Does text attract attention on e-commerce images: A novel saliency prediction dataset and method

Lai Jiang[1]*, Yifei Li[1]*, Shengxi Li[1]†, Mai Xu[1]†, Se Lei[1], Yichen Guo[1], Bo Huang
[1] School of Electronic and Information Engineering, Beihang University, Beijing, China

Abstract

E-commerce images are playing a central role in attracting people's attention when retailing and shopping online, and an accurate attention prediction is of significant importance for both customers and retailers, where its research is yet to start. In this paper, we establish the first dataset of saliency e-commerce images (SalECI), which allows for learning to predict saliency on the e-commerce images. We then provide specialized and thorough analysis by highlighting the distinct features of e-commerce images, e.g., non-locality and correlation to text regions. Correspondingly, taking advantages of the non-local and self-attention mechanisms, we propose a salient SWin-Transformer backbone, followed by a multi-task learning with saliency and text detection heads, where an information flow mechanism is proposed to further benefit both tasks. Experimental results have verified the state-of-the-art performances of our work in the e-commerce scenario.

1. Introduction

Nowadays, online retailing has revolutionized the shopping habits in daily life, which provides significantly improved efficiency and hands-on experiences for both costumers and retailers. The sudden break of Coronavirus-19 further emphasized the importance and popularity of online shopping. Since "*a picture is worth a thousand words*", the e-commerce image, exhibiting rich and heuristic content, has been a workhorse in promoting products on online shopping and it therefore plays a vital role in the shopping activities, including introducing products, aiding visual search, attracting costumers, and affecting their final decisions.

Due to the intrinsic nature of retailing, the main goal of e-commerce images is to *attract costumer attentions at a glimpse*; this is mainly two-fold: 1) attracting costumers to focus on the product when they are wandering on shopping, and 2) attracting to focus on specific parts in images that highlight distinguishable and "have-to-buy" features of the product. Consequently, the e-commerce image is typically a combination of pictures and texts, to achieve the goal of both effectively attracting and introducing to the costumers. Thus, saliency prediction on e-commerce images is of significant importance to provide both enhanced guiding information and shopping experience for costumers.

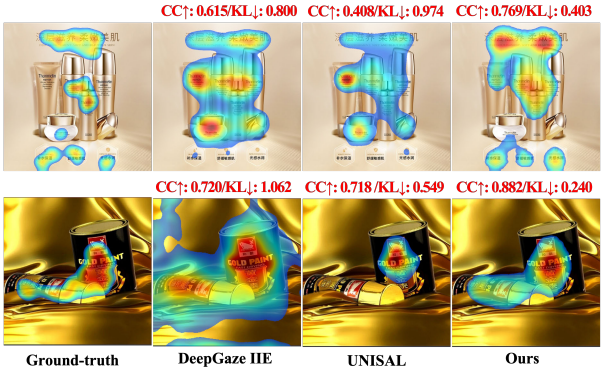


Figure 1. Saliency maps of ground-truth, DeepGace IIE [26], UNISAL [7], and the proposed method. As shown, the existing methods on natural images trend to over- or under-predict the saliency values of text regions in e-commerce images.

The existing works of saliency prediction almost focus on the natural images, from the perspective of either low-level handcrafted cues [14, 44] or data-driven deep neural networks (DNNs) [7, 18, 27, 29]. However, with the fundamental goal of retailing, the e-commerce images are specially designed, especially with short but precise texts. Thus, the e-commerce images are basically different from the widely analysed natural images, which for example, are created by the goals of beauty, recording, etc. Consequently, the existing methods are inadequate in predicting saliency of the e-commerce images. For example, the object region in natural images, one of the most important high-level cues when predicting saliency in existing methods [13, 17], may be equally and even less salient than the text regions in e-commerce images that highlight key features and brands of products. Fig. 1 illustrates the limitation of the existing saliency prediction methods on e-commerce images, where the most recent methods on natural images, i.e., DeepGace IIE [26] and UNISAL [7], trend to over- or

*Both authors contributed equally to this research.
†Corresponding authors: Shengxi Li (LiShengxi@buaa.edu.cn), Mai Xu (MaiXu@buaa.edu.cn)

under-predict the saliency value of the text regions. Thus, it is necessary and important to develop a new DNN for e-commerce images, by addressing the text priors. Meanwhile, the lack of e-commerce image dataset also impedes applying DNN-based models for saliency prediction.

To this end, we propose a novel work on saliency prediction for e-commerce images. More specifically, we establish the first eye-tracking dataset of e-commerce images, called SalECI, with recorded fixations from the eye-tracking experiments. We further analyse the proposed dataset, obtaining 4 observations. Upon the observations, we validate the non-locality nature of saliency maps. Due to the fact that saliency explicitly points out human attentions when looking at e-commerce images, we take the advantages of non-local and attention mechanisms of Transformers [40], and propose a salient Swin-Transformer (SSwin-Transformer) backbone that incorporates the saliency information to improve the learnt self-attention maps. More importantly, our observations also point out a consistent and strong relationship between the salient and text regions. Therefore, we propose to simultaneously predict saliency and detect text regions by two learnable heads, together with an information flow to let the two heads interact between each other. The experimental results have verified the state-of-the-art performance of our work. A further application to e-commerce image compression further achieves significant bit-rate saving. Our main contributions are as follows,

- We establish the first SalECI dataset, enabling advanced data-driven architectures to predict saliency on e-commerce images;
- We provide thoroughgoing and comprehensive analysis on the SalECI dataset, paving the way of specialized and insightful methods on e-commerce images;
- We propose a novel multi-task learning framework including the SSWin-Transformer, multiple heads and information flow mechanism, achieving the state-of-the-art performances on e-commerce scenarios.

