IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 8, 2017, accepted November 4, 2017, date of publication November 22, 2017,
date of current version February 14, 2018.

Digital Object Identifier 10.1109/ACCESS.2017.2776344

Deep Saliency Quality Assessment Network With Joint Metric

LIANGZHI TANG, QINGBO WU, (Member, IEEE), WEI LI, AND YINAN LIU
School of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China

This work was supported by the National Natural Science Foundation of China under Grant 61525102, Grant 61601102, and Grant 61502084.

ABSTRACT Saliency detection aims to find the most conspicuous regions in an image, which highly catches the users' attention. High-quality saliency map plays an important role in boosting many other computer vision tasks, such as object detection and segmentation. To assess a saliency map's quality, the only way is to utilize a full reference metric, i.e., compute it with the ground-truth reference map. However, in the real-world applications, the ground-truth reference map for the saliency region is unavailable, which brings urgent demands for developing no reference saliency quality metric. In this paper, we propose a deep saliency quality assessment network (DSQAN) to predict the saliency quality scores directly from saliency maps. Furthermore, a joint metric is developed to better depict the quality of a saliency map. The proposed joint metric can not only lead better quality prediction accuracy, but also bring out more robust results. As a direct application of the proposed DSQAN, the predicted saliency quality scores are first utilized to choose the optimal saliency map from a set of saliency map candidates. The experimental results on the MSRA10K data set demonstrate that our proposed method could precisely predict the saliency quality. Particularly, when the DSQAN is applied to recommend optimal saliency map to feed an object segmentation algorithm from multiple candidates, its segmentation accuracy significantly outperforms the results outputted from the best saliency detection algorithms.

INDEX TERMS Saliency quality assessment, deep convolutional neural network, saliency quality prediction, regression neural network, joint metric.

I. INTRODUCTION

Salient object detection has drawn lots of attentions in recent years, which aims to extract the salient object from an image. Many computer vision problems can benefit from this, such as image compressing [1], picture collage [2], movie summarization [3], face segmentation [4], etc. The common procedure of these applications is to generate saliency map in early stages and use it to locate the foreground objects or regions of interest. Generally, the subsequent procedures are dominated by the performance of saliency detection algorithms, where poor results of saliency detection will inevitably lower the ceiling performance of such saliency-based applications.

Although the existing saliency detection algorithms have achieved impressive performances in terms of a holistic metric, there exist significant differences when dealing with images across diverse visual contents. Figure 1 shows an example of diverse saliency detection results when four saliency detection DSR [5], MC [6], RBD [7] and GR [8], are applied to different input images. The four input images are chosen from MSRA10K [9], where the aforementioned algorithms achieve state-of-the-art performances range from 95.4 to 95.5 in terms of AUC score. In spite of the similar objective performance, these algorithms behave quite different when facing different input images. We denote the optimal saliency map with red rectangles for the corresponding image. Take MC [6] as an instance, it obtains the best results for the second image, but it totally misses the salient object in the last image. Therefore, it is risky to adopt only one single algorithm to detect the salient objects on all test images.

Based on the aforementioned analysis, it will be great helpful if one could be aware of a saliency map's quality. In general, the quality score of a saliency map is calculated by comparing it with its corresponding ground truth saliency map. However, the ground truth saliency map is unavailable for a test image. In this paper, we propose to directly predict a saliency map's quality score where the ground truth saliency map is unavailable. It is apparent that once the saliency quality scores can be predicted precisely, these results will

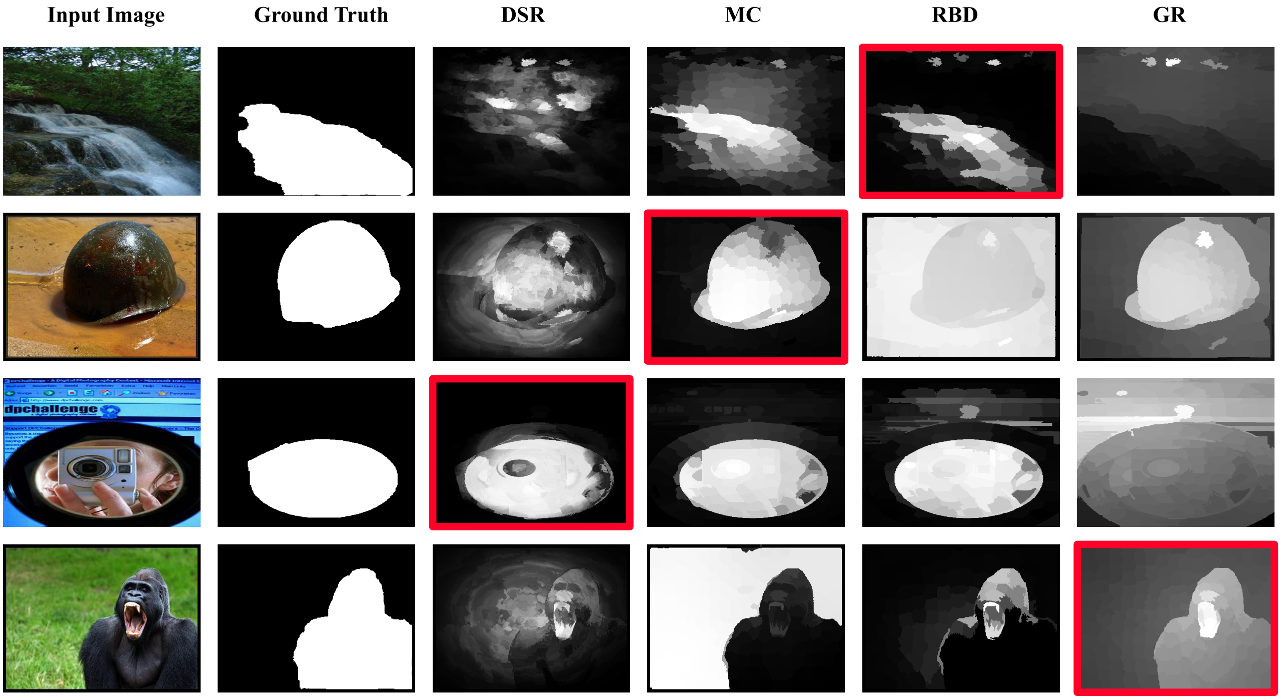


FIGURE 1. Examples of diverse salient detection results when four saliency detection algorithms, DSR [5], MC [6], RBD [7] and GR [8], are applied to different input images. The four input images are chosen from MSRA10K [9], where the aforementioned algorithms achieve state-of-the-art performances range from 95.4 to 95.5 in terms of AUC score. The saliency maps highlighted with red rectangles represent the best results for the corresponding images.

greatly improve the performance of saliency-based applications or even improve the saliency detection algorithms in return. For example, before applying saliency map, one could only choose the optimal saliency map from a set of candidate saliency maps, or only apply it when the quality of saliency map is tolerable. Alternatively, one could just tweak the algorithm's parameters to get a saliency map with satisfied quality.

To achieve this goal, we model saliency quality assessment problem as a multi-output regression problem. We observed that a good saliency map should not only locate the salient objects accurately but also highlight the whole salient object uniformly. Intuitively, when we look at a good saliency map from a multi-scale perspective, it should give us such impressions: 1. rough location from the low scale map. 2. fuzzy shape of a salient object from the middle scale map. 3. precise shape with a clear boundary from the finest scale. To leverage such observations, we utilize the state-of-the-art deep convolutional neural network technique since it possesses the ability to encode the aforementioned characteristics in its hierarchical layers [10]. For example, the initial layers tend to deal with the smooth boundary of salient objects [10]. It won't activate the following layer if there is no clear shape in the saliency map. Hence we take advantage of deep convolutional neural work to directly predict the quality score of a saliency map. The proposed network, derived from canonical network architecture, is referred as deep saliency quality assessment network (DSQAN) in this paper.

