divided the DSV benchmark database into two sets, including a training set and a test set. In our experiments, 80% of the distorted saliency maps are used as the training set and the remaining 20% of the distorted saliency map are used as the test set. Also, the training set consists of 70% samples for training and 10% samples for validation. In terms of the number of samples for fine-tuning the model, we used the entire DSV benchmark which contains 270 distorted saliency maps (note these saliency maps are originated from 18 references, i.e., each reference corresponds to 15 distorted maps). For each run (i.e., eight runs in total), we divided (randomly as per references) the dataset into 80%-training (including validation) and 20%-test sets. Effectively, this data split gives a training set of 210 maps (i.e., 195 maps (originated from 13 references) for training and 15 maps (originated from 1 reference) for validation) and a test set of 60 maps (i.e., originated from 4 references). Because the random split takes place based on the references, the generated training, validation, and test sets do not overlap at all in terms of content, which ensures a rigorous training strategy. This process is repeated eight times

TABLE XI
The performance of different saliency similarity and structural similarity/visual fidelity metrics on the DSV benchmark database.

Metric	AUC-Borji	s-AUC	NSS	IG	CC	SIM	KL	EMD	SSIM	MS-SSIM	VIF	FSIMc	DIQaM-FR	GraphIQA	ST-DSV
SROCC	0.527	0.586	0.657	0.336	0.736	0.716	0.343	0.569	0.470	0.553	0.533	0.521	0.697	0.735	**0.820**
PLCC	0.535	0.586	0.672	0.343	0.766	0.748	0.339	0.610	0.501	0.560	0.534	0.541	0.712	0.750	**0.836**
KROCC	0.384	0.418	0.492	0.238	0.562	0.538	0.247	0.412	0.337	0.408	0.375	0.390	0.501	0.542	**0.632**

TABLE XII
The performance of saliency similarity and structural similarity/visual fidelity metrics for five different distortion types on the DSV benchmark database.

Metric	FF			GB			JP2K			JPEG			WN		
	SROCC	PLCC	KROCC	SROCC	PLCC	KROCC	SROCC	PLCC	KROCC	SROCC	PLCC	KROCC	SROCC	PLCC	KROCC
AUC-Borji	0.512	0.588	0.356	0.620	0.649	0.458	0.600	0.659	0.504	0.520	0.546	0.398	0.433	0.416	0.331
s-AUC	0.583	0.614	0.451	0.596	0.653	0.428	0.589	0.610	0.462	0.549	0.575	0.410	0.542	0.504	0.417
NSS	0.664	0.667	0.542	0.618	0.723	0.477	0.611	0.669	0.462	0.663	0.702	0.519	0.554	0.515	0.446
IG	0.121	0.319	0.080	0.295	0.368	0.205	0.422	0.514	0.333	0.297	0.350	0.220	0.306	0.373	0.238
CC	0.767	0.797	0.655	0.608	0.697	0.485	0.710	0.691	0.568	0.740	0.774	0.587	0.704	0.720	0.585
SIM	0.686	0.710	0.557	0.566	0.680	0.428	0.734	0.761	0.591	0.729	0.770	0.576	0.731	0.755	0.590
KL	0.210	0.494	0.163	0.309	0.369	0.254	0.458	0.392	0.352	0.410	0.364	0.326	0.243	0.344	0.184
EMD	0.575	0.609	0.443	0.518	0.593	0.394	0.637	0.635	0.500	0.630	0.684	0.473	0.488	0.533	0.378
SSIM	0.447	0.544	0.355	0.459	0.399	0.349	0.635	0.640	0.502	0.473	0.511	0.381	0.606	0.648	0.476
MS-SSIM	0.554	0.677	0.450	0.581	0.631	0.423	0.535	0.549	0.407	0.590	0.594	0.481	0.575	0.580	0.420
VIF	0.568	0.632	0.407	0.457	0.424	0.326	0.440	0.543	0.333	0.495	0.506	0.381	0.671	0.654	0.515
FSIMc	0.512	0.666	0.411	0.567	0.575	0.436	0.536	0.540	0.416	0.506	0.546	0.411	0.594	0.651	0.463
DIQaM-FR	0.632	0.650	0.594	0.615	0.631	0.608	0.584	0.592	0.577	0.567	0.585	0.541	0.629	0.647	0.612
GraphIQA	0.642	0.633	0.621	0.635	0.657	0.638	0.604	0.612	0.579	0.580	0.587	0.560	0.641	0.655	0.628
ST-DSV	**0.851**	**0.881**	**0.712**	**0.790**	**0.861**	**0.614**	**0.864**	**0.824**	**0.735**	**0.781**	**0.833**	**0.617**	**0.767**	**0.808**	**0.614**

TABLE XIII
The performance of saliency similarity and structural similarity/visual fidelity metrics for three different distortion levels on the DSV benchmark database.