2. Related Work

Saliency Prediction. Traditional saliency prediction methods aim to predict pixel-wise human attention, mainly relying on low-level hand-designed features, including contrast [4], color [14], luminance [38], and texture [44]. As one of the most pioneer works, Itti *et al.* [14] proposed a bottom-up method for image saliency prediction, by constructing multi-scale feature maps of color, intensity and orientation. Different from Itti's method [14], Guo *et al.* [8] transformed the images/videos into a quaternion Fourier domain, to extract spatio-temporal features for saliency prediction. Most recently, with the rapid development of deep learning, large-scale eye-tracking datasets [13, 17, 42] and advanced DNN structures [5,7,16,26] were proposed to significantly improve the performance on saliency prediction. Specifically, Huang *et al.* [13] conducted mouse-contingent experiments to collect clicks over the images as the representations of fixations, and established a saliency dataset with $20k$ images. On the other hand, DNN architectures were developed and verified to be effective on saliency prediction, e.g., fully convolutional network [19], generative adversarial network (GAN) [37], convolutional long short-term memory network [17], dilated convolution [18], complex-valued network [16], transformer [33], etc. However, none of the above datasets and methods deal with the saliency prediction for the e-commerce images, which play an important role in digital shopping. To this end, the brand-new dataset and method are proposed in this paper for saliency prediction on the e-commerce images.

Text Detection. Traditional text detection methods are mainly based on connected components analysis (CCA) [12,36,45] and sliding windows [20,41]. For example, Neumann *et al.* [36] proposed to first extract candidate components, and then a support vector machine (SVM) was developed to filter out the non-text candidates. In [20], windows with different sizes were used to slide over the image, and each window was classified by morphological operations. Most recently, DNN based text detectors have been developed by relating text detection to object detection [23, 43, 46, 48] and instance segmentation [6, 9, 24, 32]. Specifically, Liao *et al.* [23] adapted the object detection framework called the SSD [28], to capture the texts with various orientations and shapes. Similarly, Zhang *et al.* [46] used FPN [25] to detect the text candidates, and then conducted a localization branch to progressively refine the bounding boxes. Moreover, EAST [48] directly detected the quadrangles of words in an end-to-end manner without proposals and anchors. In addition to the detection-based methods, the segmentation-based methods aim to detect text regions in the pixel level. For instance, PixelLink [6] straightforwardly extracted the text bounding areas from the segmentation maps. Besides, Long *et al.* [32] proposed to detect text instances by predicting the geometry attributes and the center line of the text, based on the segmentation network of FCN [31]. Liao *et al.* [24] sped up the traditional segmentation-based pipeline by developing threshold maps. Different from the above methods, Baek *et al.* [1] proposed a character level text detector, in which both components and links between characters were predicted.

3. SalECI Dataset

To study human perceptual behaviours on the e-commerce images, we establish the new SalECI dataset, including 972 e-commerce images with collected fixations and annotated text boundaries. All images collected from mainstream platforms, including Taobao, Amazon and

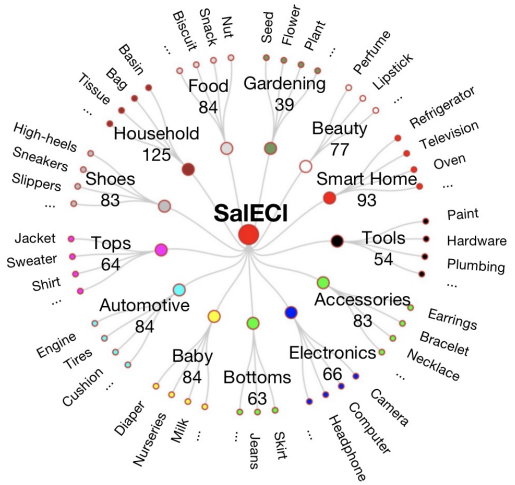


Figure 2. Main categories and sub-categories in our SalECI dataset. The number of images in each main category is also listed.

Wish. Fig. 2 shows 13 categories and their corresponding image numbers in SalECI. In summary, our SalECI dataset includes $257,302$ fixations from the eye-tracking experiment of 25 subjects, and $10,833$ text bounding boxes annotated by 3 volunteers. The details about dataset establishment is introduced in the supplementary. SalECI is released in https://github.com/leafy-lee/E-commercial-dataset.

3.1. Data Analysis

Observation 1: *In e-commerce images, the visual attention is easy to be attracted by the texts, in comparison with semantic objects.*

Analysis: According to the previous works [17, 42], the visual attention is more likely to be attracted by semantic objects. Here, given the ground-truth saliency maps and text bounding boxes, we further evaluate the correlation between visual attention and texts in SalECI. As for the comparison, we also apply YOLOv5 [39] to detect the bounding boxes of the objects in SalECI. Some examples in our SalECI are shown in Fig. 3-(a), with text and object regions in blue and green, respectively. As shown in this figure, the text regions are more consistent with the visual attention. To verify this, we calculate the fixation densities (per $1,000$ piexels) of the text and object regions, by counting the number of fixations falling into the text and object regions, respectively. Fig. 3-(b) shows the above fixation densities of each category and all images in SalECI. Note that the fixation density of randomly extracted regions is also illustrated in the figure as a baseline. As shown in Fig. 3-(b), the fixation density of the text regions is considerably larger than that of object regions. For some categories, such as gardening, tops, bottoms, and tools, the text regions averagely draw around 5 times as much attention as the object regions, implying that these categories are more sensitive to the text. The above results indicate that, for e-commerce images, the visual attention is highly likely to be attracted by the texts.

