IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 December 2022, accepted 5 January 2023, date of publication 13 January 2023, date of current version 27 January 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3236807

RESEARCH ARTICLE

Advertising Image Saliency Prediction Method Based on Score Level Fusion

QIQI KOU[1], RUIHANG LIU[2], CHEN LV[2], HE JIANG[2], AND DEQIANG CHENG[2], (Member, IEEE)

[1]School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China
[2]School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China

Corresponding authors: He Jiang (jianghe@cumt.edu.cn) and Deqiang Cheng (chengdq@cumt.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 52204177, and in part by the Fundamental Research Funds for the Central Universities under Grant 2020QN49.

ABSTRACT At present, visual saliency prediction algorithms have been developed more and more mature, but most of the current saliency prediction algorithms are aimed at natural images. Due to the inconsistency of elements and features between natural images and advertising images, the existing saliency prediction algorithms show poor robustness and low inference speed to advertising images, which severely limits its commercial application in advertising design and evaluation. In view of this, a saliency prediction algorithm for advertisement images is proposed in this paper. In the feature extraction stage, two text candidate regions based on intensity feature and improved MESR algorithm are first obtained and further integrated to produce a two-dimensional text confidence score. Meanwhile, a saliency confidence score is also obtained by an improved natural image saliency prediction network. Then, the score level fusion strategy was adopted to fuse the two confidence scores to get the final saliency prediction map. The experimental results show that the proposed model has good accuracy and robustness in advertising images, as well as the most remarkable inference speed, which can meet the demand for real-time performance of advertising image saliency prediction, leading to great practical and commercial value.

INDEX TERMS Saliency prediction, eye-gaze assessment, advertising image, score level fusion.

I. INTRODUCTION

Advertising is a core industry in today's digital world, and consumers are constantly overwhelmed by advertising images, which have become an indispensable part of televisions, web pages, posters, magazines, etc. Getting consumers' attention is a challenge advertisers are grappling with. To evaluate the advertising effectiveness is one of the most fundamental tasks in marketing. The purpose of advertising image saliency prediction is to estimate the most compelling position in an advertisement, which provides clues for further exploration of consumer intentions and behaviors. At the same time, it is also greatly beneficial to print advertisement design, image retrieval [1], product recommendation, advertisement evaluation, marketing strategy, and other cases.

In this digital age, consumers are often inundated with advertisements. Companies are adopting aggressive marketing strategies to aggressively market their products and gain an advantage over their competitors. A key part of this strategy, however, is audience measurement. When the company is able to thoroughly evaluate the effectiveness of its marketing strategy, it is likely to have a greater impact on potential customers. Therefore, measuring the effectiveness of advertising is an important task when determining the impact of a product. Traditionally, this can be done by collecting human feedback, namely free eye-gaze assessments.

Using saliency prediction to measure the expected audience response to an advertisement will enable companies to improve their design during the production stage. Therefore, a thorough analysis of human attention mechanisms can help companies redesign advertising to ensure maximum attention to key components, such as products, company

The associate editor coordinating the review of this manuscript and approving it for publication was Joewono Widjaja.

logos, etc. Although saliency prediction is an effective tool for this kind of practical market analysis, the applicability of existing models is limited. Advanced models such as deep convolutional networks require a large number of image saliency datasets and a highly configured platform, while traditional saliency models do not properly consider multi-level features, resulting in weak predictive ability. In addition, most of the current saliency prediction methods are aimed at natural scene images. As advertisement images are designed to attract customers' attention, they have unique characteristics, so saliency models trained on natural images are not robust to advertisement images.

To solve the above problems, a saliency prediction method of advertising images based on score level fusion is proposed. Firstly, the global saliency features on the advertising image dataset are fine-tuned based on the pre-trained model of natural saliency images. In addition, the text detection algorithm is used to identify the text region in the advertising image, and then the regions of the advertising image are Gaussian blurred and normalized to get the text features. Finally, the saliency features and text features are fused by the confidence scores, which is helpful to improve the robustness of saliency algorithm in advertising image prediction. It is worth mentioning that compared with the methods based on convolutional neural network, this method has faster speed and can meet the real-time requirements in advertising applications.

II. SALIENCY MODEL FOR NATURAL IMAGES

Since the ground-breaking bottom-up attention model is put forward by Treisman and Gelade in 1980, a large number of saliency prediction algorithms have emerged from the attention model to the saliency prediction model. In the following sections, the development of saliency prediction models will be briefly reviewed, and these models can be divided into two stages. The first stage is saliency detection algorithms based on traditional scale-space features such as intensity, color, and direction. In the second stage, with the innovation and development of deep neural networks, saliency algorithms based on deep learning have mushroomed like mushrooms, greatly enriching the research content and application in the field of visual saliency detection and prediction.

Inspired by Koch and Ullman [3], the first computational model is defined by Itti et al. [2] for saliency prediction. In his method, a set of separate feature vectors representing low-level features such as color, intensity, and orientation are combined into a global saliency map. Following his pioneering work, lots of methods incorporate complementary low-level features according to different principles, such as using Support Vector Machines (SVM) to combine multiple features with trainable weights at the score level. In addition, using semantic classifiers to detect higher-level semantic information, such as faces, people, and cars, can also be applied to improve the accuracy of saliency prediction methods.

The success of Convolutional Neural Network (CNN) in large-scale object semantic recognition [4], [5], [6], [7] has brought a wave of new saliency models, whose performance is significantly better than that of traditional saliency models based on manual features. Researchers used existing CNN for scene recognition training and re-used them for saliency prediction. These models are trained in an end-to-end manner, effectively defining saliency as a regression problem. To compensate for the lack of sufficiently large scale gaze datasets, we pre-trained the deep saliency model on large image datasets and then fine-tuned it on small-scale eye movement or click datasets. In this process, saliency method applies the semantic information that has been learned from CNN to saliency features, which greatly improves the performance of saliency feature extraction.

eDN [8] is the first time to introduce CNN into the field of visual saliency. Motivated by the characteristics of the biological hierarchy, multiple network layers with abundant parameters are constructed. Then, the hyperparameter optimization method is used to find independent models with significant prediction, and linear SVM is trained to combine them into a single model. Next, DeepGaze I [9] is proposed to use a relatively deep CNN (pre-trained AlexNet) for saliency prediction. The output of the convolutional layer is used to create and train a linear model to calculate the saliency of the image. In the latter period, DeepGaze II [10] based on DeepGaze I [9] is introduced. Moreover, the model further explores the unique influence of low-level and high-level features on saliency prediction performance.

SalGAN [11] is first constructed by using a Generative Adversarial Network (GAN) in the field of saliency prediction. It consists of two modules, a generation module, and a discrimination module. The generator learns by using backpropagation of binary cross-entropy loss on the existing saliency map, which is then passed to a trained discriminator to identify whether the provided saliency map is synthesized by the generator or real eye gaze data.

With the development of deep learning technology, saliency models are no longer limited to CNN. TranSalNet [12] utilizes the transformer in self-attention mechanism to encode long-range information and integrates it into CNN to capture remote contextual visual information. This algorithm proves that transformer has better performance in saliency prediction, and it enhances the perceived relevance.