Metric	Low Level			Medium Level			High Level		
	SROCC	PLCC	KROCC	SROCC	PLCC	KROCC	SROCC	PLCC	KROCC
AUC-Borji	0.487	0.534	0.350	0.580	0.559	0.429	0.490	0.503	0.375
s-AUC	0.528	0.557	0.372	0.614	0.622	0.462	0.542	0.528	0.376
NSS	0.666	0.695	0.492	0.667	0.674	0.504	0.590	0.651	0.443
IG	0.407	0.466	0.312	0.313	0.332	0.222	0.066	0.252	0.051
CC	0.671	0.697	0.507	0.753	0.748	0.591	0.700	0.791	0.546
SIM	0.681	0.680	0.507	0.711	0.698	0.557	0.637	0.730	0.482
KL	0.091	0.244	0.072	0.241	0.244	0.186	0.338	0.379	0.241
EMD	0.415	0.395	0.291	0.539	0.531	0.407	0.540	0.624	0.391
SSIM	0.422	0.436	0.307	0.453	0.475	0.342	0.300	0.411	0.218
MS-SSIM	0.475	0.529	0.356	0.562	0.575	0.413	0.441	0.481	0.354
VIF	0.552	0.543	0.400	0.524	0.546	0.397	0.349	0.363	0.228
FSIMc	0.483	0.490	0.372	0.540	0.511	0.409	0.430	0.492	0.326
DIQaM-FR	0.655	0.658	0.540	0.625	0.640	0.503	0.609	0.614	0.522
GraphIQA	0.661	0.669	0.581	0.632	0.654	0.551	0.622	0.620	0.537
ST-DSV	**0.786**	**0.828**	**0.578**	**0.783**	**0.775**	**0.604**	**0.772**	**0.830**	**0.620**

to eliminate the performance bias. For each run, the training and test sets are randomly selected as described above. The average values of the obtained SROCC, PLCC and KROCC are reported as the final results. To have a fair comparison among different metrics, each of the other metrics is applied to the test set for eight times (as the process mentioned above) and the average performance is calculated.

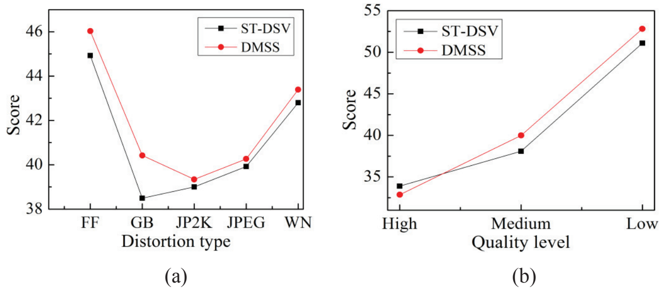
Table XI shows the performance of different saliency and image similarity/visual fidelity metrics. The best performance is shown in bold. Compared with "saliency similarity" metrics (AUC-Borji, s-AUC, NSS, IG, CC, SIM, KL and EMD), it can be seen that the proposed ST-DSV metric outperforms other metrics. As already extensively discussed in Section II, existing saliency similarity metrics are formulated on a pixel-by-pixel basis, which limits their ability of capturing complex structural variation in saliency patterns. The limitations of these pixel-based metrics are overcome by our ST-DSV method that adopts deep learning, making use of the entire visual content that composes a saliency map. The model can thereby extract strongly task-relevant deep features to represent the complex higher-level properties of saliency patterns.