Observation 2: *Although visual attention can be significantly attracted by the texts in e-commerce images, there are still remarkable fixations out of the text regions.*

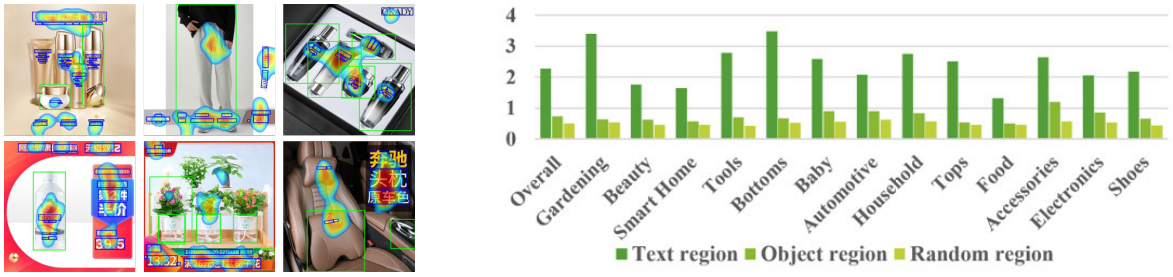
Analysis: As introduced in **Observation 1**, the visual attention can be greatly attracted by the texts. However, we might ask, *does it mean we can directly use text detection method for e-commerce image saliency prediction?* To this end, we further count the fixation numbers outside the text regions, against all fixations in the image. Fig. 4-(a) presents the proportion of fixations out of text regions. In this figure, each point represents an e-commerce image in SalECI, and the horizontal axis indicates the text area of each image. As shown in Fig. 4-(a), for most of the images, around 40% to 70% fixations are out of the text regions. This implies that, beyond the texts, the visual attention is also attracted by multiple regions with either bottom-up or top-down saliency. Similarly, in Fig. 4-(b), we calculate the proportion of text regions without any fixations inside. It is not surprising to find that a large number of text regions do not attract any visual attention. Besides, as shown Fig. 4, similar trends occur in different categories of SalECI. The above results indicate that saliency prediction for e-commerce images is complicated, and cannot be simply solved by applying text detection method.

Observation 3: *In e-commerce images, the visual attention is consistent among subjects, especially for the attention inside the text regions.*

Analysis: Regarding natural images [13] and videos [17], there exists high consistency of visual attention among subjects. Here, we measure the visual consistency in SalECI, by calculating linear correlation coefficient (CC) between the fixation maps of single subject and the rest of subjects, which is also called one-vs-rest CC. As listed in Table 1, the CCs of the whole image, text regions, and object regions are calculated, respectively. Meanwhile, in order to evaluate the location bias in the SalECI dataset, we also measure the CC between the fixation map of 2 randomly selected e-commerce images. Besides, as reported in [17], the one-vs-rest CC results of 2 other eye-tracking datasets, i.e., LEDOV [17] and Hollywood [34], are also listed as the baselines. As shown, we can conclude that the visual consistency in the SalECI dataset is similar to the other eye-tracking datasets. Moreover, the visual consistency improves when only considering the fixations inside the text regions. This again implies that the subjects tend to focus on the texts when viewing e-commerce images.

Observation 4: *The fixation transition in e-commerce images is typically much larger than the fovea region, indicating that the visual attention tends to be attracted by non-local content in e-commerce images.*

Analysis: As shown in Fig. 5-(a), the salient regions of e-commerce images tend to be separated from each other. This implies that the human visual attention may be eas-

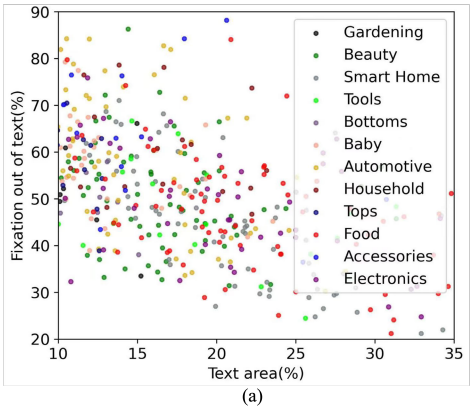


(a) (b)

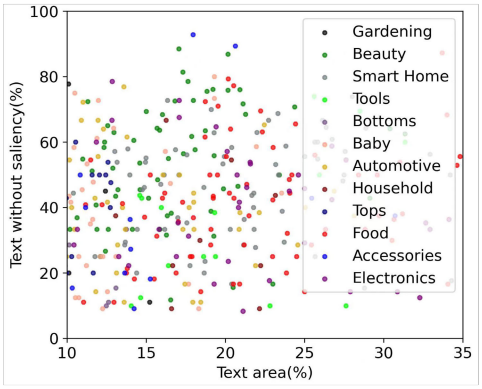
Figure 3. The analysis of ***observation 1***. (a) Images in SalECI with corresponding ground-truth saliency maps, text regions (in blue rectangulars), and object regions (in green rectangulars). (b) The fixation densities in text, object, and random regions.

Table 1. The visual consistency across subjects in our SalECI, LEDOV [17] and Hollywood [34], in terms of one-vs-rest CC.

	SalECI	SalECI Text	SalECI Object	Location Bias	LEDOV [17]	Hollywood [34]
CC	0.356	0.482	0.410	0.152	0.403	0.349



(a)



(b)

Figure 4. The analysis of ***observation 2***. (a) The proportion of fixations that are out of text regions. (b) The proportion of text regions that are without any fixations inside.

ily attracted by the non-local content. Thus, we evaluate the fixation transition in SalECI, by calculating the visual angle between two consecutive fixations of same subject. Since our eye-tracking experiment fixed the size of screen and the distance between the subject and screen, we can calculate the visual angle via the trigonometric function, as illustrated in Fig. 5-(b). As a result, the proportion of the visual angle of fixation transition are listed in Fig. 5-(c). According to [35], the human visual attention only focuses on the fovea region with visual angles less than 2 degrees. However, as shown in the figure, 26.2%, 14.9%, and 10.2% fixation transition are two, three, and four times larger than the fovea region, respectively. The distant fixation transition indicates that, in e-commerce images, the human attention is easier to be attracted by the non-local content. That is probably because the e-commerce images are specially designed to contain the semantic objects and texts all over the image, rather than in a part of the image.