To depict a saliency map's quality, we propose joint saliency quality metric, which is a vector concatenation of multiple saliency quality scores. The motivation behind this idea is that different saliency quality metrics emphasize different kinds of quality. And as a consequence, the proposed joint metric brings us three evident advantages. The first advantage is that when the proposed DSQAN is trained with joint metric rather than single metric, it converges to a lower training loss. This is because that the decrease of one metric's prediction loss is capable of bringing down other metrics' prediction loss, since they are complementary to one another. Secondly, training with joint metric generates a more robust convergence route than only using single metric. The last benefit using joint metric prediction is that it offers us more choices to select the optimal saliency map. Specifically, we propose a simple fusion strategy to generate a fusion saliency quality score by combining joint metrics. Based on such fusion quality score, the result of our optimal saliency map selection algorithm significantly outperforms all single saliency detection algorithms in terms of all metrics. In addition, we also apply our DSQAN for salient object segmentation. The segmentation performance is improved about 3% compared with the best single method in terms of mean overlap score. This paper is an extended version of work published in [11], and it significantly expands previous work mainly in three aspects:

1) We propose joint metric to depict the quality of a saliency map. The superiority of such strategy is that it not only produces lower and more robust saliency predicted performance than single metric learning, but also introduces a more comprehensive saliency quality criterion by fusing joint metric into a single metric.

2) To explore the impacts of the architectures of DCNN, we adapt four architectures to our DSQAN. We reveal that the number of downsampling layers has a great influence on the problem of saliency quality prediction.
3) Two applications of the proposed DSQAN are presented in this paper, which are salient object detection and object segmentation. The experiments demonstrate the proposed method significantly outperforms all single algorithms in both tasks.

The remaining sections of this paper are organized as follows: We first briefly introduce the existing works that are related to our work in Section II. Then, the proposed DSQAN with joint metric is presented in Section III. Next, the experimental results are shown in Section V. Finally, the applications of DSQAN are presented.

II. RELATED WORKS

As far as we know, there are no existing works that are directly related to our work. The most related work is [12], where they develop an algorithm to rank different saliency results. They extract a set of hand-crafted features from a saliency map and its corresponding RGB image, which are considered to be related to the quality of saliency map. These features include saliency coverage, saliency map compactness, saliency histogram, color separation. After extracting these features, they adopted the pairwise-based learning-to-rank methodology to train a ranking model. Compared to their work, our proposed method can directly predict the saliency quality score of a single saliency map.

Although there are no existing works that can predict the quality of a saliency map, a large mass of works have been developed to detect the salient objects. Tang *et al.* [13] and Li and Ngan [14] propose to generate the saliency maps by integrating other saliency maps in order to get more balanced results. Li *et al.* [15] propose a novel method to discover co-salient objects from a group of images, which is accomplished by linearly fusing an intra-image saliency map and an inter-image saliency map. Li *et al.* [16] modeled saliency computation as two parts, average-to-peak ratio (APR) saliency and chrominance-aware (CA) saliency. Li and Yu [17] propose a multi-scale fully convolutional network as the first stream in our deep contrast network to infer a pixel-level saliency map directly from the raw input image. Reference [18] invented multiple saliency cues to generate saliency map separately, then fuse these saliency maps into final saliency map. Wang *et al.* [19] first train a deep neural network to predict a pixel's saliency value by considering its local context, then integrate proposal generation algorithms to boost its performance. Liu *et al.* [20] present a novel framework, named as Saliency Tree, to detect the salient objects. Specifically, they first compute the initial saliency map by combining global contrast, spatial sparsity, and object prior, then refine the saliency map under the proposed framework. Du *et al.* [21] propose to detect salient object in RGBD images. Specifically, progressive region classification is invented to model the saliency distribution and saliency map is generated via two scale integration. Saliency detection on faces images is conducted in [22], where they extract different types of face and facial features to model the salient regions in faces images. Wang *et al.* [23] propose to detect the salient objects from videos by integrating static saliency and dynamic saliency networks simultaneously. Zhang *et al.* [24] identify salient objects from super-pixel clusters rather than pixels or super-pixels, which first cluster super-pixels using Laplacian sparse subspace clustering (LSSC), then formulate the saliency detection of each super-pixel cluster as a unified low-rankness and sparsity pursuit problem. Zhang *et al.* [25] deal with co-saliency detection problem by combining intra-saliency prior transfer and deep inter-saliency mining. Specifically, a stacked denoising autoenoder (SDAE) is built to model the saliency prior knowledge while the deep inter-saliency mining is formulated by using the deep reconstruction residual obtained in the highest hidden layer of a self-trained SDAE. Zhou *et al.* [26] propose a saliency quality weighted based fusion method to improve the performance of initial saliency map. Zhou *et al.* [27] propose to boost the saliency detection performance by utilizing the predicted ranking results [12]. Ye *et al.* [28] integrate saliency and objectness to boost saliency detection result. Liu *et al.* [29] improve the saliency detection performance on unconstrained videos by construct a superpixel-level graph and spatiotemporal propagation.

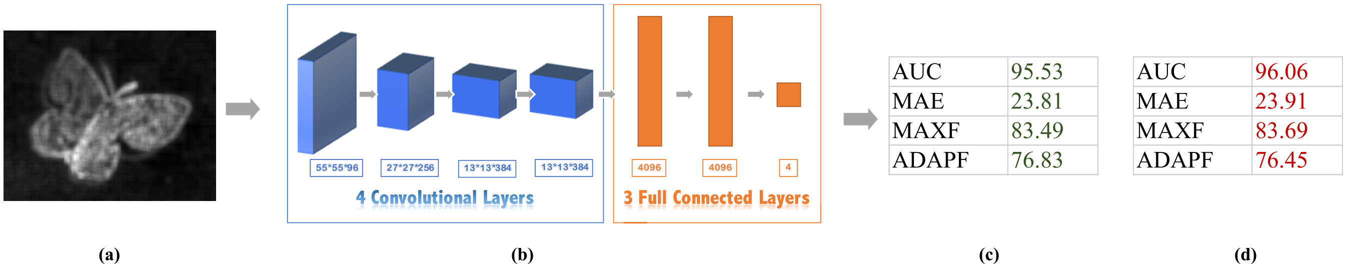
Another research field related to our work is image quality assessment. Wu *et al.* [30] propose to assess the image quality by introducing the multi-domain structural information and piecewise regression. Specifically, they fuse the local features and a complementary global feature to tackle different types of distortion. Another kind of image quality assessment is blind image quality assessment(BIQA), which is aimed to assess the quality of an image when the reference image and distortion type are not provided. In [31], they design a new feature fusion strategy to accomplish the task, which is conducted by fitting an image's colors and frequency characteristics and predicting the quality score via label transfer. Wu *et al.* [32] propose faster-than-real-time approach named local pattern statistics index (LPSI) to meet the industrial needs, which also reached competitive quality prediction performance with other state-of-the-art approaches.

III. DEEP SALIENCY QUALITY ASSESSMENT NETWORK

In this section, we present the details of the proposed deep saliency quality assessment network (DSQAN) with joint metric. We first briefly present the basic architecture of our DSQAN. Then, the motivations and objectives of introducing joint saliency quality score will be presented. Finally, to explore the impact of different CNN architectures on predicting the quality of saliency map, we utilize four classic CNN architectures as variants of our DSQAN.