Furthermore, compared with ”structural similarity/visual fidelity” methods (SSIM, MS-SSIM, VIF, FSIMc, DIQaM-FR and GraphIQA), we observe that our propose ST-DSV method is superior. This might be attributed to the fact that these existing methods are specifically designed to quantify structural similarity or visual fidelity between two natural images. The computed features (by SSIM, MS-SSIM, VIF, FSIMc) or learned features (by DIQaM-FR and GraphIQA) do not necessarily describe the variation of saliency maps, which differ from normal ”natural” images. In this respect, our ST-DSV method uses a task-specific coarse-grained and fine-grained saliency similarity transformation design, which can effectively simulate the characteristics of DSV as rendered by subjective assessment.

C. Performance on individual distortion types and levels

In Table XII, we evaluate the performance of the proposed ST-DSV method and other competing metrics on individual distortion types. Also, we compare the performance of these metrics on different distortion levels, as shown in Table XIII. The ST-DSV model was trained with all distortion types or levels contained in the training set (80%) and tested on the distortion types or levels (unseen in the training set) on the test set (20%). The best performance among all cases is shown in bold. As shown in Table XII and Table XIII, for each distortion type or level, the ST-DSV method gives the best performance. This suggests our metric is rather robust. The superior performance of our proposed method might be attributed to the fact that it captures the dependency between DSV and changes of distortion types/levels.

To verify this hypothesis, we plot the ground truth DSV (i.e., DMSS scores) versus our ST-DSV method (i.e., predicted scores) for different distortion types and again for different distortion levels, as shown in Fig. 11. It can be seen that the ST-DSV method shows a similar trend to the ground truth DSV in terms of how the scores differ in accordance with different distortion types or levels. For example, as shown in Fig. 11(a), DSV is largest for the FF distortion (i.e., FF causes the largest saliency variation), which is well predicted by the ST-DSV method (i.e., the predicted saliency variation is largest for FF amongst all distortion types). Also, as shown in Fig. 11(b), DSV monotonically increases as the image quality decrease (i.e., higher distortion level causes larger saliency variation); and this trend of saliency variation is accurately predicted by our ST-DSV method (i.e., the lower the image quality, the larger the predicted saliency variation).



(a)

(b)

Fig. 11. Performance (i.e., SROCC) of ST-DSV for different distortion types and levels on the DSV benchmark database. (a) ST-DSV performance for different distortion types. (b) ST-DSV performance for different distortion levels.

D. The ablation experiments

We conduct three types of ablation experiments to verify the advantageous properties of our proposed ST-DSV method: (1) ablation experiment to verify the superiority of the baseline network architecture, (2) ablation experiment to verify the superiority of the Siamese network core, and (3) ablation experiment to verify the superiority of the classification strategy. The results are listed in Table XIV, Table XVand Table XVI.

In terms of different baseline network architectures, we compare VGG, AlexNet [50], ResNet18, ResNet34 and ResNet50 [51]. These pre-trained models are each directly used for fine-tuning with the DSV benchmark database. Note in this paper, we mainly focus on the design of ST-DSV method by using DNN architectures with simple stack convolution layers; thereby VGG is pre-selected as a baseline architecture. Including complex DNNs with external enhancement modules (e.g., a residual block) may produce better results but at the expense of model’s simplicity, therefore, is considered outside the scope of this paper. As can be seen from Table XIV, AlexNet, ResNet34 and ResNet50 give relatively low performance because AlexNet’s architecture is too shallow, which causes poor learning ability of DMSS, and architecture of ResNet34/ResNet50 is too deep, which leads to overfitting phenomenon. VGG and ResNet18 give relatively high and comparable performance. Since VGG exhibits a much simpler architecture, we select the VGG framework as the baseline in the design of ST-DSV network.

In terms of different Siamese network core components, we compare our final ST-DSV design, ST-DSV using AlexNet as the Siamese network core (i.e., AlexNet-DSV), and ST-DSV using ResNet50 as the Siamese network core (i.e., ResNet50-DSV). This comparison reveals how different Siamese network core components, when built in to our overall metric design, can affect the metric performance. As can be seen from Table XV, ST-DSV is superior to AlexNet-DSV and ResNet50-DSV, suggesting the proposed ST-DSV network design is rather advantageous in maximizing the learning ability of saliency similarity features.

