How is Visual Attention Influenced by Text Guidance? Database and Model

Yinan Sun, Xiongkuo Min*, *Member, IEEE,* Huiyu Duan, and Guangtao Zhai*, *Senior Member, IEEE*

***Abstract*—The analysis and prediction of visual attention have long been crucial tasks in the fields of computer vision and image processing. In practical applications, images are generally accompanied by various text descriptions, however, few studies have explored the influence of text descriptions on visual attention, let alone developed visual saliency prediction models considering text guidance. In this paper, we conduct a comprehensive study on text-guided image saliency (TIS) from both subjective and objective perspectives. Specifically, we construct the first TIS database named SJTU-TIS, which includes 1200 text-image pairs and the corresponding collected eye-tracking data. Based on the established SJTU-TIS database, we analyze the influence of various text descriptions on visual attention. Then, to facilitate the development of saliency prediction models considering text influence, we construct a benchmark for the established SJTU-TIS database using state-of-the-art saliency models. Finally, considering the effect of text descriptions on visual attention, while most existing saliency models ignore this impact, we further propose a text-guided saliency prediction model (TGSal), which extracts and integrates both image features and text features to predict the image saliency under various text-description conditions. Our proposed model significantly outperforms the state-of-the-art saliency models on both the SJTU-TIS database and a generic saliency database (SALICON) in terms of various evaluation metrics. The SJTU-TIS database and the code of the proposed TGSal model will be released to facilitate further research.**

***Index Terms*—Text guidance, visual attention, image saliency, multimodal fusion.**

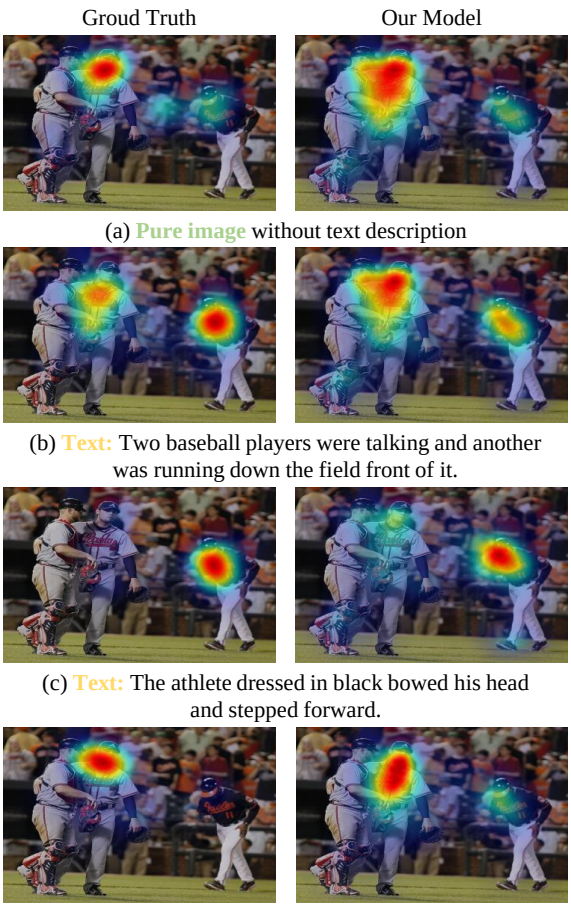
I. Introduction

HUMAN vision has the ability to select informative and conspicuous regions from external visual stimuli and attend to them, which is well known as the visual attention mechanism [1] [2]. Human vision attention can be categorized into two functions including scene-driven bottom-up (BU) and expectation-driven top-down (TD) [2], and many studies have demonstrated that eye movements are driven by the joint influence of BU and TD attention [3].

Visual attention analysis and prediction have been important tasks in multimedia and computer vision research for a long time, since they can provide new insights into the mechanisms of human attention [4], [5], and contribute to many multimedia applications [6], [7] as well as various computer vision tasks [8]–[14].

Many studies have explored the scene-driven bottom-up visual attention problem, and many corresponding saliency databases have been constructed, such as SALICON [15], MIT1003 [16], MIT300 [17], and CAT2000 [18], which are all pure image saliency databases. Based on these databases, many saliency prediction models have been proposed, which mainly include traditional handcrafted feature-based methods and deep neural network-based (DNN-based) methods. Traditional saliency prediction methods extract low-level features such as color, contrast, and semantic concepts, *etc.*, and integrate these features to generate saliency maps [19]–[21]. With the development of deep learning, many DNN-based models have also been proposed for predicting image saliency [22]–[28]. The databases and models mentioned above mainly study the scene-driven bottom-up visual attention problem, *i.e.,* the pure image saliency prediction task, however, the expectation-driven top-down visual attention task has been rarely studied.

In daily life, images are often accompanied by text descriptions, such as image captions, subtitles, audio commentary, *etc.* Since these text descriptions are strong expectation guidance, it



(a) **Pure image** without text description

(b) **Text:** Two baseball players were talking and another was running down the field front of it.

(c) **Text:** The athlete dressed in black bowed his head and stepped forward.

(d) **Text:** Two baseball players were talking.

Fig. 1. The 1_{st} column: heatmaps of the human gaze on the original image, and the images with three different text descriptions. The 2_{nd} column: corresponding prediction results of our model.