4. Proposed Method

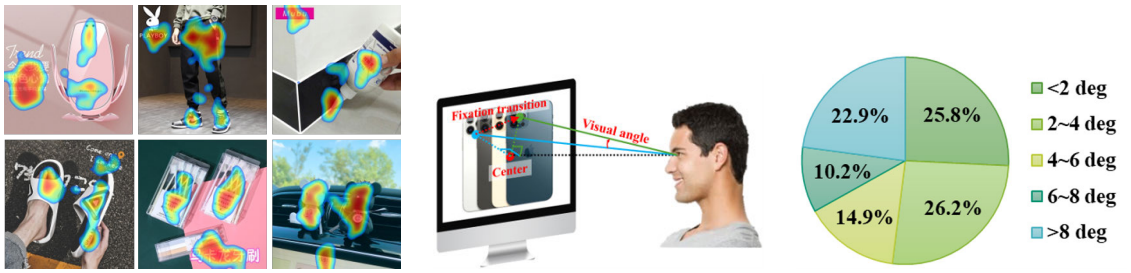
4.1. SSwin-Transformer Backbone

As analysed in **Observation 4**, the saliency information, indicating human non-local attention, is of indispensable relation with the attention mechanism of the Transformer [40] and therefore is highly potential to improve the attention backbone learnt from the Transformer. In our work, the SSwin-Transformer builds upon the Swin-Transformer, which achieves the state-of-the-art performance across a wide range of tasks [30]. More importantly, we incorporate the saliency information in each Swin Transformer block to aid shaping the attention maps, as illustrated in Fig. 6.

More specifically, towards the last basic layer of each stage, we propose an attention loss $\mathcal{L}_a$ with the usage of saliency maps, to supervise the learnt attentions in the backbone. The saliency map $\mathbf{S}_l$, in a size of $h_l \times w_l$ for the l-th basic layer, is reshaped correspondingly to keep the same size of the self-attention map processed by the patch merging operation. Then, for the l-th basic layer, the attention loss $\mathcal{L}_a$ is calculated in a channel-wise manner as follows,

$$\mathcal{L}_a = \frac{1}{2}\sum_{h=1}^{H_l} ||\mathrm{cor}(\mathbf{S}_l) - \mathrm{cor}(\mathbf{A}_{l,h})||_2^2 \qquad (1)$$

where $\mathbf{A}_{l,h}$ represents the self-attention map at the h-th multi-head output from the shifted Swin-Transformer block



(a) (b) (c)

Figure 5. The analysis of ***observation 4***. (a) Images and corresponding ground-truth saliency maps in SalECI. (b) The illustration of calculating visual angle on two consecutive fixations. (c) The statistics of fixation transition in terms of the visual angle.

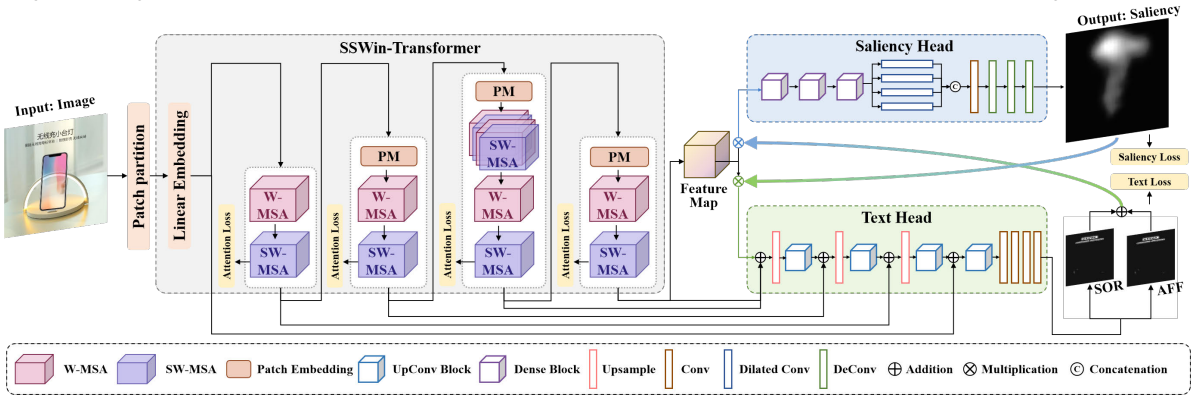


Figure 6. Illustration of our proposed framework. The framework consists of SSwin-Transformer backbone, saliency and text detection heads, together with an information flow structure that takes advantages of the fine-grained output (namely, saliency maps and text masks) as the feedback to further improve the learning of the proposed backbone and functional heads. Note that SOR represents the score of region, while AFF stands for the affinity scores.

in the l-th basic layer. H_l is the number of multi-heads for the l-th layer. Moreover, $\mathrm{cor}(\cdot)$ denotes the non-local correlation that is calculated by

$$\mathrm{cor}(\mathbf{X}) = \mathrm{softmax}\big(\mathrm{vec}(\mathbf{X}) \cdot \mathrm{vec}(\mathbf{X})^T\big), \tag{2}$$

where $\mathrm{vec}(\mathbf{X})$ vectorizes the matrix $\mathbf{X}$ of the size $h_l \times w_l$ to a vector with a size $(h_l \times w_l) \times 1$, and $\mathrm{softmax}(\cdot)$ denotes the soft-max operation.

The basic idea of the proposed attention loss is to guide our proposed SSwin-Transformer backbone so as to learn non-local correlations according to human perceptions, whilst maintaining the diversity enabled by multi-head self-attention maps. This, one one hand, encourages the proposed SSwin-Transformer to focus on non-local regions that human valued most, given the observations that people consistently focused on text regions when they are looking at the e-commercial images. On the other hand, this also imposes global and unified prior on the network that is important in multi-task learning. Consequently, the output features from the backbone are enhanced with global cues for multi-task learning and therefore benefit the following saliency and text detection heads.

4.2. Saliency and Text Detection Heads

We employ a light but effective saliency head to predict saliency maps given the feature map of our SSWin-Transformer backbone. To be more specific, the feature map is first fed into 3 dense blocks [11], and with the aim to extract multi-scale information when predicting saliency maps, we employ the atrous spatial pyramid pooling (ASPP) [3] afterwards and then use 3 deconvolution blocks to recovery the saliency map. The loss of our saliency head is calculated by the Kullback-Leibler (KL) divergence $\mathrm{KL}(\cdot||\cdot)$ between the predicted $\mathbf{S}^p$ and ground-truth $\mathbf{S}^{gt}$ saliency maps as follows,

$$\mathcal{L}_s = \mathrm{KL}(\mathbf{S}^p||\mathbf{S}^{gt}). \tag{3}$$

Additionally, the basic structure of the CRAFT [1] is developed in our work for the text detection, which is able to achieve character-level (rather than the word-level) text detections. Specifically, we first preliminarily generated the character-level annotations by precisely using the same network structure and pre-trained models in its official implementations. We further manually adjusted the annotations with an incomplete or inaccurate coverage of texts in the E-commerce images, and then obtained the ground-truth for training our text detection head. We show in Fig. 3-(a) the ground-truth text annotations by red rectangulars.