A. ARCHITECTURE

We generate our deep saliency quality assessment network (DSQAN) from the basic architecture of existing deep



AUC	95.53
MAE	23.81
MAXF	83.49
ADAPF	76.83

AUC	96.06
MAE	23.91
MAXF	83.69
ADAPF	76.45

FIGURE 2. Proposed Deep Saliency Quality Assessment Network. (a) Input saliency map. (b) Proposed Saliency Quality Assessment Network. (c) Joint saliency quality score predicted by our DSQAN. (d) Actual joint saliency quality score computed with the ground truth saliency map.

convolutional neural network. Figure 2 shows an example of our DSQAN where Alexnet is applied. First of all, we modify the size of the feature map in the first convolutional layer from 227 × 227 × 3 to 227 × 227 × 1, since the saliency map, as our network's input, is a grayscale image. As a result, all the saliency maps are resized to 227 × 227 × 1 before fed into the deep convolutional neural network. Then, the saliency map is subtracted by the mean saliency map, where the mean saliency map is calculated on the whole training set through averaging pooling. We preserve the intermediate middle layers unchanged because it has the ability to express the multi-scale processing of saliency map while not being over-complicated. And we can also replace it with other basic CNN architectures. As for the last layer, the 1000 neurons in the last layer of Alexnet is designed for image classification, while our goal is to accomplish image regression, i.e. predict the saliency quality score. Based on this requirement, we set the number of the last neuron to K, where K correspond to the number of saliency quality metrics. Therefore, the size of last fully-connected layer's weight is $1000 \times K$ (K is set to 4 here).

Regarding the original architecture, we replace the original normalization layer with batch normalization layer [33], which means to perform the normalization for each training mini-batch. The experimental results show this will decrease the mean prediction error by about 1%.

B. JOINT SALIENCY QUALITY METRIC

To measure the quality of saliency map, there are multiple standard metrics [34] used to compare the performance of saliency detection algorithms. We first briefly introduce and analyze four well-known metrics in the following.

a: AUC

The first metric is Area Under ROC Curve (AUC) score [34], denoted as S_{AUC}, where ROC is a two-dimensional representation of an algorithm's performance. AUC synthesize this information into a single scalar, calculated as the area under the ROC curve. AUC score is proportional to the quality of a saliency map.

b: MAE

Another easy and intuitive metric is mean absolute error (MAE), denoted as S_{MAE} between the continuous saliency map M and the binary ground truth M_G as the quality score of saliency map. Both saliency map and binary ground truth map are normalized in the range [0, 1]. The MAE score is defined as:

$$S_{MAE}(M) = \frac{1}{N}\sum |M(i) - M_G(i)| \tag{1}$$

where N is the number of pixels in an image, $S(i)$ and $S_G(i)$ are the saliency values in pixel i in the saliency map and binary ground truth respectively.

The lower the MAE score is, the better the quality of saliency map is. When the MAE is 0, it means that the saliency map is completely the same as the binary ground truth.

c: MAXF

To emphasize the precision of salient object detection algorithms, maximal F-measure denoted as S_{MAXF}, is defined as weighted harmonic mean of precision and recall:

$$S_{MAXF} = \max_{p} \frac{(1+\beta^2) \times precision(p) \times recall(p)}{\beta^2 \times precision(p) + recall(p)} \tag{2}$$

where p represents the number of thresholds used to binarizing the predicted saliency map, β^2 is set to 0.3 for considering *precision* more valuable.

d: ADAPF

The last metric used in this works is Adaptive threshold F-measure, denoted as S_{ADAPF}, is calculated in a similar way to MAXF, except there is only one threshold used to binarizing the saliency maps. The adaptive threshold is calculated as follows:

$$TH = \frac{1}{N}\sum |M(i)| \tag{3}$$

where N is the number of pixels in the saliency map, and the adaptive threshold here refers to the mean saliency value.

Instead of picking a single metric, we propose a joint metric to measure the quality of saliency map. Specifically, we concatenate the four aforementioned well-known metrics into a joint vector as the saliency quality measurement. The reason for adopting this strategy mainly lies on three perspectives. First of all, it can be observed from the above definitions, these metrics are all linearly associated with the saliency quality. Therefore, to predict such joint quality vector, it is intuitive to share the most previous layers in our DSQAN. Secondly, different metrics emphasize different aspects of saliency quality. For example, S_{MAE} treats salient object and backgrounds equally, when the scale of a salient object is small, even the algorithm totally fails to detect such salient object, the S_{MAE} can still be low. Under such conditions, S_{AUC} and S_{MAXF} are better choices to depict the salient quality. To express the diversity of these metrics in a thorough way, we draw the statistics of these four metrics on the training set in Figure 3. It can be seen from this figure, the distributions are quite different from each other. For example, the distribution of S_{MAE} is like a gaussian distribution while S_{AUC} is more like a beta distribution. For such observation, it it obviously more appropriate to adopt joint saliency metric. Thus, the last fully-connected layer with size $N * K$ is utilized to express such differences. The last advantage of joint metric is that it leads to better optimization since the decrease of the predicted loss of one metric will directly influence other metrics' predicted loss.

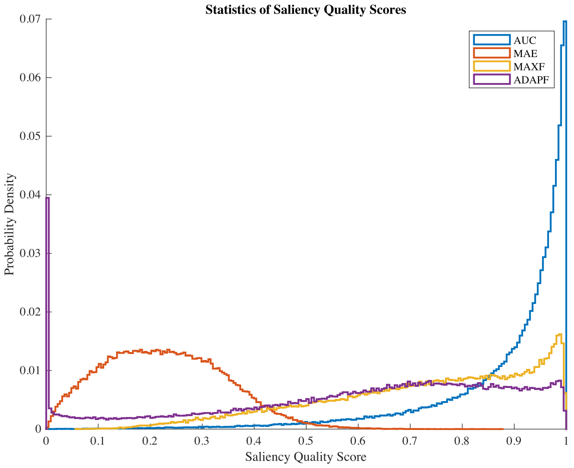
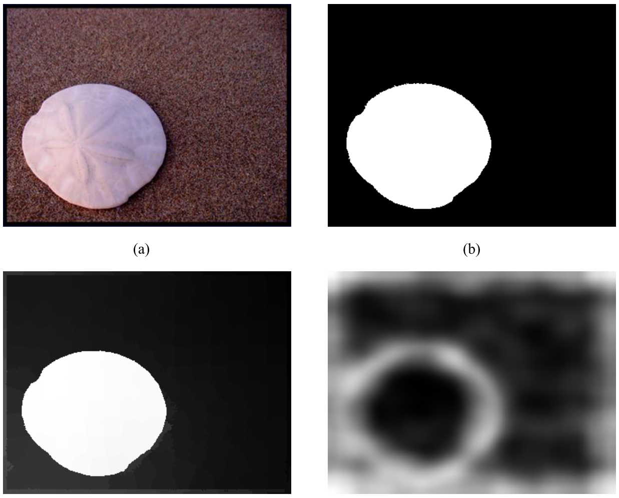


FIGURE 3. Statistics of four well-known saliency quality metrics collected on the training set. These metrics are AUC, MAE, Maximal F-measure and Adaptive threshold F-measure.

Figure 4 shows an example of predicting saliency maps' joint quality scores using our DSQAN. The above two pictures in the first row of Figure 4 are the original input image and corresponding ground truth saliency map. The two below saliency maps are generated using SER [35] and GR [8] respectively. The ground truth and predicted joint quality scores are listed under the images. For both the high and low-quality saliency map, our DSQAN is able to predict the joint quality scores with quite small prediction errors.

Figure 2 shows an example of our DSQAN, where we utilize Alexnet [36] as the basic architecture here. The basic units contain three blocks: down-sampling convolutional layers, fully-connected layers, and joint metric prediction layer. In figure 2, the joint quality metric of a saliency map is predicted directly by our DSQAN with small prediction error. The 4 neurons of the last layer are used to output the joint quality metric. We can see that the proposed DSQAN accurately predict the joint saliency score by offering very low prediction error in terms of all four metrics.



(a) (b)

(c) (d)

	AUC	MAE	MAXF	ADAPF
PD	**99.19**	**7.23**	**98.47**	**99.10**
GT	**99.93**	**7.17**	**99.87**	**99.87**

(e)

	AUC	MAE	MAXF	ADAPF
PD	**39.88**	**42.50**	**29.13**	**11.28**
GT	**39.74**	**42.30**	**26.10**	**10.66**

(f)

FIGURE 4. Examples of predicted joint saliency quality score for two saliency maps using proposed Deep Saliency Quality Assessment Network. The scores corresponding to GT are the actual saliency quality score computed with the ground truth map, while the score corresponding to PD are the predicted saliency quality score.