Due to the higher requirements for saliency prediction efficiency, lightweight saliency models are gradually emerging. MSI-Net [13] is proposed based on CNN pre-training. The architecture forms an encoder and decoder structure and includes modules with multiple convolutional layers to capture multi-scale features at different expansion rates in parallel. The model is based on a lightweight image classification backbone, the number of trainable parameters of the model is greatly reduced, and the model has better performance.

Reddy et al. proposed two new end-to-end architectures called SimpleNet and MDNSal [14], which are neater, minimal, easier to interpret, and achieve state-of-the-art performance in common saliency benchmarks. SimpleNet is an optimized encoder-decoder architecture that delivers significant performance gains on SALICON datasets. MDNSal is a parameter model that directly predicts the distribution parameters of Gaussian Mixture Models (GMM). Its purpose is to make the saliency prediction map more interpretable.

FastSal [15] can realize knowledge transfer from more computationally expensive model, such as DeepGaze II, through unlabeled datasets. Knowledge distillation is used to learn the features from a replaceable teacher network, and the results provided by this method are comparable to many state-of-the-art algorithms. The computational cost and model size are only a fraction of those methods.

Zabihi et al. proposed a compact and real-time saliency prediction model [16]. The model consists of an improved U-Net architecture, a novel fully connected layer, and a central differential convolutional layer. The revised U-Net architecture is more compact and efficient. This novel fully connected layer helps to implicitly capture location-related information. Using central differential convolutional layers at different scales can capture more robust and biologically motivated features.

AMC-SNet [17] is proposed to use a lightweight network instead of the original traditional CNN as the backbone of feature extraction to reduce the number of model parameters and floating point calculations. Moreover, a multi-scale fusion of local and global features is added to further improve the saliency prediction performance.

However, although the above methods have achieved certain results in natural images, and a few methods have been extended to the saliency prediction of advertising images, the robustness of these algorithms for advertising images still need to be improved. Furthermore, the biggest problem is that advertising images are oriented to commercial applications, while the existing saliency prediction methods for advertising images take a long time to inference, which is difficult to meet the commercial promotion and application needs of real-time interaction.

Hence, the main work of this paper is to propose a saliency prediction method that can not only comparable with the state-of-the-art algorithms but also improves the robustness of the saliency algorithm for advertising images with a faster inference speed, which can meet the demand for real-time performance of advertising image saliency prediction and has great practical and commercial value.

III. SCORE LEVEL FEATURE FUSION STRATEGY

With the support of powerful computer information technology, the saliency prediction method of natural scene images has achieved fruitful results. However, the application field of saliency is not limited to natural scene images, and more special-purpose images are needed. For example, the visual saliency of advertising images in this paper has a high application value in advertising design and audience assessment, and other tasks. However, since the artificially designed advertisement images contain different elements from the natural scene images, how to effectively fuse the features of the advertisement images based on the saliency of the natural images becomes the problem to be solved by our algorithm in this paper.

At present, researchers have proposed a large number of saliency models based on different assumptions. In practice, while following an assumption or computational principle can make the model perform well on certain types of images, it can hinder the model's performance on arbitrary images and datasets. Inspired by the successful fusion strategy of semantic analysis and multimodal biometrics, this paper proposes to fuse the most advanced saliency model at the score level to assist in promoting the learning style. The score level feature fusion strategy first uses the saliency map generated by multiple models as the confidence score. These scores are then fed to a learner, i.e. SVM, adaptive lifting, or probability density estimator, to generate the final saliency map.

There are three kinds of score level fusion strategies: transformation-based fusion, classification-based fusion, and density-based fusion.

A. SCORE LEVEL FEATURE FUSION BASED ON TRANSFORMATION

In the transform-based fusion strategy, the confidence scores are first normalized to a common range for further merging. Each individual saliency map is normalized to zero mean and unit standard deviation using the normalization method described in Equation 1 below.

$$s' = \frac{s-\mu}{\delta}, \tag{1}$$

where μ and δ are the mean and standard deviation of the input map. The fusion rules based on transformation include mean, minimum, and maximum rules. The mean value rule is to add all different saliency results and then take the average value as the final saliency prediction map. This calculation method is shown in Equation 2, where s'_i is the standardized confidence score, and S is the final confidence score.

$$S = \frac{1}{n}\sum_{i=1}^{n} s'_i \tag{2}$$

The minimum value fusion rule is to take the minimum value of all input saliency results as the final saliency prediction result, as shown in Equation 3.

$$S = min\left\{s'_1, s'_2 \ldots .s'_n\right\} \tag{3}$$

The maximum value fusion rule is to take the maximum value of all input saliency results as the final saliency prediction result, as shown in Equation 4.

$$S = max\left\{s'_1, s'_2 \ldots .s'_n\right\} \tag{4}$$

B. SCORE LEVEL FEATURE FUSION BASED ON CLASSIFICATION

The score of a single model is used as the feature vector of the classifier, and the feature vector is constructed to further improve the prediction accuracy of the classifier. There are two kinds of classifiers: linear and nonlinear. Linear classifiers are usually computationally fast, while nonlinear classifiers are usually slower but more powerful. In addition, for the input of the classifier, multiple combinations corresponding to each pixel in each image are set for each feature to be fused, and strong positive values and strong negative values are selected as classification criteria according to the sample of each pixel. Empirically, the first 5% and last 30% of the sample are usually considered as strong positive and strong negative values, respectively. The mean and unit standard deviation of the training vector are both zero after normalization, and the same parameters are used to normalize the test data.

The commonly-used classifiers are SVM and AdaBoost. Two public versions of SVM, namely liblinear and libsvm, are used in this paper to train SVM-based classifiers.

To further study the fusion ability of nonlinear classifiers, the AdaBoost algorithm can be used, which has been widely used in scene classification and object recognition. AdaBoost combines many weak classifiers to learn a strong classifier.

$$H(x) = sign(f(x)) = sign(\sum_{\tau=1}^{T} \alpha_\tau h_\tau(x)) \tag{5}$$

where α_τ is the weight of the classifier $h_\tau(x)$). Here the number of weak classifiers T is set to 10 to balance speed and accuracy.

C. SCORE LEVEL FEATURE FUSION BASED ON DENSITY

Probability density-based methods using well-known probabilistic models, such as Naive Bayes and GMM, have been widely used for model fusion. A scoring combination framework is proposed by Nandakumar et al. based on likelihood estimation [18]. The input score vector is modeled as a finite GMM. The results show that the density estimation method achieves good performance on biometric datasets such as the face, fingerprint, iris, and speech. However, it is worth noting that the method of score level fusion method based on probability density is highly dependent on the accuracy of the score level probability density estimation.

IV. PROPOSED METHOD

Current saliency prediction algorithms are basically aimed at natural images, but the features of natural images and advertising images are different. As shown in Figure 1, the left side is advertising images, and the right side is natural scene images. We can find that natural images contain a lot of structure and texture elements, but related research [19] shows that in advertising images, in addition to structure and texture elements, text and brand elements also occupy a large proportion of space, as shown in the three sub-images in the rectangular box. Therefore, the saliency prediction algorithm trained on natural images has a strong ability to extract pattern information, but it is not sensitive to text and brand information, and the saliency prediction performance will be reduced when applied to advertising images.