*Corresponding authors.

is intuitive that the human visual attention to the corresponding images will be influenced by these expectations through the top-down mechanism.

As shown in the first column of Fig. 1, when viewing the original pure image without any text description, human attention is highly attracted to the "baseball player", however, when viewing the image accompanied by different text descriptions, the human gaze is significantly changed according to the context. Thus, it is obvious that text descriptions can significantly influence the corresponding visual attention to visual stimuli. However, to the best of our knowledge, most of current saliency models, either early hand-crafted models or recent DNN models, are not able to predict the corresponding visual saliency according to different text descriptions. Therefore, it is important to investigate new robust approaches to effectively predict human visual attention in scenes with text descriptions.

In this work, we aim to thoroughly analyze human visual attention behavior under the influence of various text descriptions and build an accurate saliency prediction model for text-guided conditions. To achieve this objective, we are facing the following research challenges.

(i) Building a database for text-image saliency. Although there are many publicly available image saliency databases, such as SALICON [15] and MIT300 [17], they are all pure images without text descriptions. In addition, some other databases, such as MSCOCO [29] and Flickr30k database [30], contain images with text descriptions, but there is no corresponding ground truth visual attention data.

(ii) Understanding the effect of various text descriptions on visual saliency. As a common observation, without text descriptions, the visual attention mechanism will make people pay more attention to the informative and conspicuous areas of an image. Moreover, a text description can influence the visual attention on the image [31]. However, whether and how different text descriptions of one image influence the corresponding visual attention is still unknown.

(iii) Modeling text-image saliency. Since images are usually accompanied by texts in daily life and different descriptions cause different influences on the corresponding visual attention. It is necessary and significant to study how to integrate both image features and text features, and jointly exploit these two parts of information to build an accurate text-image saliency model.

In this work, we first construct a text-guided image saliency database, termed SJTU-TIS. Specifically, as shown in Fig. 2, in order to investigate whether a text description can influence visual attention on an image, the eye tracking experiment is conducted under two conditions including a pure image condition and a text-guided condition. Moreover, as shown in Fig. 3, in order to investigate how different text descriptions influence the corresponding visual attention, the collected 600 images are divided into two parts, including 300 images with general scenario descriptions and 300 images with three different types of object descriptions. Overall, our SJTU-TIS database has 600 images and 1200 text descriptions, which results in 1200 text-image pairs. To better predict the visual attention influenced by a text description, we propose a novel

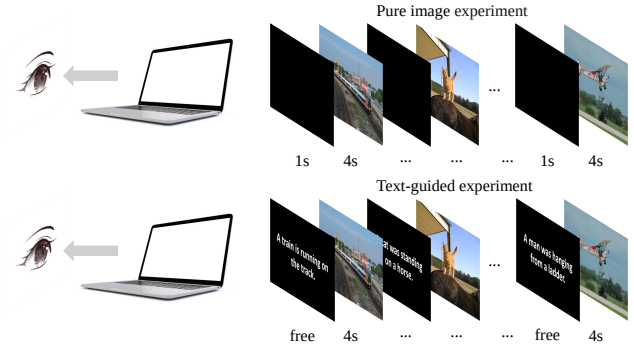


Fig. 2. Schematic diagram of our eye-tracking experiment. Comparison between pure image condition and text-guided condition is shown.

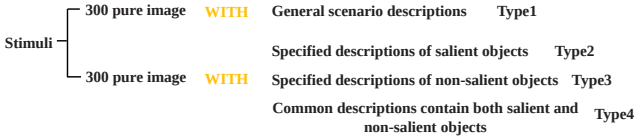


Fig. 3. Classification of the attributes of the texts.

text-guided saliency prediction model (TGSal) that can be used to predict visual saliency under both pure image and text-guided conditions. Specifically, we encode the text features and image features into the embedding space and inject the text features to image features step by step at the decoding end to get the final prediction results. Experimental results demonstrate the effectiveness of the proposed text feature fusion modules on the text-guided saliency prediction task, and our proposed TGSal achieves the best performance on both the generic saliency prediction dataset SALICON [15] and our constructed SJTU-TIS database. The contributions are summarized as follows:

- We build the first text-guided image saliency database, termed SJTU-TIS, which studies the visual saliency of one image under the influence of different text descriptions.
- We analyze the effects of different text descriptions on visual attention, and indicates that image saliency is significantly influenced by a text description, and different text descriptions for the same image may have different impacts on the corresponding visual attention.
- A new benchmark has been established for the field of saliency prediction.
- A prediction model for text-image visual saliency is proposed, which achieves the best performance under both pure image and text-guided conditions.

The rest of the paper is organized as follows. Section II introduces related works. Then we introduce the construction procedure and the analysis of the SJTU-TIS database in Section III. Section IV describes the experimental settings and the experimental results. Section V concludes the whole paper.