C. VARIANTS OF DSQAN

From the previous section, it can be seen that our saliency quality assessment network can be generated by modifying any existing DCNN. In order to explore the performance when applying different deep architectures to saliency quality assessment, we modify several existing deep convolutional networks. Specifically, four classic DCNN architectures, Alexnet [36], VGG-f [37], VGG-m [37], VGG-s [37], are adopted in this paper.

Alexnet is the winning model in ISLVRC 2012 and also the first well-known deep convolutional neural model. The size of input to this network is 227 × 227 × 3. The input is first filtered with 96 kernels with size 11 × 11 × 3 with a stride of 4 pixels. After the first convolutional layer, ReLu layers and normalization layer are followed. Next, three similar convolutional architectures are applied with different size of kernels. After these convolutional blocks, three fully-connected layer are followed, and the numbers of neurons are 4096, 4096 and 1000 respectively. 1000 neurons in the last layer correspond to 1000 categories.

VGG-f is similar to Alexnet, which consists of 8 layers in total, 5 convolutional layers and three fully-connected layers. The input image size is 224 × 224, which is slightly smaller than that of Alexnet. The difference to Alexnet is that VGG-s uses fewer filters in the first convolutional layer. Since the input of our saliency quality assessment network is gray scale, which has only one feature channel, the number of filters can be reduced in the first place theoretically. And this architecture can be used to verify such assumption by checking if the performance will decrease.

VGG-m is similar to ZFNet [10], which is defined by smaller stride and receptive field in the first convolutional layer. The second convolutional layer uses larger stride to keep the computation time reasonable [37]. Spatial support is a key factor in saliency detection algorithms [5], [7], [38]–[40]. Because both the receptive field and stride are smaller, the most critical point is that this architecture adds another scale of saliency map with a downsampling factor 2 rather than 4. Such discrepancy can explore the effectiveness of multi-scale processing.

VGG-s is related to the accurateĂŹ network from the overFeat package. It also uses 7 × 7 filters with stride 2 in the first convolutional layer. Unlike VGG-s, the stride in the second convolutional layer is smaller (1 pixel), but the max-pooling windows in layer 1 and 5 are larger (3 × 3) to compensate for the increased spatial resolution. Compared to [27], 5 convolutional layers are used as in the previous architectures ([27] used 6), and fewer filters in the 5th convolutional layer (512 instead of 1024);

IV. TRAINING

The goal of training is to find the parameters of DSQAN that minimize the average predicted quality loss. To train our DSQAN, we first generate a large number of saliency maps on the MSRA10K [9] dataset. For each input image, we generate 15 saliency maps using 15 state-of-the-art saliency detection algorithms. The details of training data will be presented in the Experiments Section.

A. LOSS FUNCTION

Since we modify the DCNN to regress the saliency quality score, it's necessary to define the task specific regression loss function. Considering most regression loss functions depend on the residual between ground truth value and predicted value:

$$r = l - \hat{l} \tag{4}$$

where l refers to the predicted continuous value, while $\hat{l}$ refers to ground truth continuous value.

In our DSQAN, we choose the square ($l2$) Loss as our regression loss:

$$L = \sum_i \sum_k \sqrt{(S_{ik} - \hat{S_{ik}})^2} \tag{5}$$

where i represents the ith saliency map i, k refers to the kth saliency quality metric, S refers to the ground truth saliency quality score and $\hat{S}$ refers to the corresponding predicted score. In conclusion, the objective of learning DSQAN is to minimize the average error between multiple predicted saliency scores and ground truth saliency quality scores over the training set.

B. OPTIMIZATION

Optimization is conducted by stochastic gradient descent using mini-batches of N samples [36], and here N is up to the architecture of DSQAN. In most cases, the weights of modified network will use the learned weights on some larger dataset. Considering the goal of our network is completely different from image classification models, the weights of our network are randomly initialized. Specifically, the weights of the filters in our DSQAN were initialized by random sampling from a Gaussian distribution with zero mean and 10^2 standard deviation. The training images were resized to 227 × 227. To deal with image classification tasks, the network is fed with patches cropped from these images (where crops change every time an image is sampled). However, our saliency quality score is corresponding to the whole image. Thus, we don't apply the cropping strategy.

At test time, we fetch out the last layer, adopting the output of last fully-connected layer as the continuous saliency quality score of a saliency map.

V. EXPERIMENTS

A. SETUP

We evaluate our proposed method on MSRA10K [9] dataset, which contains 10,000 images and corresponding ground truth binary masks. To train DSQAN, we randomly spilt this dataset into train, validation and test, with 6,000, 2,000, and 2,000 images respectively. We select 15 state-of-the-art saliency detection algorithms to generate saliency maps, which are DSR [5], MC [6], RBD [7], GR [8], SeR [35], CA [41], FES [42], AC [43], PCA [44], SEG [45], SIM [46], SR [47], SUN [48], and SWD [49]. This gives us 150,000 saliency maps in total. The corresponding saliency maps are generated from the code provided by the author or directly downloaded from the author's homepage. The DSQAN is learned via cross validation on training set and validation set. We utilize matconvnet [50], a CNN toolbox developed on MATLAB, to implement our DSQAN.

Before the performance of proposed method is presented, we propose two baseline algorithms to predict the saliency map's quality score. As the first baseline algorithm, it assigns a random saliency quality score in [0, 1] for a given saliency map, referred as *RndUnif*. The second one is to assign a random saliency quality score based on the estimated gaussian distribution, which is obtained by fitting all the saliency quality scores on the training set, and this algorithm is referred as *RndGaus*.

To evaluate the generation of proposed method, we conduct the same experiment on DUT-OMRON [51] dataset, which has 5166 images. Generally, DUT-OMRON is used to compare models on a large scale and has more complex backgrounds. Similarily, we generate the saliency maps using the aforementioned 15 saliency detection algorithms on all the images of DUT-OMRON dataset. It is worth noting that the salient objects on these two datasets are different from those on the MSRA10k dataset.

TABLE 1. Mean prediction errors in terms of four metrics on the test set of MSRA10k dataset (%). For one metric, the best three results are shown in red, green and blue, respectively.

	$RndUnif$	$RndGaus$	Mai's [12]	Alexnet	VGG-f	VGG-m	VGG-s
AUC	33.67	16.01	9.35	**4.71**	4.74	**4.47**	**4.54**
MAE	33.95	13.06	6.55	**3.33**	3.34	**3.17**	**3.18**
MAXF	34.16	23.02	11.93	8.12	**7.93**	**7.62**	**7.62**
ADAPF	33.65	31.32	14.36	8.96	**8.73**	**8.48**	**8.58**

TABLE 2. Mean prediction errors in terms of four metrics on the DUT-OMRON Dataset (%). For one metric, the best three results are shown in red, green and blue, respectively.

	$RndUnif$	$RndGaus$	Mai's [12]	Alexnet	VGG-f	VGG-m	VGG-s
AUC	38.55	20.03	16.66	14.12	**14.01**	**13.03**	**13.45**
MAE	33.18	14.52	8.11	5.01	**4.89**	**4.10**	**4.24**
MAXF	32.29	30.25	16.78	14.29	**14.09**	**13.60**	**13.56**
ADAPF	34.41	32.78	20.08	14.55	**14.44**	**13.76**	**13.61**

B. METRIC

We evaluate the quantitative results in terms of mean of prediction errors, and the predicted quality error of a saliency map is calculated as follows:

$$MPE = \frac{1}{N}\sum_{i}|S(i) - \hat{S}(i)| \tag{6}$$

where $S(i)$ is the saliency quality score of saliency map i, which is computed with the ground truth map, while $\hat{S}_i$ is the saliency quality score of saliency map predicted by our DSQAN. The mean prediction error is calculated over the whole test set, and N is the number of saliency maps on the test set.