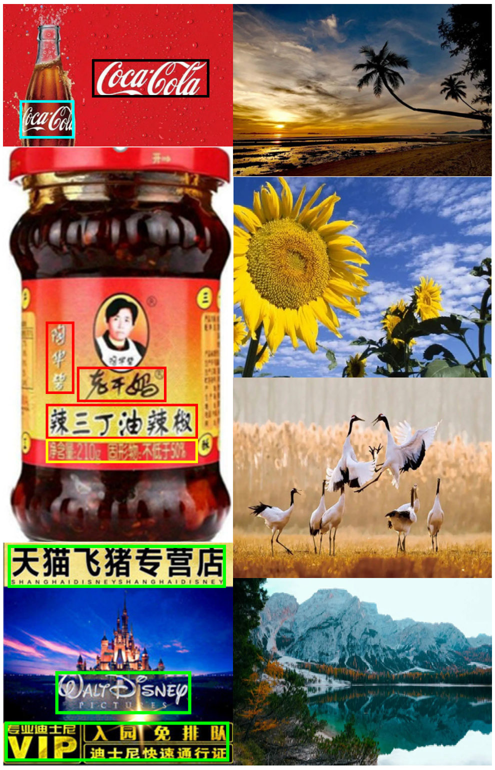

FIGURE 1. Comparisons between advertisement images and natural scene images.

Method [20] shows that text enhancement based on the saliency prediction algorithm of natural images can significantly improve the performance of the algorithm on advertising images. Therefore, on the basis of the lightweight natural image saliency prediction method, i.e. AMC-SNet, the score level feature fusion method is used to fuse the unique feature information of advertisements to improve the saliency prediction performance of advertising images. The model structure is shown in Figure 2.

Among the text elements and brand elements in advertising images, the text elements have obvious features, but the brand elements are often highly designed and innovative and do not have obvious and unified features. However, brand elements are often composed of patterns and text information. Therefore, in this algorithm, we still adopt text element features to improve the performance of the algorithm, instead of extracting brand or logo features separately.

Humans have a strong centrality when viewing natural images. Research in advertising image saliency dataset ADD1000 [20] proves that human gaze on advertising images also has obvious center bias. In machine vision, this imaging center is called the extended focus or perspective projection center. In both traditional saliency algorithms and

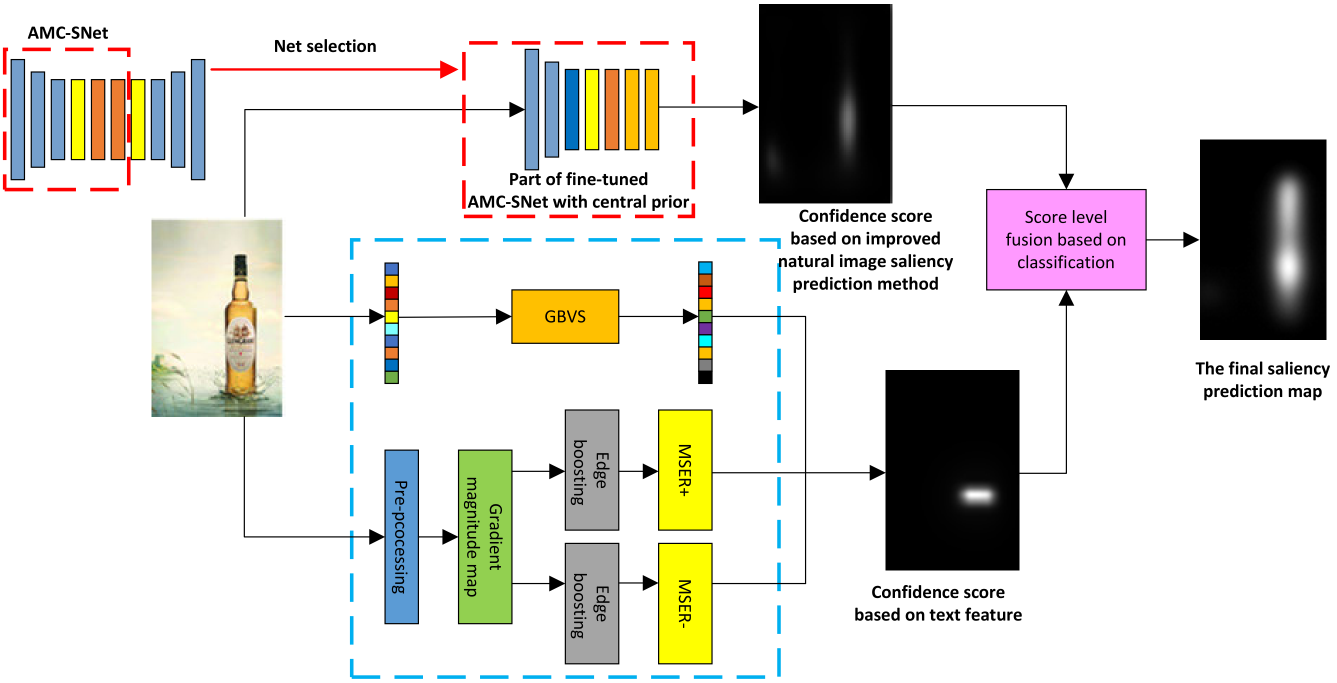

FIGURE 2. Method architecture of this paper. Each functional part of the method is marked with different colors. The red rectangle dotted box is the image feature extraction part, and the blue rectangle dotted box is the text feature extraction part.

depth-based saliency algorithms, central prior is often added into some models as a high-level prior feature. In this paper, the central prior is considered in feature fusion to further improve the model performance.

The score level fusion strategy adopted in this paper are SVM-based classifers. Here, two classifiers, namely liblinear and libsvm, are trained using the publicly available Matlab version of SVM. Linear kernel and nonlinear kernel are used, and the used nonlinear kernel is Radial Basis Function (RBF), see Equations 6 and 7, where α and γ are parameters of the kernel functions, and x_i/x_j are the eigenvectors of the fractional scale. They have been demonstrated to have good performance in a wide range of image applications. During testing, instead of predicting binary labels, labels with values in the range [0,1] are generated, so the final output is a saliency probability map with a continuous distribution.

$$\mathcal{K}_{linear}(x_i, x_j) = \alpha x_i^{\top} x_j \tag{6}$$

$$\mathcal{K}_{nonlinear}(x_i, x_j) = e^{-\gamma ||x_i - x_j||_2^2} \tag{7}$$

A. TEXT SALIENCY CONFIDENCE SCORE EXTRACTION BASED ON IMAGE INTENSITY FEATURES

In this paper, to reduce the complexity of the overall algorithm and further optimize the real-time performance of advertising image saliency, the traditional algorithm is used to extract text saliency. Since the purpose of this paper is to finally get the saliency region, it is not necessary to accurately box the text region, so the text candidate region is taken as the text saliency confidence score. A combination of traditional saliency and text region detection algorithm is used to extract text saliency, and paper [20] has proved that this mode can improve the accuracy of text detection. Relevant study [21] has shown through experiments that, among the three features of color, intensity, and direction extracted by the ITTI algorithm, saliency maps based on intensity features are used to characterize text candidate regions. Since the ultimate goal of this algorithm is to get a saliency map, text region is only fittingly used as an intermediate feature.Subsequently, to obtain text saliency confidence score with higher accuracy, on the basis of the extracted intensity features, the Graph Based Visual Saliency algorithm, namely GBVS, is used to generate activation maps to further optimize text saliency. Finally, the confidence score is obtained and input into the SVM-based classifer.