II. Related Work

A. Eye-tracking Databases

1) Traditional Saliency Databases: In order to understand and model visual attention behavior, many eye-tracking

databases have been constructed under the free-viewing task. MIT1003 [16] is a large-scale saliency database that contains 1003 images coming from Flickr and LabelMe. MIT300 [17] and CAT2000 [18] are two widely used benchmark databases, which contain 300 and 2000 test images respectively. SALICON [15] is currently the largest crowd-sourced saliency database, which contains 10000 training images, 5000 validation images, and 5000 test images, which is widely adopted for pretraining saliency prediction models. The saliency data is collected through mouse tracking using Amazon Mechanical Turk (AMT).

2) Text-image Saliency Database: In our previous work [31], we have established a text-image saliency database, however, this database is just a pilot database, which only has one description for an image. This makes it hard to analyze the influence of different texts on image saliency. This paper is based on our previous paper [31]. Compared with the previous work, we have made four new contributions, which are mentioned in Section I.

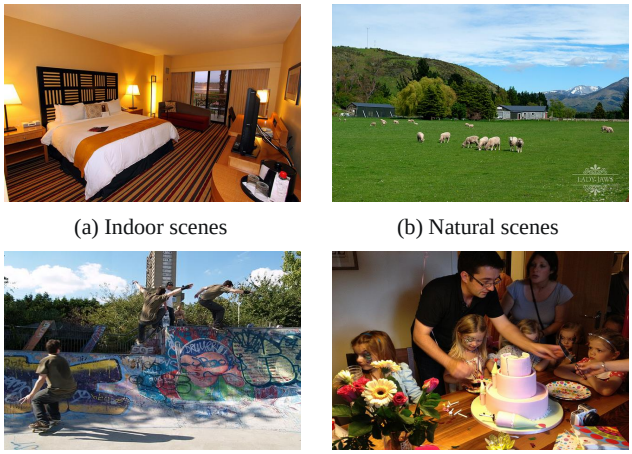
B. Saliency Prediction Models

1) Classical Models: Most traditional methods model visual saliency based on the bottom-up mechanism, which generally extract simple low-level feature maps, such as intensity, color, direction, *etc.*, and integrate them to generate saliency maps. Itti *et al.* [32] considered underlying features on multiple scales to predict saliency maps. Harel *et al.* [19] introduced a graph-based saliency model, which defined Markov chains on various image maps and regarded the balanced distribution of map locations as activation values and saliency values. Many other classical methods such as AIM [33], SMVJ [34], CovSal [35], SeR [36], HFT [37], *etc.*, are also commonly used saliency prediction models.

2) Deep Models: With the development of deep neural networks (DNN), saliency prediction tasks have made significant progress in recent years [38], [39]. Huang *et al.* [23] used VGG as the backbone and proposed a two-stream network to extract coarse features and fine features to calculate saliency maps. Cornia *et al.* [40] proposed an Attentive ConvLSTM, which focuses on different spatial positions of a bunch of features to enhance prediction. Pan *et al.* [41] proposed the generative adversarial network (GAN) to calculate the saliency map. Che *et al.* [42] studied the influence of transformation on visual attention and proposed a GazeGAN model based on the U-Net for saliency prediction. Duan *et al.* [1] proposed a vector quantized saliency prediction method and generalized it for AR saliency prediction. These top-down-based saliency prediction models have been widely used in various research fields in recent years [43].

III. SJTU-TIS Database

Due to the absence of a text-image saliency database, in this paper, we construct the first text-guided image saliency database, denoted as the SJTU-TIS database. We first select 600 images from MSCOCO [29] and Flickr30k [30] with the corresponding text descriptions. Then, a subjective experiment is conducted to obtain the eye movement data, which is processed to obtain the visual attention map of the SJTU-TIS database.



(a) Indoor scenes (b) Natural scenes

(c) Urban scenes (d) Party scenes

Fig. 4. Examples of the collected different scenes. (a) Indoor scenes. (b) Natural scenes. (c) Urban scenes. (d) Party scenes.

How the data is derived

A. Text-image Pair Collection

To collect images with diverse scenes, we first selected 4 scenarios from the MSCOCO database [29] and the Flickr30k database [30], including indoor scenes, natural scenes, urban scenes, and party scenes, as shown in Fig. 4. Although each image in the MSCOCO database [29] and the Flickr30k database [30] has multiple corresponding text descriptions, the semantic meanings of these text descriptions are similar, which does not meet the study requirements. Since our objective is to study the visual saliency of an image under different descriptions, we manually modified these text descriptions to four conditions as shown in Fig. 3. For natural scenes, since they usually do not include salient or non-salient objects, we only produce general scenario descriptions. For other scenes that have salient objects and non-salient objects, we produce three text descriptions for each image, including specified descriptions for salient objects, specified descriptions for non-salient objects, and common descriptions for both salient and non-salient objects. Therefore, our SJTU-TIS database contains 600 images and 1200 text descriptions, which results in 1200 text-image pairs (300+300×3) in total.

Fig. 5 shows some representative images from the SJTU-TIS database. The first row represents the first group of images (Type1), and the second to fourth rows demonstrate the second group of images (Type2, 3, and 4). The red rectangular box represents the specified descriptions of salient objects, the blue rectangular box represents the specified descriptions of non-salient objects, and the yellow rectangular box represents the common descriptions containing both salient and non-salient objects. It is obvious that different text descriptions correspond to different image areas, which may significantly influence the corresponding visual attention.