In terms of different classification strategies, we compare VGG, ST-DSV and VGG-C. The VGG method means that the VGG network is directly used for fine-tuning with the DSV benchmark database. The VGG-C method means that the VGG network is first used to classify 4 saliency similarity levels and then the last fully-connected layer of the network with 4 classifications (VGG-C) is modified to a one-dimensional output for fine-tuning with the DSV benchmark database. It can be seen from Table XVI that the classification strategy (i.e., pair-wise labeling classification) of our ST-DSV method is superior to both VGG (no classification) and VGG-C (simple classification) methods. This indicates the importance of developing sophisticated DSV-specific classification methods.

TABLE XIV
The performance of different DNN baseline network architectures for DSV prediction.

Method	SROCC	PLCC	KROCC
VGG(Baseline)	0.734	0.745	0.535
AlexNet	0.708	0.715	0.521
ResNet18	0.740	0.747	0.535
ResNet34	0.722	0.730	0.528
ResNet50	0.702	0.744	0.501

TABLE XV
The performance ST-DSV model using different Siamese network core components.

Method	SROCC	PLCC	KROCC
AlexNet-DSV	0.725	0.737	0.544
ResNet50-DSV	0.713	0.750	0.516
ST-DSV	0.802	0.836	0.632

V. Discussion

A. The rationality of coarse-grained label

The proposed coarse-grained ”similarity level” label is a unique feature of our proposed method, so we give a further discussion on this point. First, the labeling method has been designed to capture DSV in a coarse-grained manner, as already explained above. Second, to verify the rationality, we conduct a new experiment by shuffling similarity-level labels so that each distorted saliency map could be assigned by any available label and all combinations of label assignment are covered in the experiment. For each instance of label assignment, we calculate the network performance as shown in Table XVII. By doing this, the results indicate the label assignment of the proposed ST-DSV is reliable and gives the best network performance.

B. The overflow and underflow

To verify whether the network has overflow/underflow issues we conduct further experiments, as also suggested by [52].

TABLE XVI
The performance of different classification strategies for DSV prediction.

Method	SROCC	PLCC	KROCC
VGG	0.734	0.745	0.535
VGG-C	0.750	0.796	0.550
ST-DSV	**0.820**	**0.836**	**0.632**

TABLE XVII
The model performance (i.e., SROCC) using different coarse-grained ”similarity level” label assignment instances. Note that PES means ”perfect” similarity, GOS means ”good” similarity, POS means ”poor” similarity and BAS means ”bad” similarity.

Method	PES	GOS	POS	BAS	SROCC
1	Label 0	Label 1	Label 3	Label 2	0.712
2	Label 0	Label 2	Label 1	Label 3	0.771
3	Label 0	Label 2	Label 3	Label 1	0.620
4	Label 0	Label 3	Label 2	Label 1	0.605
5	Label 0	Label 3	Label 1	Label 2	0.609
ST-DSV	**Label 0**	**Label 1**	**Label 2**	**Label 3**	**0.820**

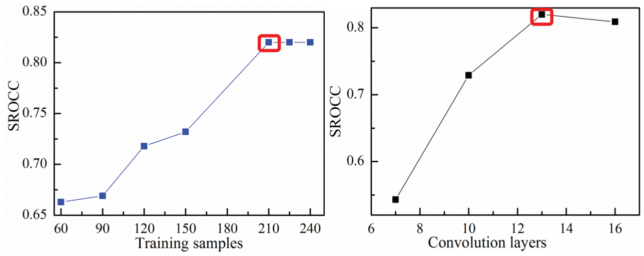
In terms of preventing overflow/underflow from the standpoint of training sample size, we compare the performance of the proposed ST-DSV method using different sizes of training samples (i.e., from 60 to 240 saliency maps from the DSV benchmark database). Note the theoretical maximum size of training data is 270 (0 for testing data), we decided not to go over the size of 240 for training data, otherwise the test data is too small. Fig. 12(a) shows training sample data size vs prediction performance (i.e., SROCC). The training sample size selected in our model is highlighted in red color. As can be seen in Fig. 12(a), when the training sample size increases from 60 through to 210, the prediction performance of the model increases, which indicates the model has no underflow problem. When the training sample size increases from 210 through to 240, the prediction performance remains unchanged, which indicates that the network is not subject to overflow. So, we selected the training sample size of 210 so the standard train-test split (80%-20%) is maintained.