The intensity feature-based text saliency detection method used in this paper consists of three steps: 1. obtaining the multi-scale intensity feature map $\mathcal{M}_F$. 2. obtaining the activation map $\mathcal{M}_A$ based on the intensity feature map. 3. obtaining the final saliency map $\mathcal{M}_S$ by normalized activation map. Three color channels, i.e. r, g, and b of the input image, are used to generate the color Gaussian pyramid. On this basis, Equation 8 is used to output the intensity pyramid Ip.

$$Ip = \frac{r + g + b}{3} \tag{8}$$

Since small-scale images will lose more detailed features, which are more prominent for background features, and are not friendly to the saliency detection of text regions, three scales, i.e. 480 × 640, 240 × 320 and 120 × 160, are obtained through Gaussian downsampling. Thus, an intensity pyramid containing three scales are obtained. Then, the intensity

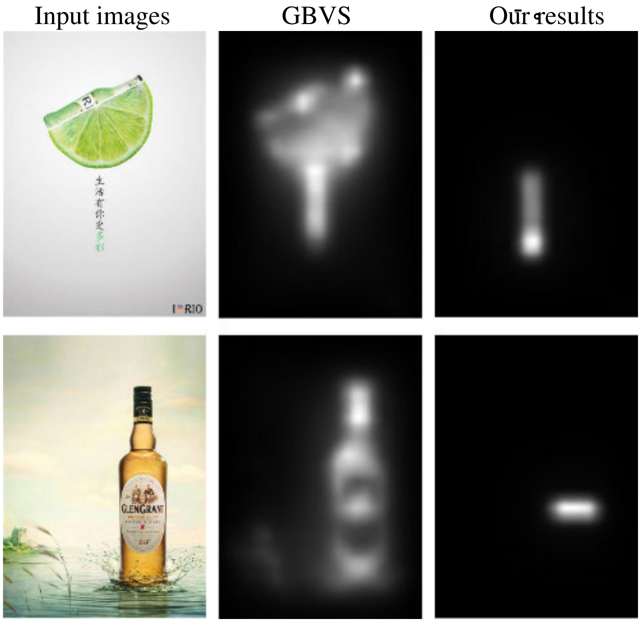

FIGURE 3. **Visualization results of saliency maps based on intensity features. Above each column of images is the name of the algorithm that generate them.**

feature map $\mathcal{M}_F$ is obtained by the center surrounding method.

The fully connected directed graph is used to construct the activation map. The edge weight $w((i,j),(p,q))$ between two nodes, i.e. (i,j) and (p,q), in a feature map $\mathcal{M}_F$ is defined in Equation 9, where δ is a free parameter.

$$w((i,j),(p,q)) = |log_{10}\frac{\mathcal{M}_F(i,j)}{\mathcal{M}_F(p,q)}|e^{-\frac{(i-p)^2+(j-q)^2}{2\delta^2}} \tag{9}$$

The activation map $\mathcal{M}_A$ of the feature map $\mathcal{M}_F$ can be obtained by traversing the operation with this method, and the saliency map $\mathcal{M}_S$ can be obtained by normalizing the activation map $\mathcal{M}_A$. Saliency information will preferentially flow to nodes with high activation. This method can effectively concentrate the saliency information and finally get the desired saliency map. Figure 3 shows the visualization process map. The middle column is the saliency nap obtained by GBVS [23] method, and the last column is the text saliency map obtained by our method. It can be found that the saliency map based on intensity features is obviously more focused on the text area.

B. TEXT SALIENCY CONFIDENCE SCORE EXTRACTION BASED ON IMPROVED MSER ALGORITHM

The Maximally Stable Extremal Regions (MSER) algorithm is a traditional and classical text detection method. The input image I is grayed, and then a random number from 0 to 255 is selected as the threshold to binarize the grayscale image I_{gray}. In our method, the binarization operation is canceled, because the saliency map $\mathcal{M}_S$ of the image is continuous. After the threshold is selected, the pixels above the threshold are retained, and the pixels in other positions are directly blackened. As the threshold changes, the Maximally Stable Extremal Region, MSER+, is reached when the retained region does not change. However, for images with bright characters on dark background and dark characters on a bright background, the grayscale image should be reversed and then the above operations should be carried out to obtain the Maximally Stable Extremal Region, MSER−.

Due to the poor accuracy of the traditional MSER algorithm in text detection, some researchers put forward an improved MSER algorithm [22]. Since the final purpose of the task in this paper is inconsistent with text detection, the subsequent steps, namely obtaining text candidate regions through a multi-mechanism suppression strategy, are ignored, and MSER candidate regions are directly taken as the confidence score.

The influence of noise is smoothed from the input image I in the preprocessing stage, and I is grayed to get I_{gray} by using the weights in Equation 10, where R, G, B are its three channels.

$$I_{gray} = 0.299 \times R + 0.587 \times G + 0.114 \times B \tag{10}$$

On this basis, gradient operation is carried out to obtain the gradient map $\mathcal{M}_G = \nabla I_{gray}$, and the edge enhancement operation enhances the gradient map, and the specific realization method is shown in Equations 11 and 12. Here, $\mathcal{M}_{EB1}$ and $\mathcal{M}_{EB2}$ are edge boosting maps, β ranges from 0 to 1, and ∇ is the gradient operator.

$$\mathcal{M}_{EB1} = I_{gray} + \beta \times \mathcal{M}_G \tag{11}$$
$$\mathcal{M}_{EB2} = I_{gray} - \beta \times \mathcal{M}_G \tag{12}$$

Then, the MSER+ and MSER− regions are obtained from the grayscale image after edge boosting, and the two regions are fused to obtain the text candidate region, which is used as the confidence score of the text feature to further improve the saliency performance on the advertising image.

Finally, by using the score level feature fusion strategy based on classification, the above obtained two text confidence scores based on intensity feature and improved MESR algorithm are further fused with a saliency confidence score to get the final saliency prediction map. Moreover, it is worth mentioning that the saliency confidence score is a 16-dimensional high-level feature map, which can be extracted by the fine-tuned AMC-SNet with central prior.

V. EXPERIMENTAL SETTINGS

A. DATASETS

The saliency algorithm in this paper is aimed at advertising images, so ADD1000 [20], the saliency dataset of advertising images, is used in this paper. This dataset contains 1000 advertising images, which are divided into four dimensions: (calm type, exciting type), (retro type, fashion type), (artistic type, practical type), and (emotional type, rational type), with a total of eight attributes. The advertising categories in the dataset include wine, tea, home appliances, cars, food, cosmetics, beverages, electronics, real estate,

TABLE 1. Quantitative tests. Combination of six metrics recommended by the MIT Public benchmark as evaluation metrics. All algorithms run on the dataset named ADD1000. The best and second best results are shown in black bold and **blue bold**.