In our text detection head, since the input of the head is the feature map from our SSWin-Transformer backbone, we employ the deconvolution modules whilst discarding the

feature encoding module as in [1]. Furthermore, to enhance the information aggregation across different resolutions, 5 middle attention maps at multi-scale resolutions in our SSwin-Transformer backbone are fed into the deconvolution modules, as illustrated in Fig. 6. Therefore, in the proposed text detection head, 4 deconvolution modules are used to output the final region and affinity scores, and their difference from the ground-truth scores are evaluated by the mean squared error (MSE). However, due to the intrinsic nature of advertising, texts in e-commerce images are mainly in short and precise forms (e.g., phrases and logos), leading to rather sparse text regions and affinity score maps. Therefore, directly applying MSEs as our text head loss can cause the overwhelming on negative predictions, that is, outputting zero values almost everywhere. To overcome this issue of imbalanced samples, we develop a balanced MSE loss BMSE to relieve the ill-posed training where the network simply outputs 0. More specifically, when calculating the MSE, we randomly select N_{pos} positive samples and N_{neg} negative samples, and calculate the balanced MSE as follows,

$$\mathrm{BMSE}(\mathbf{X},\mathbf{Y}) = \frac{\sum_{(i,j)\in\mathcal{P}\cup\mathcal{N}} ||\mathbf{X}(i,j)-\mathbf{Y}(i,j)||_2^2}{N_{pos}+N_{neg}}, \quad (4)$$

where $\mathcal{P}$ and $\mathcal{N}$ denote the indices of positive and negative samples. Note that $\mathbf{X}(i,j)$ represents the position (i,j) given the matrix $\mathbf{X}$. Upon the balanced MSE, we employ the following loss $\mathcal{L}_t$ for our text detection head in training

$$\mathcal{L}_t = \mathrm{BMSE}(\mathbf{T}_r^p, \mathbf{T}_r^{gt}) + \mathrm{BMSE}(\mathbf{T}_a^p, \mathbf{T}_a^{gt}), \quad (5)$$

where $\mathbf{T}_r^p$ and $\mathbf{T}_r^{gt}$ are the predicted and ground-truth scores of region (SORs), whereas $\mathbf{T}_a^p$ and $\mathbf{T}_a^{gt}$ represent the predicted and ground-truth affinity scores (AFFs), respectively.

4.3. Information Flow

As indicated by **Observations 1** and **2**, in E-commerce images, although not fully attracting human fixation points, the text regions consistently attract visual attention. The detection of text regions, is therefore helpful in tailoring the accuracy of saliency prediction. In contrast, an accurate prediction of saliency can also be of help on precisely detecting text regions since the saliency information almost includes all text regions in the E-commerce images.

Correspondingly, we flow the information output from the saliency head back to the input of the text detection head, and also flow the output from the text detection head to the input of saliency head. Such an interactive information flow is able to improve both the learning of saliency prediction and text detection. Specifically, the initial prediction is made by the saliency head to refine the text head. The forward process of the saliency head now becomes

$$\mathbf{S}^p = \mathrm{SalHead}\big(f(\frac{\widetilde{\mathbf{T}}_r^p + \widetilde{\mathbf{T}}_a^p}{2}) \odot \mathbf{F}\big), \quad (6)$$

where $\mathrm{SalHead}(\cdot)$ represents the saliency head, $\widetilde{\mathbf{T}}_r^p$ and $\widetilde{\mathbf{T}}_a^p$ denote the resized predicted region and affinity scores so that they have the same size of the feature map $\mathbf{F}$, and $\odot$ denotes the element-wise product. More importantly, $f(\cdot)$ is an element-wise scaling function, which is set as $f(x) = \rho \cdot (x - 0.5) + 1$ in our work given $0 \le x \le 1$ and ρ being a scaling factor. In this way, compared to the background regions with zero outputs, the detected text regions that have positive values can properly increase the importance of corresponding locations so that the saliency prediction can be further improved through this extra information.

Afterwards, the prediction from the text head is employed to refine the saliency head, completing the interaction of information flow. Given the resized saliency output $\widetilde{\mathbf{S}}^p$, the text detection head process is as follows,

$$\mathbf{T}_r^p, \mathbf{T}_a^p = \mathrm{TextHead}\big(f(\widetilde{\mathbf{S}}^p) \odot \mathbf{F}\big). \quad (7)$$

Therefore, after our SSWin-Transformer backbone, we first initialize the text head output by all ones matrix when forwarding the saliency head of (6) and obtain a coarse prediction $\mathbf{S}^p$, after which is fed into (7) to obtain text detections $\mathbf{T}_r^p$ and $\mathbf{T}_a^p$. A fine-grained saliency prediction is then obtained by forwarding (6) again upon the obtained text detections $\mathbf{T}_r^p$ and $\mathbf{T}_a^p$. We may need to point out that although iterating this procedure might obtain further enhanced predictions, we empirically find that the gain is slight, whereas at the cost of computational complexity. For the ease of computation efficiency, we only flow the information once for each head. Finally, our cost is given by

$$\mathcal{L} = \lambda_a \mathcal{L}_a + \lambda_s \mathcal{L}_s + \lambda_t \mathcal{L}_t, \quad (8)$$

where λ_a, λ_s and λ_t are used to adjust the scales of losses.