In terms of preventing overflow/underflow from the standpoint of number of network layers, we construct a DNN in the proposed ST-DSV method using different numbers of convolution layers (C) (i.e., from 7 to 16 C contained in the VGG network). Note 7 means the C of third group of VGG [8], 13 means the C of fifth group of VGG, 16 means that we add on the same C architecture of fifth group. We decided not to go over 17 layers, otherwise, the network is too complex. Fig. 12(b) shows the number of C vs prediction performance (i.e., SROCC). The number of convolution layers selected for our final model is highlighted in red color. As can be seen in Fig. 12(b), when the number of layers C increases from 7 (the third group of VGG) to 13 (the fifth group of VGG), the prediction performance increases, which suggests that the network is not prone to underflow. When C increases from 13 to 16, the prediction performance decreases, which indicates that going over 13 layers would potentially lead to overflow. Therefore, we selected 13 layers for our network.

C. The k-fold cross-validation

In training a DNN model, k-fold cross-validation is often used to tune hyper parameters and prevent from overfitting [53]. In k-fold cross-validation, data is partitioned into k subsets, the model is trained on k-1 folds iteratively while using the remaining fold as the test set. We conduct 5-fold cross-validation for our proposed ST-DSV model, where the total of 270 saliency maps (originated from 18 references) are partitioned into 5 non-overlapped subsets, i.e., 3 subsets of 60 maps each (originated from 4 references) and 2 subsets of 45 maps each (originated from 3 references). By doing so, the fairness of the procedure is ensured as there is no data leakage. The results are shown in Table XVIII. It can be seen that the model's performance of using k-fold cross-validation and our training strategy is similar.

It is worth nothing that our training strategy as the way it is specifically designed and detailed in Section IV.B is essentially similar to a 5-fold cross-validation. In our design, the train-validation-test split ratio is 7:1:2 and these subsets per run do not overlap in content at all (so no data leakage). Also, the data splitting is randomly repeated eight times to generate an average model output so to eliminate the performance bias. In this respect, we expect our training strategy and k-fold cross-validation should give similar results.



(a) (b)

Fig. 12. Model's performance (i.e., SROCC) using different sizes of training samples and different numbers of convolution layers. (a) Model's performance using different sizes of training samples. (b)Model's performance using different numbers of convolution layers.

TABLE XVIII
Performance (i.e., SROCC) of the proposed model using k-fold cross-validation versus our training strategy

Method	SROCC
K-fold (K=5) cross-validation for ST-DSV	0.816
Our training strategy for ST-DSV	0.820

VI. Conclusion

In this paper, based on the ground truth benchmark of the distortion-induced saliency variation (DSV), we have found that the use of existing mathematical algorithms for measuring DSV is rather limited. These algorithms fail in quantifying the degree of difference/similarity between the saliency of a reference image and that of its distorted image. To achieve a reliable metric for DSV, we have proposed a new algorithm based on a deep neural network. Our algorithm uses a coarse-grained to fine-grained saliency similarity transformation approach. Experiments demonstrate that the proposed algorithm can accurately predict ground truth DSV (i.e., DMSS scores). Further, the research will investigate the improvement of current metric (e.g., by exploiting state-of-the-art deep learning techniques including generative adversarial network (GAN) based methods) and the application of DSV metric for advanced image quality assesment algorithms.

References

[1] F. Li, S. Fu, Z. Li and X. Qian, "A cost-constrained video quality satisfaction study on mobile device," *IEEE Trans. Multimedia*, vol. 20, no. 5, pp. 1154–1168, 2018.

[2] Q. Jiang, W. zhou, X. Chai, G. Yue, F. Shao and Z. Chen, "A full-reference stereoscopic image quality measurement via hierarchical deep feature degradation fusion," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 12, pp. 9784-9796, 2020.

[3] Q. Jiang, W. Gao, S. Wang, G. Yue, F. Shao, Y. Ho and S. Kwong, "Blind image quality measurement by exploiting high-order statistics with deep dictionary encoding network," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 10, pp. 7398-7410, 2020.

[4] A. Angelis, A. Moschitta, F. Russo and P. Carbone, "A vector approach for image quality assessment and some metrological considerations," *IEEE Transactions on Instrumentation and Measurement*, vol. 58, no. 1, pp. 14-25, 2009.

[5] G. Yue, C. Hou, T. Zhou and X. Zhang, "Effective and efficient blind quality evaluator for contrast distorted images," *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 8, pp. 2733-2741, 2019.