C. COMPARISON AND ANALYSIS

In this section, we present and analyze the experimental results when conducted on DUT-OMRON dataset [51] and compare our DSQAN with [12]. Although [12] is designed to rank more than one saliency maps, their hand-crafted features are capable of describing the quality of a saliency map. Therefore, we implement the features proposed in [12] by ourself (the author doesn't provide the code), including Saliency Coverage, Saliency Map Compactness, Saliency Histogram, Color Separation, Segmentation Quality and Boundary Quality. All the parameters are specified according to their paper, and it generates a 41 dimension feature for one saliency map. Then, support vector regressor (SVR) [52] are used to predict the saliency quality scores of saliency maps against these features. The optimal parameters of SVR are learnt via cross-validation on the same training and validation sets.

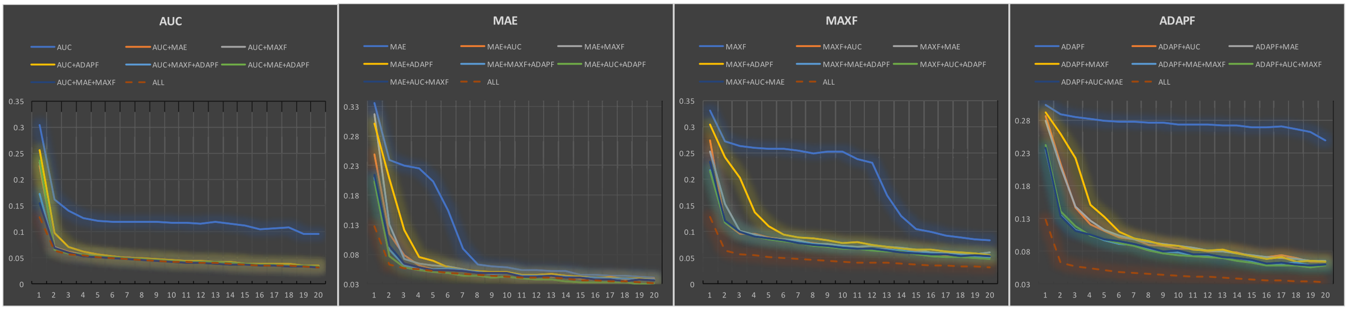
The results are shown Table 1 and 2 in terms of mean prediction errors of four saliency quality metrics. First of all, it can be observed that the proposed DSQAN has the ability to predict the quality of a saliency map with a low prediction error in terms of all four metrics. Specifically, the mean prediction error of the proposed DSQAN is about 3.3% in terms of MAE score on MSRA10k dataset, which roughly means that for a given saliency map, the predicted saliency quality score is near the actual saliency score by only 3.3% error.

For the comparison and generalization, we can see that our DSQAN consistently outperforms Mai's method [12] in terms of four metrics on both MSRA10k and DUT-OMRON datasets. For example, the proposed DSQAN is 4.5% lower than Mai's method [12] in terms of AUC score on the MSRA10k dataset. Another phenomenon is that the performances of both our DSQAN and Mai's [12] method are quite different when evaluated on DUT-OMRON dataset compared to MSRA10k dataset. We presume there are two reasons causing this diversity. The first reason is that the salient objects are very different on various aspects, such as scale, number and locations. Another reason is that the aforementioned saliency detection algorithms show inconsistent performances on DUT-OMRON dataset since DUT-OMRON dataset is much harder for most saliency detection algorithms [34].

In the end, we also exhibit the performances of different network variants in terms of mean prediction errors are listed in Table 1 and 2. From these results, it can be observed that the architectures of VGG-m and VGG-s slightly outperform Alexnet and VGG-f. Incorporating the assumptions in Section III-C, first of all, it demonstrates that the decrease in the number of filters in the first convolutional layer has little influence on the performance. On the other hand, increasing the number of downsampling layers will elevate the prediction accuracy according to the results of VGG-m and VGG-s. This proves our idea presented in the very beginning that multi-scale representation of a saliency map greatly influences its saliency quality.

D. THE EFFECTS OF THE NUMBER OF METRICS

In this section, we justify the choice of designing saliency quality assessment network with joint metric, instead of utilizing single metric separately. An evident superiority is that such architecture can output multiple metrics using only one network. To explore the impacts of the number of metrics, we design a thorough and extensive experiment. For clarity, we take metric AUC as an example to explain the details of this experiment. We first train the proposed DSQAN where only AUC is used as the predicted quality metric, and plot the



(a) (b) (c) (d)

FIGURE 5. **The training error curves over a fixed number of iterations when the number of metrics is varying. From left to right: (a) AUC, (b) MAE, (c) MAXF and (d) ADAPF. The dashed line refers to the training error curve using all the four metrics.**

training error curve of AUC over training iterations. Then, we add another metric (MAE, MAXF or ADAPF) to repeat the same training process, which is designed to observe the performance variation of AUC when adding another metric. Next is to add another one from the remaining metrics until all metrics are used. For other three metrics, we conduct the same experiments. As a result, this gives us four illustrations shown in Fig. 5. These experiments not only produce the performances variations of adding or moving any metrics but also evaluate the performances of all combinations of multiple metrics. All these training processes are carried out with the same parameters.

Observing from Fig. 5, adding more metrics improves the performances in terms of four metrics in most cases, and we explain our results in three ways. First of all, it brings us much lower training error after the first iteration, which obviously leads to quick convergence. Take metric MAXF (the third plot in Fig. 5) as an instance, the training error is reduced to 0.27 from 0.33 after the first iteration when AUC is added. Furthermore, when four metric are joint learnt, the training error is dramatically reduced to 0.13, which is 0.3 lower than single metric. Secondly, adding more metrics draws a lower final loss. Take ADAPF (the fourth plot in Fig. 5) as an example, it is hard to reach a satisfied local optimal when only ADAPF is used. When AUC is joint learnt with ADAPF, the decrease of the training loss of AUC significantly drag down the training loss cure of ADAPF. While four metrics are adopted together, it leads to the lowest training losses at all iterations and also the final loss. The final observations is that adding more metrics shows diversity boost for different metrics' performances. For metric AUC (the first plot in Fig. 5), although adding more metrics leads to better performances, the improvement is limited after adding two metrics. In conclusion, these results prove the our assumption that since different saliency quality metrics emphasize different kinds of quality, the decrease of one metric prediction loss is capable of bringing down other metric prediction loss. In addition, training with joint metric generates a more robust convergence route than only using single metric. In the following sections, we will show another benefit using joint metric prediction, which offers us more choices to select the optimal saliency map.

VI. APPLICATIONS

A. OPTIMAL SALIENCY MAP SELECTION

To further evaluate the effectiveness of our proposed DSQAN, we apply DSQAN in choosing the best saliency map from a set of candidate saliency maps. For simplicity, we directly choose the saliency map with the highest predicted quality score as the optimal saliency map, then calculate mean of the chosen saliency maps' ground truth scores. For a fair comparison, we adopt the standard way used to compare the performance between different saliency detection algorithms, which are mean scores of AUC, MAE, MAXF and ADAPF. To express the ceiling performance of our optimal saliency map selection algorithm, we add another method, denoted as *ORACLE*, to show the theoretical upper limit.