Method	AUC	sAUC	NSS	CC	KLD	SIM
EML-NET [24]	**0.871**	0.724	**2.771**	0.765	0.870	0.696
MSI-NET [13]	0.870	**0.752**	2.510	0.753	**0.460**	0.669
SAM-VGG [25]	**0.873**	0.716	2.26	0.669	1.220	0.607
SALICON [33]	0.853	0.706	**2.673**	**0.767**	0.520	**0.709**
GazeGAN [26]	0.830	0.714	2.462	0.736	1.390	0.644
SALGAN [11]	0.807	0.645	2.651	0.690	0.770	0.602
DeepGaze II [10]	0.806	0.643	2.457	0.630	**0.430**	0.567
DVA [27]	0.806	0.705	2.130	0.607	1.280	0.423
ML-NET [28]	0.805	0.720	2.094	0.613	0.820	0.569
eDN [8]	0.801	0.613	1.127	0.452	1.210	0.406
GBVS [23]	0.817	0.586	2.266	0.466	0.900	0.569
AIM [29]	0.786	0.611	2.212	0.517	1.270	0.535
SR [30]	0.779	0.529	1.993	0.432	1.520	0.514
SDSP [31]	0.756	0.544	2.069	0.366	1.550	0.512
ITTI [2]	0.752	0.561	1.983	0.425	1.430	0.533
SUN [32]	0.751	0.540	2.235	0.463	1.560	0.509
Our method	**0.871**	**0.725**	2.664	**0.773**	1.290	**0.715**

clothing, and other categories. Eye gaze data of 57 different personality observers are used as the ground truth.

B. EVALUATION METRICS

Previous studies on saliency evaluation show that using multiple metrics can improve the fairness of evaluation. Therefore, to evaluate the saliency model fairly, the model evaluation results reported by the saliency benchmark dataset are based on various saliency evaluation metrics. In this paper, metrics from the MIT saliency benchmark report are used, including KL Divergence (KLD), Pearson's Correlation Coefficient (CC), structure SIMilarity (SIM), Normalized Scanpath Saliency (NSS), and Area Under the receiver operating characteristic Curve (AUC) and its variants named shuffled AUC (sAUC).

C. EXPERIMENTAL DETAILS

The simulation experiments in this paper are conducted based on MATLAB R2016b platform, the system CPU is Intel(R) Core(TM) i7-7700HQ, whose main frequency is 2.80GHz, and the system memory capacity is 8GB RAM. The new method of lightweight natural image salience AMC-SNet is first pre-trained on the SALICON dataset, and then fine-tuned on the advertising dataset. For the 1000 images in the advertising dataset, 800 images are randomly selected as the training set, 100 images are selected as the validation set, and the remaining 100 images are selected as the test set. Since the distribution of image classes in the dataset is very uniform, the algorithm proposed in this paper does not use cross-validation. The advertising image is not a fixed size and the ratio of width and height is not consistent, to meet the fine-tuning training requirements, the advertising image is padded until the width and height ratio are consistent with 480 : 640, then the image is scaled to the required size. After fine-tuning, a new model is acquired, through which the confidence scores are obtained with advanced saliency features.

For score level fusion, we have three types of input: the confidence score of fine-tuned advanced saliency feature, the confidence score of the intensity-based text feature, and the confidence score of the text feature based on the improved MSER algorithm. For advanced saliency features, AMC-SNet is used, but to make better use of multi-dimensional saliency features, the last 1 × 1 convolutional layers in AMC-SNet are not considered, and 16-dimensional advanced feature maps for each image are treated as the output. For the other two text feature confidence scores, each corresponds to a feature map, so the input of SVM contains 18-dimensional feature information.

To train and test our model, the ADD1000 dataset is divided into 900 training images and 100 test images. From each image, we randomly selected 10 positively labeled pixels from the top 20% saliency position of the human gaze truth saliency map and 10 negatively labeled pixels from the bottom 70% saliency position to obtain a training set with 18000 samples and a test set with 2000 samples.

D. EXPERIMENTAL COMPARISONS

As a saliency method for advertising images, we only conduct objective and subjective comparison experiments on the saliency dataset of advertising images. In the comparison experiments, the saliency algorithm in this paper is compared with other current representative saliency algorithms based on deep learning and traditional saliency algorithms in various forms, and a combination of six metrics recommended by the MIT Public Benchmark is used as the evaluation metrics. To prove the effectiveness of score level fusion of advertising image features proposed in this paper, ablation experiments are conducted.

Table 1 shows that our model achieves the highest performance in AUC, CC, and SIM. In the other three metrics, it also achieves a relatively high ranking, indicating that our model has a competitive and excellent performance in the