B. Subjective Eye-tracking Experiment

We conducted a subjective eye-tracking experiment to obtain the visual attention maps of the images in the SJTU-

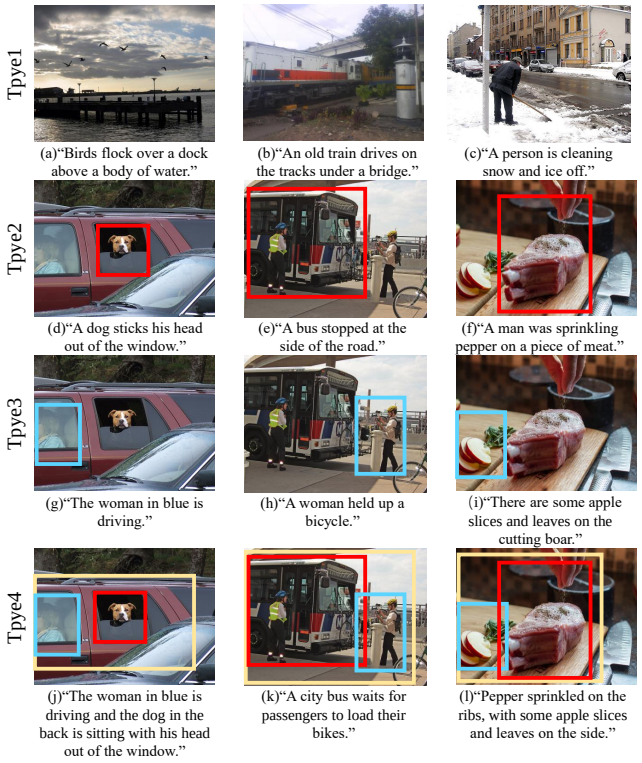


Fig. 5. An illustration of the example images with different text descriptions from 4 categories given in Fig. 3. Red boxes represent salient objects, blue boxes represent non-salient objects, and yellow boxes represent common descriptions contain both salient and non-salient objects.

TIS database using the Tobii Pro X3-120 eye-tracker [44]. Tobii Pro X3-120 is an ultra-thin and lightweight portable eye tracker that can be directly combined with various types of screens. The eye-tracking device can be directly connected to a computer through USB. We recorded eye movements at a sampling rate of 120Hz. The resolution of the screen is 1920×1080, and the screen resolution can be freely adjusted. The size of the head sports box is (wide×high) 50×40cm, 80cm in length, with a tracking distance of 50-90cm.

We designed the experimental process using the software provided by the Tobii Pro X3-120 eye tracker. The eye-tracking experiment is divided into 5 sessions. We set Type1 to Type4 into four sessions (each with 300 text-image pairs), and set all pure images into one session, which is the fifth session (containing 600 pure images). All images were displayed at their raw resolutions. To avoid fatigue, subjects had a break time after viewing every 100 images. As shown in Fig. 2, during the experiment, each image was displayed for 4 seconds. In the pure image condition, each image stimuli was followed by a 1-second black screen interval. In the text-guided condition, there was a text-description viewing while before displaying each image. The duration of viewing the text was controlled by the participants through the mouse click, *i.e.*, after reading and understanding the text, they can click the mouse to view the corresponding image. The display time of each image is still four seconds. After four seconds, the next text description was automatically shown. The distance between the subjects and the eye tracker was roughly maintained between the range of 2 and 2.3 feet during each session. The eye-tracking experiment was conducted in a quiet room.

A total of 60 subjects were recruited to participate in the experiment. Since subjects may remember an image if they have seen this image before, which may affect the reliability of the collected eye movement data, we ensured that each subject only viewed each image once. Specifically, the 60 subjects were divided into 4 groups with 15 subjects in each group, and the first group participated in the pure image experiment, the second group participated in the "Type1" and "Type2" text-guided eye-tracking experiment, the third and the fourth group participated in the "Type3" and "Type4" text-guided eye-tracking experiment, respectively. Therefore, each pure image or text-image pair was viewed by 15 participants without repeating. Before the experiment, each subject first read an instruction explaining the experimental process and experimental requirements, and then experienced a brief training session to be familiar with the experimental procedure. All subjects had normal or corrected normal vision.

C. Data Processing and Analysis

1) Image Attribute Analysis: We analyze four image attributes, including contrast, colorfulness, spatial information, and brightness as follows to characterize the content diversity of the images in the SJTU-TIS database.

a) Contrast: The contrast metric is measured simply by the standard deviation of gray-scale pixel intensities [45].

b) Colorfulness: Letting R, G, B indicates the RGB channels of an image, the colorfulness metric first computes two matrices $rg = R - G$ and $yb = 0.5 \times (R + G) - B$. Then, the colorfulness is calculated as $\sqrt{(\sigma_{rg})^2 + (\sigma_{yb})^2} + 0.3 \times \sqrt{(\mu_{rg})^2 + (\mu_{yb})^2}$, where σ and μ denote the standard deviation and mean respectively.

c) Spatial information (SI): Spatial information is obtained by applying a Sobel filter to each frame to extract the gradient magnitude G_{mag} and the gradient direction G_{dir} for each pixel and then the metric is calculated as $0.5 \times (\sigma_{G_{mag}} + \sigma_{G_{dir}})$, where $\sigma_{G_{mag}}$ and $\sigma_{G_{dir}}$ denote the standard deviation of G_{mag} and G_{dir}.

d) Brightness: The brightness level of an image is calculated as $Y = 0.299 \times R + 0.587 \times G + 0.114 \times B$, where R, G, B indicate the RGB channels.