To choose the optimal saliency map, there are multiple alternatives which can be utilized in our DSQAN. First of all, we use the predicted saliency quality scores to choose the optimal saliency map separately. The corresponding methods are denoted as $DSQAN_{AUC}$, $DSQAN_{MAE}$, $DSQAN_{MAXF}$ and $DSQAN_{ADAPF}$. For example, $DSQAN_{AUC}$ represents that the optimal saliency map is chosen with highest AUC predicted score. To better exploit the joint metric, we add another new method to choose the optimal saliency map. Concretely, we propose a simple strategy to fusion joint saliency quality scores, denoted as S_{JOINT}, as follows:

$$S_{JOINT} = e^{S_{AUC}} + e^{1-S_{MAE}} + e^{S_{MAXF}} + e^{S_{ADAPF}} \tag{7}$$

1) QUANTITATIVE RESULTS

Table 3 shows the quantitative performances when compare our optimal saliency map selection results with other saliency detection algorithms. The bold numbers with black colors in this table represent the best performances within 15 saliency detection algorithms. It can be observed that these methods perform non-consistent across different saliency quality metrics, and none of them could obtain the best result in all metrics. For our first group of optimal selection algorithms($DSQAN_{AUC}$, $DSQAN_{MAE}$, $DSQAN_{MAXF}$ and $DSQAN_{ADAPF}$), each of them obtains the best performance on the corresponding metric, bold numbers with red colors. For example, $DSQAN_{MAE}$ acquires 9.95 in terms of MAE score, about 1% lower than the best single method performance

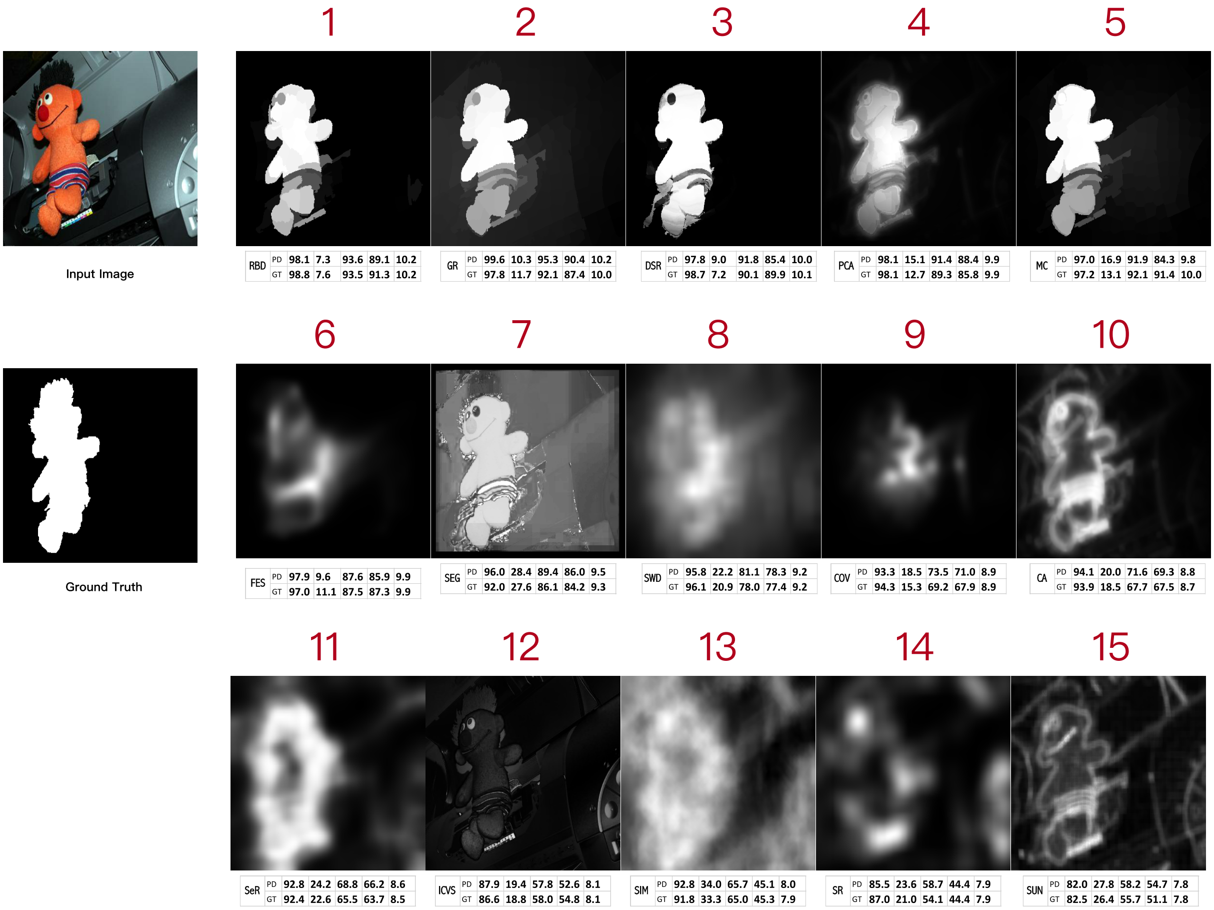

1
2
3
4
5
Input Image
RBD PD 98.1 7.3 93.6 89.1 10.2
GT 98.8 7.6 93.5 91.3 10.2
GR PD 99.6 10.3 95.3 90.4 10.2
GT 97.8 11.7 92.1 87.4 10.0
DSR PD 97.8 9.0 91.8 85.4 10.0
GT 98.7 7.2 90.1 89.9 10.1
PCA PD 98.1 15.1 91.4 88.4 9.9
GT 98.1 12.7 89.3 85.8 9.9
MC PD 97.0 16.9 91.9 84.3 9.8
GT 97.2 13.1 92.1 91.4 10.0
6
7
8
9
10
Ground Truth
FES PD 97.9 9.6 87.6 85.9 9.9
GT 97.0 11.1 87.5 87.3 9.9
SEG PD 96.0 28.4 89.4 86.0 9.5
GT 92.0 27.6 86.1 84.2 9.3
SWD PD 95.8 22.2 81.1 78.3 9.2
GT 96.1 20.9 78.0 77.4 9.2
COV PD 93.3 18.5 73.5 71.0 8.9
GT 94.3 15.3 69.2 67.9 8.9
CA PD 94.1 20.0 71.6 69.3 8.8
GT 93.9 18.5 67.7 67.5 8.7
11
12
13
14
15
SeR PD 92.8 24.2 68.8 66.2 8.6
GT 92.4 22.6 65.5 63.7 8.5
ICVS PD 87.9 19.4 57.8 52.6 8.1
GT 86.6 18.8 58.0 54.8 8.1
SIM PD 92.8 34.0 65.7 45.1 8.0
GT 91.8 33.3 65.0 45.3 7.9
SR PD 85.5 23.6 58.7 44.4 7.9
GT 87.0 21.0 54.1 44.4 7.9
SUN PD 82.0 27.8 58.2 54.7 7.8
GT 82.5 26.4 55.7 51.1 7.8

FIGURE 6. An example of saliency quality prediction and saliency maps ranking according to the predicted fusion quality scores. Each row contains five saliency quality scores, which are listed from left to right: AUC, MAE, MAXF, ADAPF and our Fusion score. The row denoted as "GT" is the actual saliency quality score computed with the ground truth map, while the row denoted as the "PD" is the predicted saliency quality score. The number with red color above each saliency map refers to its ranking order.

and only 1% higher than the *ORACLE* method, which is the theoretical upper limit. For our fusion method, $DSQAN_{JOINT}$, although it performs slightly lower than the previous four selection methods in terms of separate metric, it outperforms all saliency detection methods in terms of four metrics, which strongly proves the efficacy of the joint metric learning strategy.

2) QUALITATIVE RESULTS

To better visualize the results of proposed method, we plot an example saliency quality prediction results and optimal saliency map selection in Figure 6. In this figure, each row contains five saliency maps with their corresponding joint quality scores, where the joint saliency scores are listed from left to right: AUC, MAE, MAXF, ADAPF and our Fusion score. The row denoted as "GT" is the actual saliency quality score computed with the ground truth map, while the row denoted as the "PD" is the predicted saliency quality score. For better visual effect, we rank these saliency maps according to their predicted saliency quality fusion scores. Observing from this figure, the superiority of our joint saliency quality scoring method not only lies in choosing the good saliency maps with high performance in terms of objective metrics but also gives us better subjective saliency quality. Concretely, the saliency quality scores of fist five saliency maps are significantly higher than that of last five ones. In the meanwhile, they have more clear contours and they are more uniformly highlighted. These characteristics are very important when saliency map is used in other applications under most circumstances. As a contrasting example, to depict the disadvantage of single metric, we take the saliency map generated by SIM [46] as an example, which has a 92.8 AUC score, only about 5 % lower that the one generated by RBD [7]. However, the result of SIM looks significantly poorer than that of RBD.