5. Experiments

5.1. Implementation Details

In our experiments, SalECI are randomly divided into the training and test sets with 871 and 101 images, respectively. For stable training, Batch normalization, leaky ReLU, and GeLU [10] are used as the normalization and activation functions in SSwin-transformer. The resolution of input and output images are set as 896×896, and the channel number of the embed feature from SSwin-transformer is 96. The scaling factor ρ and the loss weights $\{\lambda_a, \lambda_s, \lambda_t\}$, are set to be 0.2 and $\{1, 1, 3\}$. During training, the proposed method is optimized based on stochastic gradient descent with a Adam optimizer. Besides, the initial learning rate is 5×10^{-7}, and a warm-up cosine learning schedule is conducted for the first 20 epochs. The whole training process takes about 1.5 hours on a RTX 3090Ti GPU for 50 epochs.

5.2. Evaluation on SalECI Dataset

In this section, we evaluate the performance of our method on e-commerce image saliecny prediction, com-

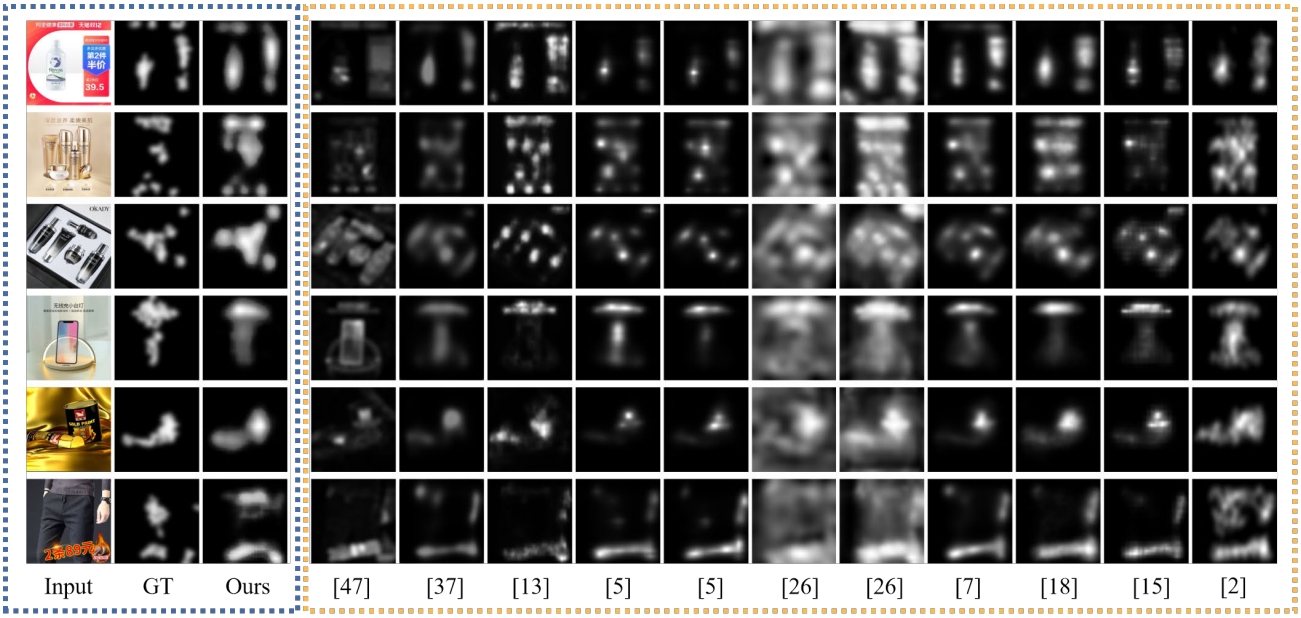


Figure 7. Qualitative results of our and 11 other compared methods over 8 randomly selected e-commerce images in SalECI. From the left to right are: input images, the saliency maps of ground-truth, ours, BMS [47], SalGAN [37], SALICON [13], SAM-ResNet [5], SAM-VGG [5], DeepGaze I [26], DeepGaze IIE [26], UNISAL [7], MSI [18], EML-Net [15], and GazeGAN [2].

Table 2. Mean value and standard deviation of saliency prediction accuracy for our and 11 other methods over SalECI.

Methods	CC	KL	AUC	NSS	SIM	sAUC
BMS [47]	0.411±0.173	1.108±0.291	0.768±0.079	1.025±0.507	0.395±0.090	0.726±0.080
SalGAN [37]	0.552±0.174	0.873±0.311	0.826±0.069	1.449±0.574	0.496±0.100	0.766±0.082
SALICON [13]	0.507±0.149	0.967±0.296	0.805±0.073	1.334±0.507	0.461±0.091	0.767±0.077
SAM-ResNet [5]	0.535±0.181	0.855±0.336	0.832±0.067	1.452±0.632	0.500±0.103	0.766±0.084
SAM-VGG [5]	0.551±0.178	0.844±0.342	0.833±0.066	1.515±0.640	0.512±0.104	0.771±0.078
DeepGaze I [26]	0.407±0.158	1.250±0.225	0.760±0.093	0.969±0.404	0.326±0.052	0.741±0.092
DeepGaze IIE [26]	0.561±0.124	0.995±0.215	0.842±0.055	1.327±0.318	0.399±0.065	**0.811± 0.058**
UNISAL [7]	0.605±0.148	0.768±0.262	0.845± 0.056	1.574±0.522	0.514±0.094	0.777±0.075
MSI [18]	0.603±0.173	0.804±0.310	0.834±0.066	1.555±0.554	0.514±0.104	0.771±0.086
EML-Net [15]	0.597±0.154	0.788±0.328	0.841±0.063	1.595±0.561	0.534±0.101	0.780±0.080
GazeGAN [2]	0.522±0.194	0.987±0.453	0.797±0.090	1.321±0.575	0.481±0.117	0.706±0.114
SSwin transformer(Ours)	**0.687±0.175**	**0.652±0.478**	**0.868±0.072**	**1.701±0.497**	**0.606±0.101**	0.783±0.064