The image attributes of the SJTU-TIS and SALICON databases [15] are shown in Fig. 6. It can be observed that our SJTU-TIS database covers a wide range of content diversity.

2) Eye-tracking Data Processing and Analysis: If the overall sampling rate of the eye movement is less than 90%, the data and the subject will be regarded as outlier. None of the 60 subjects is identified as outlier and removed from the experiment. We first overlay all fixation points of one image fixated by all viewers into one map to generate the fixation map of this image. Then the fixation map is smoothed with a 1° Gaussian kernel to obtain a continuous fixation density map (visual attention map).

Fig. 7 shows some schematic diagrams of the fixation maps. We use green points to represent the fixations obtained under

先看文字再看图像

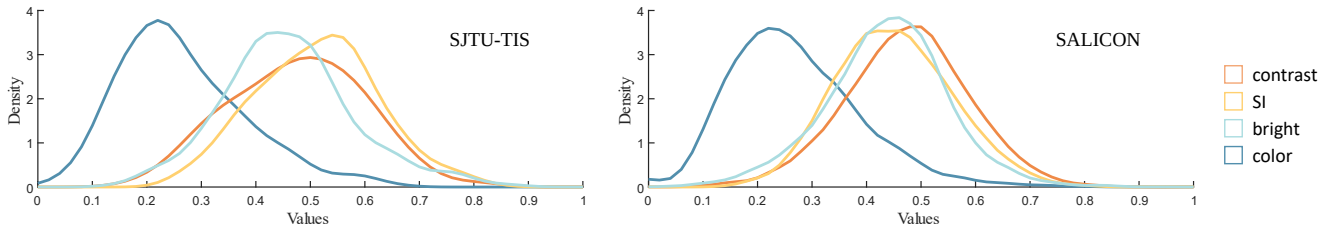


Fig. 6. Distribution comparisons of the image attributes between two databases: SJTU-TIS, SALICON [15].

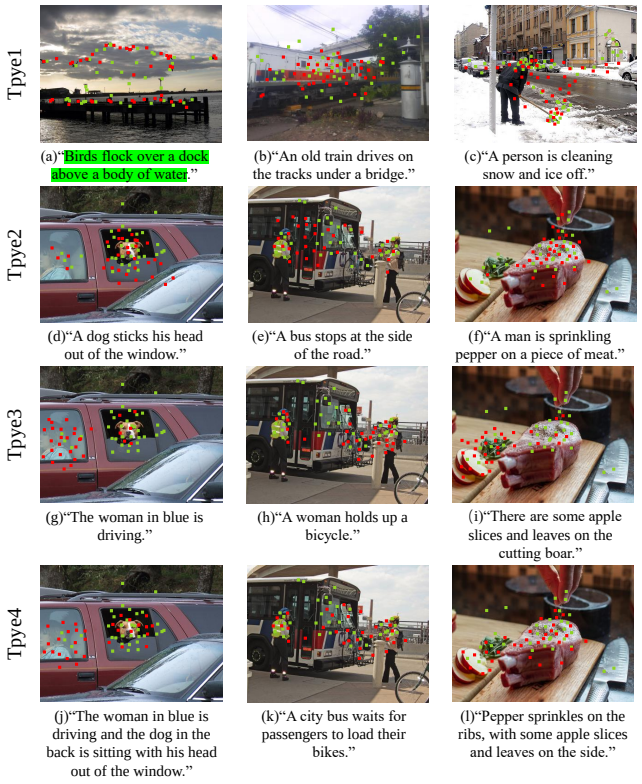


Fig. 7. Schematic diagrams of the stimuli and fixations in text-guided and pure image conditions for all 4 types. Red points: text-guided condition. Green points: pure image condition.

the pure image condition and red points to represent the fixations obtained under the text-guided condition. The first line represents the general scenario descriptions (Type1), while the second to fourth rows represent three different descriptions of the same image (Type2-Type4). It can be observed that for Type1, there is not much difference between the distributions of red and green points, which are both relatively uniform. For Type2, under the specified descriptions of salient objects, red points are more concentrated on the described object compared to green points, such as the meat in the third image. For Type3, when describing non-salient objects, there is a significant difference between the red points and green points. Green points are rarely distributed on the described non-salient object, while red points are concentrated on the described object, such as the apple slices in the third image. For Type4, under the condition of common descriptions containing both salient and non-salient objects, the red points will partially shift to non-salient objects, while most of them concentrated on salient objects.

TABLE I
DIFFERENT METRICS USE DIFFERENT FORMATS OF GROUND TRUTH FOR EVALUATING SALIENCY MODELS.