B. SALIENT OBJECT SEGMENTATION

In this section, we apply our DSQAN to segment salient objects from input image and corresponding saliency map. We apply our DSQAN to choose the best saliency map from the aforementioned 15 results, then apply the following salient object segmentation method. In order to associate

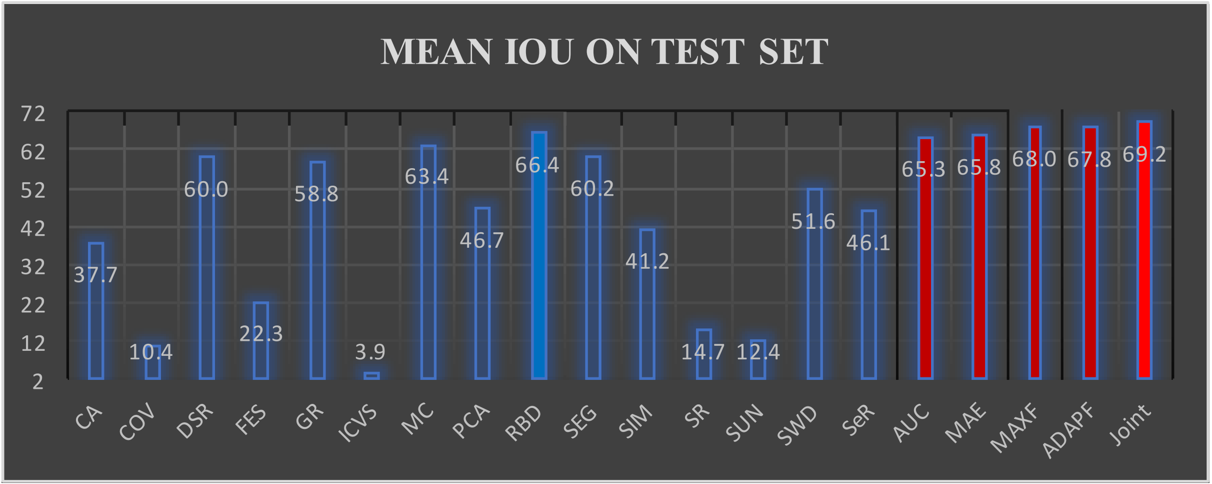


FIGURE 7. Mean IoU scores of different salient object segmentation algorithms on test set. The blue bars represents the performances of single saliency detection algorithms, and the sold blue represents the highest score. The red bars represents the performances of our proposed methods, where AUC means use predicted AUC score to choose the best saliency map and so on. (Best viewed in colors).

TABLE 3. Mean saliency quality scores of optimal saliency map selection on the test set (%).

METHOD	AUC	MAE	MAXF	ADAPF
CA	88.07	23.07	66.68	58.27
COV	90.78	19.02	69.17	63.05
DSR	**95.53**	11.50	86.80	80.48
FES	91.15	17.72	75.60	68.80
GR	95.48	19.46	89.59	74.49
ICVS	77.74	22.44	58.83	50.44
MC	95.35	14.27	89.58	**82.93**
PCA	94.93	17.88	82.11	74.73
RBD	95.35	**10.73**	**89.94**	82.42
SEG	89.43	31.20	77.74	52.58
SIM	80.65	38.51	56.77	24.50
SR	80.99	24.15	57.12	48.86
SUN	76.55	28.47	55.79	42.28
SWD	91.23	26.49	71.39	60.89
SeR	82.04	30.06	59.39	43.77
ORACLE	98.63	8.80	94.27	90.58
$DSQAN_{JOINT}$	**96.24**	**10.28**	**90.50**	**85.07**
$DSQAN_{AUC}$	**96.45**	13.30	89.07	82.28
$DSQAN_{MAE}$	95.38	**9.95**	89.01	83.40
$DSQAN_{MAXF}$	96.35	12.41	**90.81**	83.97
$DSQAN_{ADAPF}$	96.32	11.71	90.52	**85.25**

salient object detection and segmentation in a more consistent way, we propose an straightforward objective function for consistently segmenting salient objects:

$$E(y) = \sum_{p \in \mathcal{V}} U(y_p) + \lambda \sum_{p,q \in \mathcal{E}} V(y_p, y_q) \tag{8}$$

where y_p and y_q refer to segmentation variables at pixel p and pixel q respectively. When y_p equals to 1, it means p belongs to the salient objects.

The unary term $U(y_p)$ is designed to ensure that the final segmentation result y_p is close to the saliency value s_p of given saliency map at pixel p, therefore it is computed as follows:

$$U(y_p) = \sum_{p} \log {s_p}^{y_p} (1 - s_p)^{1-y_p} \tag{9}$$

The pairwise term $V(y_p, y_q)$ is used to smooth the saliency segmentation results, defined as follows:

$$V(y_p, y_q) = d(p, q)[y_p \neq y_q] e^{-\beta |f_p - f_q|^2} \tag{10}$$

where $d(p, q)$ represents the distance between region p and q, [.] refers to the indicator function, β refers to the parameter that weights the feature distance, and f_p is the color feature vector of region q. The λ is set to 10, and the trade-off parameter β is set to 1.5 in our experiments. The above objective function is a submodular binary discrete optimization, and it can be minimized using graph cuts [53].

To compare the performances of different salient object segmentation algorithms, we adopt the standard segmentation metric mean intersection-over-union(IoU) score, which is computed between the salient object segmentation mask and the ground truth mask.

We show the quantitative results in Figure 7, where the blue bars refer to the results using all the 15 saliency detection algorithms individually and the red bars denote the results when applying our DSQAN to choose the optimal saliency map as the input to equation 9. From this figure, we can see the highest segmentation score of single method is obtain by RBD [7], whose mean IoU is 66.37%. The mean IoU of optimal saliency map selection algorithm using our fusion score denoted as $DSQAN_{JOINT}$, achieves 69.18%, remarkably outperforms the best single method by 3%. Another two methods, $DSQAN_{MAXF}$ and $DSQAN_{ADAPF}$, also outperform each single method.

VII. CONCLUSION

In this paper, we propose Deep Saliency Quality Assessment Network (DSQAN) to directly predict the joint saliency quality score of a saliency map. DSQAN is derived from the canonical state-of-the-art network through task-specific modification. To better express the saliency map's quality, we propose joint saliency quality score, which is defined as the vector concatenation of four well-known metrics. It does not only produce more accuracy and robust predicted result, but also bring better choice to rank saliency maps. To investigate the effects of different architectures on saliency quality prediction, we implement our DSQAN under different CNN architectures. We demonstrate that the number of downsampling layers has a great influence on predicting the saliency quality. As the applications of our method, we apply the learned DSQAN to both optimal saliency map selection and salient object segmentation, and the experiments strongly prove the effectiveness of proposed method.

REFERENCES

[1] C. Christopoulos, A. Skodras, and T. Ebrahimi, "The JPEG2000 still image coding system: An overview," *IEEE Trans. Consum. Electron.*, vol. 46, no. 4, pp. 1103–1127, Nov. 2000.

[2] J. Wang, L. Quan, J. Sun, X. Tang, and H. Shum, "Picture collage," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 347–354.

[3] G. Evangelopoulos *et al.*, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1553–1568, Oct. 2013.

[4] H. Li and K. N. Ngan, "Saliency model-based face segmentation and tracking in head-and-shoulder video sequences," *J. Vis. Commun. Image Represent.*, vol. 19, no. 5, pp. 320–333, 2008.

[5] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2976–2983.

[6] B. Jiang, L. Zhang, H. Lu, C. Yang, and M.-H. Yang, "Saliency detection via absorbing Markov chain," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1665–1672.

[7] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2814–2821.

[8] C. Yang, L. Zhang, and H. Lu, "Graph-regularized saliency detection with convex-hull-based center prior," *IEEE Signal Process. Lett.*, vol. 20, no. 7, pp. 637–640, Jul. 2013.