pared with 11 other state-of-the-art saliency prediction methods, i.e., BMS [47], SalGAN [37], SALICON [13], SAM-ResNet [5], SAM-VGG [5], DeepGaze I [26], DeepGaze IIE [26], UNISAL [7], MSI [18], EML-Net [15], and GazeGAN [2]. Note that except for BMS, all compared methods are based on DNN. Specifically, except BMS (not learning based) and DeepGaze (no released training code), all compared methods are fine-tuned over SalECI with the similar experimental setting as ours. Then, 6 metrics are applied to measure the performance of saliency prediction: CC, KL divergence, the area under the receiver operating characteristic curve (AUC), normalized scanpath saliency (NSS), simliarity (SIM), shuffled AUC (sAUC). Note that the larger values of CC, AUC, NSS, SIM or sAUC, and smaller KL, indicate more accurate saliency prediction. As tabulated in Table 2, our method significantly outperforms the compared methods in all metrics. Compared with the second best method UNISAL, the proposed method can achieve 0.082, 0.116, 0.023, 0.127, and 0.072 improvements in terms of CC, KL, AUC, NSS, and SIM, respectively. In addition to the quantitative results, Figure 7 shows the qualitative results of our and 11 other methods over 8 randomly selected test images in SalECI. We can see from figure that our method is capable of well locating the salient regions, making the predicted saliency map closer to the ground-truth than the other methods. Particularly, as shown in the first two rows of this figure, our method can correctly detect the text regions that attract visual attention, compared with other methods. That verifies the effectiveness of the multi-task learning framework in our method.

5.3. Ablation Study

Here we further conduct the ablation experiments to analyse the contribution of each component proposed in our

(a) HM-16.0: 20.90dB (b) UNISAL [7]: 26.19dB (c) MSI [18]: 26.31dB (d) Ours: 28.00dB

Figure 8. Subjective results and EW-PSNRs of the original HM-16.0 and the perceptual codec [21] fed by saliency prediction from the UNISAL [7], MSI [18] and ours methods.

Table 3. Ablation study on the proposed method in terms of CC, KL, AUC and NSS.

Modules	CC	KL	AUC	NSS
A + B	0.629	0.847	0.833	1.537
A + B + C	0.670	0.759	0.848	1.662
A + B + D	0.633	0.780	0.843	1.603
A + B + C + D	0.678	0.703	0.854	1.690
A + B + D + E	0.661	0.677	0.863	**1.709**
A + B + C + D + E	**0.687**	**0.652**	**0.868**	1.701

Module notations: A: SSwin-Transformer; B: Saliency head; C: Attention loss; D: Text head; E: Information flow

method. Each ablated model is trained and tested in the same experimental setting. As a result, the CC, KL, AUC and NSS values of each ablated model are listed in Table 3. Note that the basic components of SSwin-trnasformer and saliency head can not be ablated, otherwise the saliency map is inaccessible. We can see from this table that, for different ablated models, the multi-scale attention loss can stably bring the improvements. For instance, compared with the model only with SSwin-trnasformer and saliency head, the attention loss is able to further improves the performance of saliency prediction, by 0.041 CC, 0.088 KL, 0.015 AUC, and 0.125 NSS. This indicates that the non-local attention in the Transformer can benefit from the supervision of real human attention. Besides, similar improvements are achieved by adding text head or the developed information flow. This again verifies one of the main motivations of this work, that is, there exists a strong correlation between the salient and text regions in e-commerce images.

5.4. Application in Video Compression

In this section, we further show that the improvement on the saliency prediction on e-commerce images could witness practical gains when compressing e-commerce images. More specifically, we followed the work of [21] to compress the images upon the state-of-the-art high efficiency video coding (HEVC) standard, with the aim of improving the perceptual quality of compressed images. We implemented the perceptual codec [21] on the official platform HM-16.0 and evaluated the perceptual quality in terms of the eye-tracking weighted PSNR (EW-PSNR), which has been verified with a strong relationship with the subjective quality [22]. The subjective results as well as the EW-PSNR are shown in Fig. 8, with more results shown in the supplementary material. From this figure, it is obvious that for the e-commerce images, our method, being able to more accurately prediction saliency, achieves the best subjective quality, with clear details on the product and brand.

6. Conclusion

This paper has set out a first attempt for saliency prediction on e-commerce images. The first eye-tracking e-commerce image dataset, called SalECI, has been established to enable training DNN for e-commerce image saliency prediction. Based on the newly-built dataset, we conducted thorough data analysis, leading to 4 important observations on e-commerce images. Insipred by the observations, we proposed a new multi-task learning framework for e-commerce image saliency prediction, which is composed of the developed SSWin-Transformer, saliency head, text head, and information flow mechanism. The experimental results showed that our methods significantly outperform the state-of-the-art image saliency prediction.

Acknowledgments

This work is supported by NSFC under Grants 61922009, 61876013, 62050175, Beijing Natural Science Foundation under Grant JQ20020, and Alibaba Innovative Research.

References

[1] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. Character region awareness for text detection. In *Proceedings of the IEEE/CVF Conference on Com-*

puter Vision and Pattern Recognition, pages 9365–9374, 2019. 2, 5, 6

[2] Z. Che, A. Borji, G. Zhai, X. Min, G. Guo, and P. L. Callet. How is gaze influenced by image transformations? dataset and model. 2019. 7

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 5

[4] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):569–582, 2015. 2

[5] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. SAM: Pushing the limits of saliency prediction models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1890–1892, 2018. 2, 7

[6] Dan Deng, Haifeng Liu, Xuelong Li, and Deng Cai. Pixellink: Detecting scene text via instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 2

[7] Richard Droste, Jianbo Jiao, and J Alison Noble. Unified image and video saliency modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 419–435. Springer, 2020. 1, 2, 7, 8

[8] Chenlei Guo and Liming Zhang. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE Transactions on Image Processing*, 19(1):185–198, Jan. 2010. 2

[9] Pan He, Weilin Huang, Tong He, Qile Zhu, Yu Qiao, and Xiaolin Li. Single shot text detector with regional attention. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3047–3055, 2017. 2