Metrics	Location-based	Distribution-based
Similarity	AUC-J, sAUC, NSS, IG	SIM, CC
Dissimilarity	-	KL

IV. Experiments and Results

In this section, we first introduce our experimental settings, including the test databases, evaluation metrics, and implementation details. Secondly, we quantitatively and qualitatively compare the proposed method with the benchmark saliency prediction models on a common saliency prediction dataset, *i.e.,* SALICON, and the constructed SJTU-TIS database. Then, we introduce our ablation studies to validate the effectiveness of each module of our proposed model.

A. Experimental Setup

1) Datasets: In order to understand and predict visual attention behavior, many eye-tracking databases have been constructed in recent years. In this paper, two databases are used to validate the effectiveness of the proposed TGSal model, including the largest publicly available saliency database SALICON [15] which is constructed for the general saliency prediction purpose, and the proposed SJTU-TIS database which is established for the text-guided saliency prediction purpose.

2) Evaluation Metrics: In the field of visual attention and saliency prediction, many consistency metrics are generally adopted to evaluate the performance of saliency algorithms. We select seven commonly used metrics including AUC-J, sAUC, CC, IG, KL, NSS, and SIM [46]. These saliency evaluation metrics can be categorized into two types including the location-based metrics and the distribution-based metrics [46]–[48], as summarized in Table I. Location-based metrics consider saliency values at discrete fixation locations, while distribution-based metrics consider both ground truth fixation density maps and predicted saliency maps as continuous distributions.

B. Comparison with State-of-the-art on SALICON

We first compare the performance of the proposed TGSal with seven state-of-the-art saliency models including SALICON [23], ML-Net [25], SalGAN [41], SAM-VGG [40], SAM-ResNet [40], GazeGAN [42] on the SALICON database.

TABLE II
QUANTITATIVE COMPARISON RESULTS OF DIFFERENT MODELS ON THE SALICON DATABASE [15].

Model\Metric	AUC-J↑	sAUC↑	CC↑	IG↑	KL↓	NSS↑	SIM↑
SALICON [23]	0.8304	0.6569	0.7486	34.8575	5.7630	1.5003	0.6664
ML-Net [25]	0.8094	0.5857	0.6746	34.7837	5.7142	1.5021	0.5858
SalGAN [41]	0.8601	0.6569	0.8601	35.2235	5.4493	1.7989	0.7520
SAM-VGG [40]	0.8524	0.6092	0.8247	34.8683	5.6555	1.7637	0.7300
SAM-ResNet [40]	0.8543	0.6019	0.8398	34.9310	5.6120	1.8121	0.7376
GazeGAN [42]	0.8522	0.6175	0.8056	35.1258	5.9386	1.6241	0.7014

For fair comparison, all these comparative models are re-trained on SALICON using their officially released code. Table II shows the performance of the baseline models and our proposed model on the SALICON database. It can be observed that our TGSal model achieves the best performance compared to other state-of-the-art models under the pure image condition, which manifests the superiority of the proposed model.

C. Comparison with State-of-the-art on SJTU-TIS

We further conduct experiments on our SJTU-TIS database to validate the effectiveness and superiority of our proposed model on the text-guided image saliency prediction tasks. We compare the proposed TGSal model with 10 classical saliency models including IT [32], AIM [33], GBVS [19], SMVJ [34], SUN [49], Hou [50], SeR [36], CA [20], HFT [37], CovSal [35], and seven DNN saliency models including SALICON [23], ML-Net [25], SalGAN [41], SAM-VGG [40], SAM-ResNet [40], GazeGAN [42]. It should be noted that all of the DNN models are first pretrained on SALICON [15] and then fine-tuned on the five groups of our SJTU-TIS database. Table III shows the quantitative comparisons of different models under the pure image condition and the general scenario description condition. Table **??** presents the quantitative comparisons of different models under the condition of specified descriptions of salient objects and the condition of specified descriptions of non-salient objects. Table **??** demonstrates the quantitative comparisons of different models under the condition of common descriptions containing both salient and non-salient objects.

V. CONCLUSION

Visual attention analysis and prediction are important tasks in multimedia systems. In this work, we conduct an in-depth exploration of text-induced visual attention and saliency prediction. Specifically, we construct the first text-guided image saliency database termed SJTU-TIS, where an image has multiple different text descriptions. Our constructed SJTU-TIS database contains 1200 text-image pairs and the correspondingly collected eye movement data. Through qualitative and quantitative analysis, we conclude that text descriptions do have influence on the visual attention, and different types of text descriptions of the same image may have different influences on the corresponding visual attention, which mainly depends on the objects being described. A novel text-guided saliency prediction model, termed TGSal, is then proposed to better predict the text-guided image saliency, which extracts both text and image features and hierarchically fuses them during the decoding process. Experimental results on the SALICON database and the SJTU-TIS database validate that our proposed method outperforms the benchmark saliency prediction models under both pure image and text-guided conditions, demonstrating the superiority and generality of the model. Moreover, under the conditions of the text descriptions of objects, the performance of our proposed TGSal is significantly improved with introducing the text features into the backbone compared to without using them, therefore manifests the importance of the text-image feature fusion for the text-guided saliency prediction task.

REFERENCES

[1] H. Duan, W. Shen, X. Min, D. Tu, J. Li, and G. Zhai, "Saliency in augmented reality," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 6549–6558.