[6] H. Sellahewa and S. Jassim, "Image-quality-based adaptive face recognition," *IEEE Transactions on Instrumentation and Measurement*, vol. 59, no. 4, pp. 805-813, 2010.

[7] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 36-47, Jan. 2020.

[8] X. Yang, F. Li and H. Liu, "Deep feature importance awareness based no-reference image quality prediction," *Neurocomputing*, vol. 401, pp. 209-223, 2020.

[9] J. Xu, W. Zhou, Z. Chen, S. Ling and P. Callet, "Binocular rivalry oriented predictive auto-encoding network for blind stereoscopic image quality measurement," *IEEE Transaction on Instrumentation and Measurement*, vol. 70, 2021. [DOI:10.1109/TIM.2020.3026443]

[10] L. Shi, W. Zhou, Z. Chen and J. Zhang, "No-reference light field image quality assessment based on spatial-angular measurement," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 4114-4128, 2020.

[11] W. Zhou, J. Xu, Q. Jiang and Z. Chen, "No-reference quality assessment for 360-degree images by analysis of multifrequency information and local-global naturalness," *IEEE Trans. Circuits Syst. Video Technol.*, 2021. [DOI: 10.1109/TCSVT.2021.3081182]

[12] W. Zhou, L, Shi, Z. Chen and J. Zhang, "Tensor oriented no-reference light field image quality assessment," *IEEE Trans. Image Process.*, vol. 29, pp. 4070-4084, 2020.

[13] Z. Cheng, M. Takeuchi and J. Katto, "A pre-saliency map based blind image quality assessment via convolutional neural networks," *in Proc. IEEE International Symposium on Multimedia*, pp. 77–82, 2017.

[14] J. Guan, S. Yi, X. Zeng, W. Cham and X. Wang, "Visual importance and distortion guided deep image quality assessment framework," *IEEE Trans. Multimedia*, vol. 19, no. 11, pp. 2505–2521, 2017.

[15] S. Yang, Q. Jiang, W. Lin and Y. Wang, "SGDNet:An end-to-end saliency-guild deep neural network for no-reference image quality assessment," *in Proc. ACM International Conference on Multimedia*, pp. 1383-1391, 2019.

[16] C. Aladine, "Convolutional neural network and saliency selection for blind image quality assessment," *in Proc. ICIP*, pp. 2835-2839, 2018.

[17] X, Yang, F. Li, and H. Liu, "A study of DNN methods for blind image quality assessment," *IEEE Access*, vol. 7, no. 1, pp. 123788-123806, 2019.

[18] X. Luo, J. Zhang and Q. Dai, "Saliency-based geometry measurement for image fusion performance," *IEEE Transactions on Instrumentation and Measurement*, vol. 61, no. 4, pp. 1130-1132, 2012.

[19] W. Zhang, R. Martin and H. Liu, "A saliency dispersion measure for improving saliency-based image quality metric," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 28, no. 6, pp. 1462-1466, 2018.

[20] K. Gu, S. Wang, H. Yang, W. Lin, G. Zhai, X. Yang and W. Zhang, "Saliency-guided quality assessment of screen content images," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1098—1110, 2017.

[21] L. Gu, F. Gao, Y. Zhang, R. Judith and C. Pan, "Bling image quality assessment via vector regression and object oriented pooling," *IEEE Trans. Multimedia*, vol. 20, no. 5, pp. 1140–1153, 2018.

[22] Y. Yang, B. Li, P. Li and Q. Liu, "A two-stage clustering based 3D visual saliency model for dynamic scenarios," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 809-820, 2019.

[23] W. Zhang, A. Borji, Z. Wang, P. Le Callet and H. Liu, "The application of visual saliency models in objective image quality assessment: a statistical evaluationt," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 6, pp. 1266-1278, 2016.

[24] W. Zhang, H. Liu, "Towards a reliable collection of eye-tracking data for image quality research: challenges solutions and applications," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2424-2437, 2017.

[25] H. Sheikh, M. Sabir, and A. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, pp. 3440-3451, 2006.

[26] L. Leveque, W. Zhang, H. Liu, "Subjective assessment of image quality induced saliency variation," *in Proc. ICIP*, pp.1024-1028, 2019.

[27] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," *in Proc. IEEE 12th Int. Conf. Comput. Vis.*, pp. 2106–2113, 2009.