[10] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 6

[11] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 5

[12] Weilin Huang, Zhe Lin, Jianchao Yang, and Jue Wang. Text localization in natural images using stroke feature transform and text covariance descriptors. In *Proceedings of the IEEE international Conference on Computer Vision*, pages 1241–1248, 2013. 2

[13] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *ICCV*, pages 262–270, 2015. 1, 2, 3, 7

[14] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, Nov. 1998. 1, 2

[15] Sen Jia and Neil DB Bruce. Eml-net: An expandable multilayer network for saliency prediction. *Image and Vision Computing*, 95:103887, 2020. 7

[16] Lai Jiang, Zhe Wang, Mai Xu, and Zulin Wang. Image saliency prediction in transformed domain: A deep complex neural network method. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8521–8528, 2019. 2

[17] Lai Jiang, Mai Xu, Zulin Wang, and Leonid Sigal. DeepVS2.0: A saliency-structured deep learning method for predicting dynamic visual attention. *International Journal of Computer Vision*, 129(1):203–224, 2021. 1, 2, 3, 4

[18] Alexander Kroner, Mario Senden, Kurt Driessens, and Rainer Goebel. Contextual encoder–decoder network for visual saliency prediction. *Neural Networks*, 129:261–270, 2020. 1, 2, 7, 8

[19] Srinivas SS Kruthiventi, Kumar Ayush, and Radhakrishnan Venkatesh Babu. Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing*, 2017. 2

[20] Jung-Jin Lee, Pyoung-Hean Lee, Seong-Whan Lee, Alan Yuille, and Christof Koch. Adaboost for text detection in natural scene. In *2011 International Conference on Document Analysis and Recognition*, pages 429–434. IEEE, 2011. 2

[21] Shengxi Li, Mai Xu, Yun Ren, and Zulin Wang. Closed-form optimization on saliency-guided image compression for HEVC-MSP. *IEEE Transactions on Multimedia*, 20(1):155–170, 2017. 8

[22] Zhicheng Li, Shiyin Qin, and Laurent Itti. Visual attention guided bit allocation in video compression. *Image and Vision Computing*, 29(1):1–14, 2011. 8

[23] Minghui Liao, Baoguang Shi, Xiang Bai, Xinggang Wang, and Wenyu Liu. Textboxes: A fast text detector with a single deep neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017. 2

[24] Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. Real-time scene text detection with differentiable binarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11474–11481, 2020. 2

[25] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 2

[26] Akis Linardos, Matthias Kummerer, Ori Press, and Matthias Bethge. DeepGaze IIE: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12919–12928, 2021. 1, 2, 7

[27] Nian Liu and Junwei Han. A deep spatial contextual long-term recurrent convolutional network for saliency detection. *IEEE Transactions on Image Processing*, 27(7):3264–3274, 2018. 1

[28] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 21–37. Springer, 2016. 2

[29] Yufan Liu, Minglang Qiao, Mai Xu, Bing Li, Weiming Hu, and Ali Borji. Learning to predict salient faces: A novel

visual-audio saliency model. In *European Conference on Computer Vision*, pages 413–429. Springer, 2020. 1

[30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021. 4

[31] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 2

[32] Shangbang Long, Jiaqiang Ruan, Wenjie Zhang, Xin He, Wenhao Wu, and Cong Yao. Textsnake: A flexible representation for detecting text of arbitrary shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 20–36, 2018. 2

[33] Jianxun Lou, Hanhe Lin, David Marshall, Dietmar Saupe, and Hantao Liu. TranSalNet: Visual saliency prediction using transformers. *arXiv preprint arXiv:2110.03593*, 2021. 2

[34] Stefan Mathe and Cristian Sminchisescu. Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(7):1408–1424, 2014. 3, 4

[35] Ethel Matin. Saccadic suppression: A review and an analysis. *Psychological bulletin*, 81(12):899, 1974. 4

[36] Lukas Neumann and Jiri Matas. A method for text localization and recognition in real-world images. In *Proceedings of the Asian Conference on Computer Vision*, pages 770–783. Springer, 2010. 2

[37] Junting Pan, Cristian Canton, Kevin McGuinness, Noel E. O'Connor, Jordi Torres, Elisa Sayrol, and Xavier and Giro-i Nieto. SalGAN: Visual saliency prediction with generative adversarial networks. In *arXiv*, January 2017. 2, 7

[38] Umesh Rajashekar, Ian van der Linde, Alan C Bovik, and Lawrence K Cormack. GAFFE: A gaze-attentive fixation finding engine. *IEEE Transactions on Image Processing*, 17(4):564–573, 2008. 2

[39] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 3

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 2, 4

[41] Kai Wang, Boris Babenko, and Serge Belongie. End-to-end scene text recognition. In *2011 International Conference on Computer Vision*, pages 1457–1464. IEEE, 2011. 2

[42] Wenguan Wang, Jianbing Shen, Jianwen Xie, Ming-Ming Cheng, Haibin Ling, and Ali Borji. Revisiting video saliency prediction in the deep learning era. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):220–237, 2019. 2, 3

[43] Yuxin Wang, Hongtao Xie, Zheng-Jun Zha, Mengting Xing, Zilong Fu, and Yongdong Zhang. Contournet: Taking a further step toward accurate arbitrary-shaped scene text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11753–11762, 2020. 2

[44] Mai Xu, Lai Jiang, Zhaoting Ye, and Zulin Wang. Bottom-up saliency detection with sparse representation of learnt texture atoms. *Pattern Recognition*, 60:348–360, 2016. 1, 2

[45] Xu-Cheng Yin, Xuwang Yin, Kaizhu Huang, and Hong-Wei Hao. Robust text detection in natural scene images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(5):970–983, 2013. 2

[46] Chengquan Zhang, Borong Liang, Zuming Huang, Mengyi En, Junyu Han, Errui Ding, and Xinghao Ding. Look more than once: An accurate detector for text of arbitrary shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10552–10561, 2019. 2

[47] Jianming Zhang and Stan Sclaroff. Exploiting surroundedness for saliency detection: A boolean map approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(5):889–902, 2015. 7

[48] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: An efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5551–5560, 2017. 2