[2] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 35, no. 1, pp. 185–207, 2012.

[3] F. Katsuki and C. Constantinidis, "Bottom-up and top-down attention: different processes and overlapping neural systems," *The Neuroscientist*, vol. 20, no. 5, pp. 509–521, 2014.

[4] X. Ren, H. Duan, X. Min, Y. Zhu, W. Shen, L. Wang, F. Shi, L. Fan, X. Yang, and G. Zhai, "Where are the children with autism looking in reality?" in *Proceedings of the CAAI International Conference on Artificial Intelligence (CICAI)*. Springer, 2022, pp. 588–600.

[5] H. Duan, X. Min, Y. Fang, L. Fan, X. Yang, and G. Zhai, "Visual attention analysis and prediction on human faces for children with autism spectrum disorder," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 15, no. 3s, pp. 1–23, 2019.

[6] H. Duan, G. Zhai, X. Min, Y. Fang, Z. Che, X. Yang, C. Zhi, H. Yang, and N. Liu, "Learning to predict where the children with asd look," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 704–708.

[7] H. Duan, G. Zhai, X. Min, Z. Che, Y. Fang, X. Yang, J. Gutiérrez, and P. L. Callet, "A dataset of eye movements for the children with autism spectrum disorder," in *Proceedings of the ACM Multimedia Systems Conference (ACM MMSys)*, 2019, pp. 255–260.

[8] Y. Zhu, G. Zhai, Y. Yang, H. Duan, X. Min, and X. Yang, "Viewing behavior supported visual saliency predictor for 360 degree videos," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 32, no. 7, pp. 4188–4201, 2021.

[9] Y. Fang, H. Duan, F. Shi, X. Min, and G. Zhai, "Identifying children with autism spectrum disorder based on gaze-following," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 423–427.

[10] Y. Cao, X. Min, W. Sun, and G. Zhai, "Subjective and objective audio-visual quality assessment for user generated content," *IEEE Transactions on Image Processing (TIP)*, 2023.

[11] ——, "Attention-guided neural networks for full-reference and no-reference audio-visual quality assessment," *IEEE Transactions on Image Processing (TIP)*, vol. 32, pp. 1882–1896, 2023.

[12] Y. Gao, X. Min, Y. Zhu, J. Li, X.-P. Zhang, and G. Zhai, "Image quality assessment: From mean opinion score to opinion score distribution," in *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, 2022, pp. 997–1005.

[13] Y. Gao, X. Min, Y. Zhu, X.-P. Zhang, and G. Zhai, "Blind image quality assessment: A fuzzy neural network for opinion score distribution prediction," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2023.

[14] Y. Gao, X. Min, W. Zhu, X.-P. Zhang, and G. Zhai, "Image quality score distribution prediction via alpha stable model," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2022.

[15] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "Salicon: Saliency in context," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1072–1080.

[16] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2009, pp. 2106–2113.

TABLE III

QUANTITATIVE COMPARISONS BETWEEN OUR PROPOSED TGSAL MODEL AND BENCHMARK METHODS UNDER PURE IMAGE CONDITION AND THE GENERAL SCENARIO DESCRIPTION CONDITION. WE **BOLD** THE BEST RESULT AND UNDERLINE THE SECOND-BEST RESULT FOR EACH METRIC IN EACH TYPE (TRADITIONAL SALIENCY MODEL OR DNN-BASED SALIENCY MODEL), RESPECTIVELY.

Type	Pure images							General scenario descriptions						
Model\Metric	AUC-J↑	sAUC↑	CC↑	IG↑	KL↓	NSS↑	SIM↑	AUC-J↑	sAUC↑	CC↑	IG↑	KL↓	NSS↑	SIM↑
IT [32]	0.7713	0.5913	0.4931	34.6916	6.8440	1.0248	0.5329	0.7900	0.6052	0.4950	34.7876	6.7735	1.1235	0.5045
AIM [33]	0.7119	0.6145	0.3494	34.3621	7.0724	0.7551	0.4635	0.7318	**0.6285**	0.3594	34.4300	7.0213	0.8376	0.4317
GBVS [19]	0.8188	**0.6169**	0.6255	**34.9341**	**6.6759**	1.2715	0.5942	0.8327	0.6249	0.6140	35.0264	6.6080	1.3578	0.5621
SMVJ [34]	0.6497	0.5434	0.2412	34.1852	7.1880	0.5610	0.4384	0.6522	0.5440	0.2277	34.2292	7.1605	0.5806	0.4033
SUN [49]	0.6453	0.5465	0.2632	34.2230	7.0723	0.5728	0.4582	0.6628	0.5527	0.2923	34.2836	7.0273	0.5927	0.4263
Hou [50]	0.6085	0.5017	0.1230	20.5622	16.6378	0.3253	0.2556	0.6110	0.5016	0.1196	20.9027	16.3977	0.3417	0.2474
SeR [36]	0.6246	0.5323	0.1905	33.8000	7.4620	0.4642	0.4093	0.6264	0.5314	0.1809	33.8824	7.4009	0.4816	0.3807
CA [20]	0.7462	0.5974	0.4290	34.5718	6.9270	0.9349	0.5085	0.7602	0.6045	0.4290	34.6780	6.8498	1.0194	0.4812
HFT [37]	0.8058	0.5834	0.5997	34.9208	6.6851	1.2851	0.5858	0.8248	0.5945	0.6026	**35.0600**	**6.5847**	1.4177	0.5630
CovSal [35]	**0.8288**	0.5629	**0.6557**	34.3128	7.1066	**1.4102**	**0.6112**	**0.8448**	0.5672	**0.6500**	34.7237	6.8178	**1.5139**	**0.6046**
SALICON [23]	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ML-Net [25]	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SalGAN [41]	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SAM_VGG [40]	0.8644	0.6290	0.8635	35.2599	6.5456	1.8835	0.7403	0.8892	0.6706	0.8321	35.5403	6.2497	2.0446	0.6954
SAM_ResNet [40]	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GazeGAN [42]	-	-	-	-	-	-	-	-	-	-	-	-	-	-