[28] A. Borji, D. N. Sihite, and L. Itti, “Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study,” *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 55–69, Jan. 2012.

[29] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, “Sun: A bayesian framework for saliency using natural statistics,” *J. Vis.*, vol. 8, no. 7, pp. 1–20, 2008.

[30] R. J. Peters, A. Iyer, L. Itti, and C. Koch, “Components of bottom-up gaze allocation in natural images,” *Vis. Res.*, vol. 45, no. 18, pp. 2397–2416, 2005.

[31] M. Kummerer, T. Wallis, and M. Bethge, “Information-theoretic model comparison unifies saliency metrics,” *Proceedings of the National Academy of Sciences*, vol.112, no. 52, pp. 16054–16059, 2015.

[32] Y. Rubner, C. Tomasi, and L. J. Guibas, “The earth mover’s distance as a metric for image retrieval,” *Int. J. Comput. Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[33] O. Meur, P. Callet, and D. Barba, “Predicting visual fixations on video based on low-level visual features,” *Vis. Res.*, vol. 47, no. 19, pp. 2483–2498, 2007.

[34] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, “What do different evaluation metrics tell us about saliency models?,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 740–757, 2018.

[35] O. Pele and M. Werman, “A linear time histogram metric for improved sift matching,” *in Proc. ECCV*, pp. 495–508, 2008.

[36] N. Riche, M. Duvinage, M. Mancas, B. Gosselin and T. Dutoit, “Saliency and human fixations: state-of-the-art and study of comparison metrics,” *in Proc. ICCV*, pp. 1153-1160, 2013.

[37] H. Lin, V. Hosu and D. Saupe, “KADID-10K: a large-scale artificially distorted IQA database,” *in Proc. QoMEX*, 2019.

[38] J. Ilan, “Comparing rankings of search results on the web,” *Information processing and management*, pp. 1511-1519, 2005.

[39] “Final report from the video quality experts group on the validation of objective models of video quality assessment,” VQEG, Tech. Rep., 2000.

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

[41] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multi-scale structural similarity for image quality assessment,” *in Proc. ACSSC.*, pp. 1398–1402, 2003.

[42] H. Sheikh, and A. C. Bovik, “Image information and visual quality,” *IEEE Trans. Image Process.*, pp. 430–444, 2006.

[43] L. Zhang, L. Zhang, X. Mou, and D. Zhang, “FSIM: A feature similarity index for image quality assessment,” *IEEE Trans. Image Process.*, pp. 2378–2386, 2011.

[44] S. Bosse, D. Maniry, K. Muller, T. Wiegand and W. Samek, “Deep neural networks for no-reference and full-reference image quality assessment,” *IEEE Trans. Image Process.*, pp. 206–219, 2018.

[45] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *in Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

[46] W. Zhang and H. Liu, “Learning picture quality from visual distraction: psychophysical studies and computational models,” *Neurocomputing.*, vol. 247, pp. 183–191, 2017.

[47] L. Bertinetto, J. Valmadre, J. Henriques, A. Vedaldi and P.H. Torr, “Fully convolutional siamese networks for object tracking,” *in Proc. ECCV*, pp. 850–865, 2016.

[48] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional architecture for fast feature embedding,” *in Proc. ACM MM*, pp. 675-678, 2014.

[49] L. Kang, P. Ye, Y. Li, and D. Doermann, “Convolutional neural networks for no-reference image quality assessment,” *in Proc. CVPR*, pp. 1733–1740, 2014.

[50] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *in Proc. NIPS*, pp. 1097–1105, 2012.

[51] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *in Proc. CVPR*, pp. 770–778, 2016.

[52] J. Neumann, “Model selection and overfitting,” *Nature*, pp. 703-704, 2016.

[53] T. Wong and N. Yang, “Dependency analysis of accuracy estimates in k-fold cross validation,” *IEEE Transaction on Knowledge and Data Engineering*, pp. 2417-2427, 2017.

[54] S. Sun, T. Yu, J. Xu, J. Lin, W. Zhou and Z. Chen, “GraphIQA: Learning distortion graph representations for blind image quality assessment,” *arXiv preprint arXiv:2103.07666*, 2021.

[55] Methodology for the Subjective Assessment of the Quality of Television Pictures Jun. 2002, Recommendation ITU-R BT.500-11.