[9] T. Liu *et al.*, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, Feb. 2011.

[10] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.

[11] L. Tang, Q. Wu, W. Li, and Y. Liu, "Deep saliency quality assessment network," in *Proc. IEEE Int. Conf. Multimedia Expo Workshop*, Jul. 2017, pp. 567–572.

[12] L. Mai and F. Liu, "Comparing salient object detection results without ground truth," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 76–91.

[13] L. Tang, H. Li, and T. Chen, "Extract salient objects from natural images," in *Proc. IEEE Int. Symp. Intell. Signal Process. Commun. Syst.*, Dec. 2010, pp. 1–4.

[14] H. Li and K. N. Ngan, "A co-saliency model of image pairs," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3365–3375, Dec. 2011.

[15] H. Li, F. Meng, and K. N. Ngan, "Co-salient object detection from multiple images," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1896–1909, Dec. 2013.

[16] H. Li, L. Xu, and G. Liu, "Two-layer average-to-peak ratio based saliency detection," *Signal Process., Image Commun.*, vol. 28, no. 1, pp. 55–68, 2013.

[17] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 478–487.

[18] H. Li *et al.*, "Using mid-high level cues to detect salient object," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2014, pp. 1–6.

[19] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2015, pp. 3183–3192.

[20] Z. Liu, W. Zou, and O. Le Meur, "Saliency tree: A novel saliency detection framework," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 1937–1952, May 2014.

[21] H. Du, Z. Liu, H. Song, L. Mei, and Z. Xu, "Improving RGBD saliency detection using progressive region classification and saliency fusion," *IEEE Access*, vol. 4, pp. 8987–8994, Dec. 2016.

[22] Y. Ren, Z. Wang, and M. Xu, "Learning-based saliency detection of face images," *IEEE Access*, vol. 5, pp. 6502–6514, Mar. 2017.

[23] W. Wang, J. Shen, and L. Shao, "Video salient object detection via fully convolutional networks," *IEEE Trans. Image Process.*, vol. 27, no. 1, Jan. 2018.

[24] Q. Zhang, Y. Liu, S. Zhu, and J. Han, "Salient object detection based on super-pixel clustering and unified low-rank representation," *Comput. Vis. Image Understand.*, vol. 161, pp. 51–64, Aug. 2017.

[25] D. Zhang, J. Han, J. Han, and L. Shao, "Cosaliency detection based on intrasaliency prior transfer and deep intersaliency mining," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1163–1176, Jun. 2016.

[26] X. Zhou, Z. Liu, G. Sun, L. Ye, and X. Wang, "Improving saliency detection via multiple kernel boosting and adaptive fusion," *IEEE Signal Process. Lett.*, vol. 23, no. 4, pp. 517–521, Mar. 2016.

[27] X. Zhou, Z. Liu, G. Sun, and X. Wang, "Adaptive saliency fusion based on quality assessment," *Multimedia Tools Appl.*, vol. 76, no. 22, pp. 23187–23211, Nov. 2016.

[28] L. Ye, Z. Liu, L. Li, L. Shen, C. Bai, and Y. Wang, "Salient object segmentation via effective integration of saliency and objectness," *IEEE Trans. Multimedia*, vol. 19, no. 8, pp. 1742–1756, Aug. 2017.

[29] Z. Liu, J. Li, L. Ye, G. Sun, and L. Shen, "Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2527–2542, Dec. 2017.

[30] Q. Wu, H. Li, F. Meng, K. N. Ngan, and S. Zhu, "No reference image quality assessment metric via multi-domain structural information and piecewise regression," *J. Vis. Commun. Image Represent.*, vol. 32, pp. 205–216, Oct. 2015.

[31] Q. Wu *et al.*, "Blind image quality assessment based on multichannel feature fusion and label transfer," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 3, pp. 425–440, Mar. 2016.

[32] Q. Wu, Z. Wang, and H. Li, "A highly efficient method for blind image quality assessment," in *Proc. Int. Conf. Image Process.*, Sep. 2015, pp. 339–343.

[33] S. Ioffe and C. Szegedy. (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift." [Online]. Available: https://arxiv.org/abs/1502.03167

[34] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, Dec. 2015.

[35] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *J. Vis.*, vol. 9, no. 12, p. 15, 2009.

[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[37] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. BMVC*, 2014, p. 1.

[38] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 409–416.

[39] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2083–2090.

[40] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1155–1162.

[41] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1915–1926, Oct. 2012.

[42] H. R. Tavakoli, E. Rahtu, and J. Heikkilä, "Fast and efficient saliency detection using sparse sampling and kernel density estimation," in *Proc. Scandin. Conf. Image Anal.*, 2011, pp. 666–675.

[43] R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk, "Salient region detection and segmentation," in *Computer Vision Systems* (Lecture Notes in Computer Science), vol. 5008, A. Gasteratos, M. Vincze, and J. K. Tsotsos, Eds. Berlin, Germany: Springer, 2008, pp. 66–75.

[44] R. Margolin, A. Tal, and L. Zelnik-Manor, "What makes a patch distinct?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1139–1146.

[45] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 366–379.

[46] N. Murray, M. Vanrell, X. Otazu, and C. A. Parraga, "Saliency estimation using a non-parametric low-level vision model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 433–440.

[47] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[48] 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, p. 32, Dec. 2008.

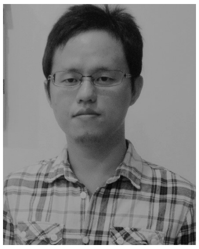
[49] L. Duan, C. Wu, J. Miao, L. Qing, and Y. Fu, "Visual saliency detection by spatially weighted dissimilarity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 473–480.

[50] A. Vedaldi and K. Lenc, "MatConvNet: Convolutional neural networks for MATLAB," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 689–692.

[51] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 3166–3173.

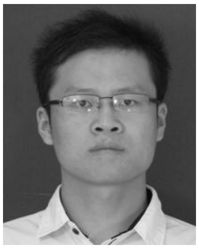
[52] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011.

[53] V. Kolmogorov and R. Zabin, "What energy functions can be minimized via graph cuts?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 147–159, Feb. 2004.

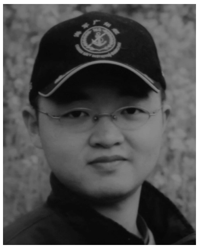


LIANGZHI TANG received the B.Sc. and M.Sc. degrees from the School of Electronic Engineering, University of Electronic Science and Technology of China, in 2008 and 2011, respectively, where he is currently pursuing the Ph.D. degree under the supervision of Prof. H. Li. His research interests include saliency detection, object segmentation, and deep convolutional neural network.

QINGBO WU received the B.E. degree in education of applied electronic technology from Hebei Normal University in 2009 and the Ph.D. degree in signal and information processing from the University of Electronic Science and Technology of China in 2015. In 2014, he was a Research Assistant with the Image and Video Processing Laboratory, The Chinese University of Hong Kong. From 2014 to 2015, he served as a Visiting Scholar with the Image and Vision Computing Laboratory, University of Waterloo. He is currently a Lecturer with the School of Electronic Engineering, University of Electronic Science and Technology of China. His research interests include image/video coding, quality evaluation, and perceptual modeling and processing.



WEI LI received the B.Sc. degree in electrical and information engineering from Henan Polytechnic University, Jiaozuo, China, in 2011. He is currently pursuing the Ph.D. degree with the Intelligent Visual Information Processing and Communication Laboratory, University of Electronic Science and Technology of China, Chengdu, China. His research interests include image recognition, object detection, and machine learning.



YINAN LIU received the B.Sc. and the M.Sc. degrees from the School of Electronic Engineering, University of Electronic Science and Technology of China, in 2009 and 2012, respectively, where he is currently pursuing the Ph.D. degree under the supervision of Prof. H. Li. His research interests include human action recognition, video representation, and deep learning.