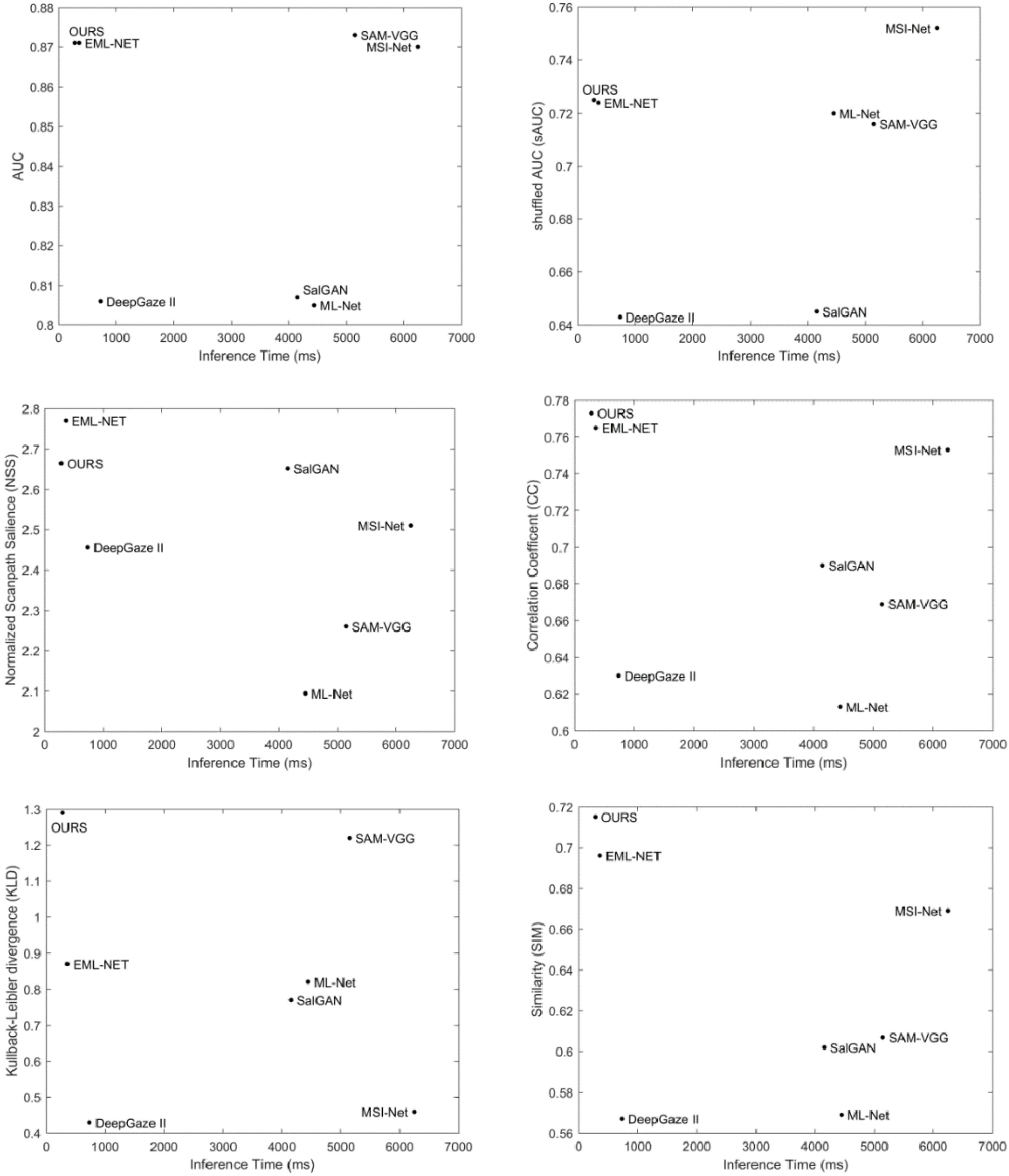

FIGURE 4. Comparisons of inference time and evaluation metrics of different algorithms.

saliency model on advertising images. Conclusion can be drawn from Table 2, i.e. the inference time of our model is much shorter than that of the deep model. Although various evaluation metrics of EML-Net are also outstanding, the model size is very large due to cascading multiple deep convolutional neural networks, which have high requirements for the hardware equipment, thus is limited for the advertising image application. The applicability of EML-Net is not as

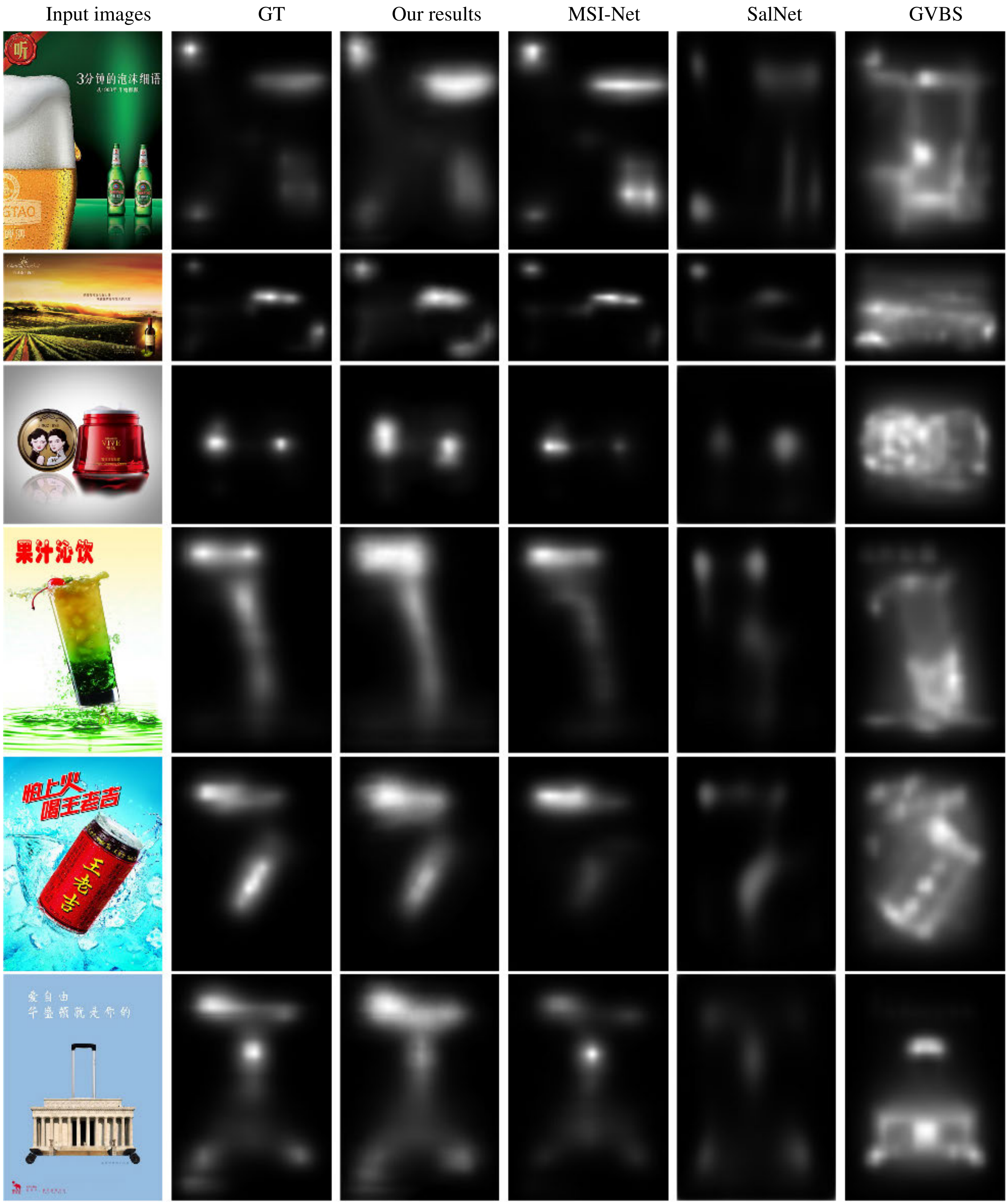

FIGURE 5. Comparisons of subjective performance. Above each column of images is the name of the algorithm that generate them.

strong as our model. Moreover, Table 2 shows that the inference time of EML-Net is about 22 times longer than the proposed algorithm. In comparison, our model can better meet the requirements of real-time performance. Therefore, it can be seen that compared with other models, our model achieves both accuracy and efficiency for advertising images.