[17] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations," 2012.

[18] A. Borji and L. Itti, "Cat2000: A large scale fixation dataset for boosting saliency research," *arXiv preprint arXiv:1505.03581*, 2015.

[19] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, vol. 19, 2006.

[20] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 34, no. 10, pp. 1915–1926, 2011.

[21] J. Zhang and S. Sclaroff, "Saliency detection: A boolean map approach," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 153–160.

[22] M. Kümmerer, L. Theis, and M. Bethge, "Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet," *Computer Science*, 2014.

[23] X. Huang, C. Shen, X. Boix, and Q. Zhao, "Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 262–270.

[24] S. Jetley, N. Murray, and E. Vig, "End-to-end saliency mapping via probability distribution prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5753–5761.

[25] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "A deep multi-level network for saliency prediction," in *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 3488–3493.

[26] Z. Che, A. Borji, G. Zhai, S. Ling, J. Li, Y. Tian, G. Guo, and P. Le Callet, "Adversarial attack against deep saliency models powered by non-redundant priors," *IEEE Transactions on Image Processing (TIP)*, vol. 30, pp. 1973–1988, 2021.

[27] Q. Zhang, X. Wang, S. Wang, Z. Sun, S. Kwong, and J. Jiang, "Learning to explore saliency for stereoscopic videos via component-based interaction," *IEEE Transactions on Image Processing (TIP)*, vol. 29, pp. 5722–5736, 2020.

[28] S. Yang, G. Lin, Q. Jiang, and W. Lin, "A dilated inception network for visual saliency prediction," *IEEE Transactions on Multimedia (TMM)*, vol. 22, no. 8, pp. 2163–2176, 2019.

[29] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 740–755.

[30] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2641–2649.

[31] Y. Sun, X. Min, H. Duan, and G. Zhai, "The influence of text-guidance on visual attention," in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2023, pp. 1–5.

[32] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 20, no. 11, pp. 1254–1259, 1998.

[33] N. Bruce and J. Tsotsos, "Saliency based on information maximization," *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, vol. 18, 2005.

[34] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, vol. 20, 2007.

[35] E. Erdem and A. Erdem, "Visual saliency estimation by nonlinearly integrating features using region covariances," *Journal of Vision*, vol. 13, no. 4, pp. 11–11, 2013.

[36] H. J. Seo and P. Milanfar, "Nonparametric bottom-up saliency detection by self-resemblance," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR)*. IEEE, 2009, pp. 45–52.

[37] J. Li, M. D. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 35, no. 4, pp. 996–1010, 2012.

[38] S. S. Kruthiventi, K. Ayush, and R. V. Babu, "Deepfix: A fully convolutional neural network for predicting human eye fixations," *IEEE Transactions on Image Processing (TIP)*, vol. 26, no. 9, pp. 4446–4456, 2017.

[39] D. Tu, X. Min, H. Duan, G. Guo, G. Zhai, and W. Shen, "End-to-end human-gaze-target detection with transformers," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022, pp. 2192–2200.

[40] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting human eye fixations via an lstm-based saliency attentive model," *IEEE Transactions on Image Processing (TIP)*, vol. 27, no. 10, pp. 5142–5154, 2018.

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

[42] 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 Transactions on Image Processing (TIP)*, vol. 29, pp. 2287–2300, 2019.

[43] A. Borji, "Saliency prediction in the deep learning era: Successes and limitations," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 43, no. 2, pp. 679–700, 2019.
[44] Tobii pro. [Online]. Available: https://www.medicalexpo.com.cn/prod/tobii/product-125319-909357.html
[45] D. Hasler and S. E. Suesstrunk, "Measuring colorfulness in natural images," in *Human Vision and Electronic Imaging VIII*, vol. 5007. SPIE, 2003, pp. 87–95.
[46] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 41, no. 3, pp. 740–757, 2018.
[47] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit, "Saliency and human fixations: State-of-the-art and study of comparison metrics," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 1153–1160.
[48] M. Kümmerer, T. S. Wallis, and M. Bethge, "Information-theoretic model comparison unifies saliency metrics," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 112, no. 52, pp. 16 054–16 059, 2015.
[49] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "Sun: A bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, no. 7, pp. 32–32, 2008.
[50] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, vol. 21, 2008.