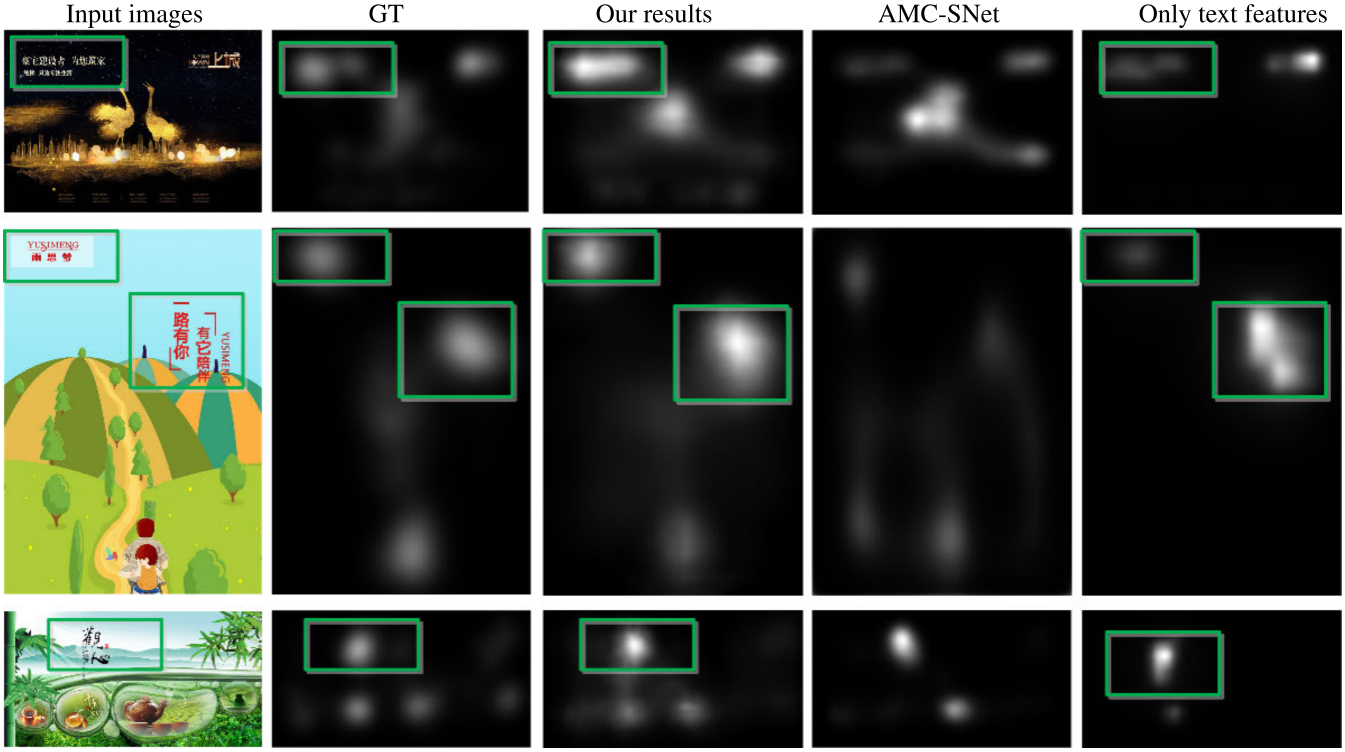

FIGURE 6. Comparisons of subjective performance of ablation experiments. Above each column of images is the name of the algorithm that generate them.

TABLE 2. Comparisons of the inference time. The best and second best results are shown in black bold and **blue bold**.

Method	Inference time (ms)
MSI-NET [13]	6251
SAM-VGG [25]	5147
ML-NET [28]	4448
SalGAN [11]	4152
DeepGazeII [10]	737
EML-NET [24]	**359**
Our method	**283**

Figure 4 puts evaluation metrics and inference time in six scatter plots, which show the advantages of our algorithm in both performance and efficiency more clearly. In each scatter plot, the abscissa is the inference time, and the ordinate is the evaluation metric value. Except for the metric KLD, the larger the value of other metrics, the better. Therefore, except for the metric KLD, the algorithm in the upper left corner of the scatter plot has superior performance. As can be seen, our model does not perform well in the KLD metric, but it is in the top left corner in other metrics, which intuitively indicates that it has achieved competitive and robust results in saliency performance and inference time.

Objective data prove that our model is superior in both performance and efficiency. Figure 5 shows the subjective performance of our model. It can be seen intuitively from Figure 5 that compared with the deep saliency algorithm MSI-Net, SalGAN, and the traditional saliency algorithm GBVS, our method is more accurate in the saliency extraction of the text elements in advertising images, and closer to the truth saliency map viewed by human eyes.

E. ABLATION EXPERIMENTS

To verify the effectiveness and importance of fine-tuning model and central prior for advertising image saliency prediction, ablation experiments are conducted. Table 3 shows that the fine-tuned model has improved AUC, sAUC, KLD, and SIM compared with the pre-trained model, i.e. AMC-SNet. Both AUC and sAUC increased by 1% to 2% and SIM increased by 7%. The calculation method of growth rate is shown in Equation 13, where gr means the growth rate, $data_{ours}$ represents the data obtained by our algorithm, and $data_{AMC-SNet}$ represents the data of the pre-trained AMC-SNet in the same test environment. Our model, namely fine-tuned AMC-SNet with central prior, is effective and has a significant effect on improving the saliency prediction performance of advertising images.

$$gr = \frac{data_{ours} - data_{AMC-SNet}}{data_{AMC-SNet}} \tag{13}$$

Table 4 shows that the model based on score level fusion is faster than the lightweight saliency network AMC-SNet, and the running time is about one-third of it, which greatly improves the efficiency of the system.

The above figures and tables, have proved the effectiveness of introducing the saliency features of natural images into advertising images by means of score level fusion.

TABLE 3. Comparisons of results in ablation experiments The best and second best results are shown in black bold and blue bold.

Method	AUC	sAUC	NSS	CC	KLD	SIM
Pre-trained AMC-SNet [17]	0.856	0.717	**2.763**	0.755	**1.35**	0.668
Fine-tuned AMC-SNet without central prior	**0.864**	**0.726**	2.663	**0.771**	**1.29**	**0.690**
Fine-tuned AMC-SNet with central prior	**0.871**	**0.725**	**2.664**	**0.773**	**1.29**	**0.715**

TABLE 4. Comparisons of the inference time.

Method	Inference time (ms)
AMC-SNet	892
Our method	283

In addition, the results of the ablation experiments are shown in Figure 6, and the green rectangular boxes are used to demonstrate the visual superiority of our algorithm. As can be seen in Figure 6, for the special case of advertisement images, text features are taken into account to make the final results closer to the ground truth saliency maps.

VI. CONCLUSION

In this paper, a saliency prediction algorithm for advertising images based on score level fusion is proposed. In the proposed method, the image intensity features and the improved MSER algorithm are applied to obtain the text candidate regions. Then, the corresponding confidence score of the two text candidate regions is integrated to produce a two-dimensional text confidence score, which can better characterize the advertising image features. Subsequently, the final text confidence score are further fused with the saliency confidence score obtained by an improved natural image saliency prediction network by the score level fusion strategy. Experimental data show that the proposed method improves the accuracy and robustness of the saliency prediction algorithm of advertising images. Moreover, the inference speed of the proposed method is significantly faster than that of all the algorithms compared, which meet the commercial promotion and application needs of real-time interaction. In future work, to facilitate the loading and deployment of the proposed method, we plan to do further lightweight processing in the pre-processing and decoding stage, and consider building an end-to-end saliency model for advertising images.

VII. DECLARATION OF COMPETING INTEREST

The authors declare that they are not aware of the possibility of competing for financial interests or personal relationships affecting the work reported in this paper.

REFERENCES

[1] D. Cheng, H. Zhang, M. Jiang, and Q. Kou, "Color image retrieval method fusing principal curvature and color information," *J. Comput.-Aided Design Comput. Graph.*, vol. 33, no. 2, pp. 223–231, Feb. 2021.

[2] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[3] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Matters Intell.*, vol. 188, pp. 115–141, Apr. 1987.

[4] D. Cheng, J. Xu, and Q. Kou, "Lightweight network based on residual information for foreign body classification on coal conveyor belt," *J. China Coal Soc.*, vol. 47, no. 3, pp. 1361–1369, Mar. 2022.

[5] J. Li, D. Cheng, R. Liu, Q. Kou, and K. Zhao, "Unsupervised person re-identification based on measurement axis," *IEEE Signal Process. Lett.*, vol. 28, pp. 379–383, 2021.

[6] K. Zhao, D. Cheng, Q. Kou, J. Li, and R. Liu, "Sequences consistency feature learning for video-based person re-identification," *Electron. Lett.*, vol. 58, no. 4, pp. 142–144, Feb. 2022.

[7] N. Zhang, D. Cheng, Q. Kou, H. Ma, and J. Qian, "Person re-identification based on random occlusion and multi-granularity feature fusion," *J. Beijing Univ. Aeronaut. Astronaut.*, vol. 389, pp. 1–12, Apr. 2022, doi: 10.13700/j.bh.1001-5965.2022.0091.

[8] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 2798–2805.

[9] M. Kümmerer, L. Theis, and M. Bethge, "Deep gaze I: Boosting saliency prediction with feature maps trained on ImageNet," 2014, *arXiv:1411.1045*.

[10] M. Kummerer, T. S. A. Wallis, L. A. Gatys, and M. Bethge, "Understanding low- and high-level contributions to fixation prediction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 4799–4808.

[11] J. Pan, C. Canton Ferrer, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. Giro-i-Nieto, "SalGAN: Visual saliency prediction with generative adversarial networks," 2017, *arXiv:1701.01081*.

[12] J. Lou, H. Lin, D. Marshall, D. Saupe, and H. Liu, "TranSalNet: Towards perceptually relevant visual saliency prediction," 2021, *arXiv:211003593*.

[13] A. Kroner, M. Senden, K. Driessens, and R. Goebel, "Contextual encoder–decoder network for visual saliency prediction," *Neural Netw.*, vol. 129, pp. 261–270, Sep. 2020.

[14] N. Reddy, S. Jain, P. Yarlagadda, and V. Gandhi, "Tidying deep saliency prediction architectures," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Las Vegas, NV, USA, Oct. 2020, pp. 10241–10247.

[15] F. Hu and K. McGuinness, "FastSal: A computationally efficient network for visual saliency prediction," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Milan, Italy, Jan. 2021, pp. 9054–9061.

[16] S. Zabihi, H. R. Tavakoli, A. Borji, and E. Mansoori, "A compact deep architecture for real-time saliency prediction," *Signal Process., Image Commun.*, vol. 104, May 2022, Art. no. 116671.

[17] D. Cheng, R. Liu, J. Li, S. Liang, Q. Kou, and K. Zhao, "Activity guided multi-scales collaboration based on scaled-CNN for saliency prediction," *Image Vis. Comput.*, vol. 114, Oct. 2021, Art. no. 104267.

[18] K. Nandakumar, Y. Chen, S. C. Dass, and A. K. Jain, "Likelihood ratio-based biometric score fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 342–347, Feb. 2008.

[19] Z. Ma, L. Qing, J. Miao, and X. Chen, "Advertisement evaluation using visual saliency based on foveated image," in *Proc. IEEE Int. Conf. Multimedia Expo.*, New York, NY, USA, Jun. 2009, pp. 914–917.

[20] S. Liang, R. Liu, and J. Qian, "Fixation prediction for advertising images: Dataset and benchmark," *J. Vis. Commun. Image Represent.*, vol. 81, Nov. 2021, Art. no. 103356.

[21] D. Wang, R. Cui, and J. Jin, "Text detection in natural scene based on visual attention model and multi-scale MSER," *J. Appl. Sci.*, vol. 38, no. 3, pp. 496–506, May 2020.

[22] Y. Li, T. Quan, and W. Liu, "Adaptive learning text detection in the natural scene based on MSER," *J. Chinese Comput. Syst.*, vol. 41, no. 9, pp. 1966–1971, Sep. 2020.

[23] B. Schölkopf, J. Platt, and T. Hofmann, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2007, pp. 545–552.

[24] S. Jia and N. Bruce, "EML-NET: An expandable multi-layer network for saliency prediction," *Image Vis. Comput.*, vol. 95, Mar. 2020, Art. no. 103887.

[25] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting human eye fixations via an LSTM-based saliency attentive model," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5142–5154, Oct. 2018.

[26] Z. Che, A. Borji, G. Zhai, X. Min, G. Guo, and P. Le Callet, "How is gaze influenced by image transformations? Dataset and model," *IEEE Trans. Image Process.*, vol. 29, pp. 2287–2300, 2020.

[27] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2368–2378, May 2018.

[28] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "A deep multi-level network for saliency prediction," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Cancun, Mexico, Dec. 2016, pp. 3488–3493.

[29] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 155–162.

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

[31] L. Zhang, Z. Gu, and H. Li, "SDSP: A novel saliency detection method by combining simple priors," in *Proc. IEEE Int. Conf. Image Process.*, Melbourne, VIC, Australia, Sep. 2013, pp. 171–175.

[32] M. Kümmerer, T. S. A. Wallis, and M. Bethge, "Information-theoretic model comparison unifies saliency metrics," *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 52, pp. 16054–16059, Dec. 2015.

[33] X. Huang, C. Shen, X. Boix, and Q. Zhao, "SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 262–270.

QIQI KOU received the B.S. and M.S. degrees from the Anhui University of Science and Technology, in 2012 and 2015, respectively, and the Ph.D. degree from the School of Information and Control Engineering, China University of Mining and Technology, in 2019. He is currently an Assistant Professor with the School of Computer Science and Technology, China University of Mining and Technology. His research interests include image processing, computer vision, and pattern recognition.

RUIHANG LIU received the B.S. and M.S. degrees from the China University of Mining and Technology, in 2016 and 2022, respectively. She is currently a Teaching Assistant with the Xuhai College, China University of Mining and Technology. Her main research interest includes image saliency detection.

CHEN LV received the B.S. degree from Shanxi University, in 2017, and the M.S. degree in electrical and information engineering from the China University of Mining and Technology, in 2021, where he is currently pursuing the Ph.D. degree with the School of Information and Control Engineering. His main research interests include pattern recognition and image processing.

HE JIANG received the B.S. degree in telecommunication engineering from the Nanjing University of Post and Telecommunications, and the M.S. degree in telecommunication engineering and the Ph.D. degree in control science and engineering from Shanghai Jiao Tong University, in 2021. He is currently working as an Assistant Professor with the School of Information and Control Engineering, China University of Mining and Technology. His research interests include machine learning and deep learning-based vision tasks, such as image restoration, image retrieval, and saliency detection.

DEQIANG CHENG (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical and information engineering from the China University of Mining and Technology. He is currently a Professor with the China University of Mining and Technology. His research interests include machine learning, video coding, image processing, and pattern recognition.