SMARTTEXT: LEARNING TO GENERATE HARMONIOUS TEXTUAL LAYOUT OVER NATURAL IMAGE

Peiying Zhang, Chenhui Li, Changbo Wang*

School of Computer Science and Technology, East China Normal University
zhangpeiying17@gmail.com, {chli, cbwang}@cs.ecnu.edu.cn

ABSTRACT

Automatic typography is important because it helps designers avoid highly repetitive tasks and amateur users achieve high-quality textual layout designs. However, there are often many parameters that need to be adjusted in automatic typography work. In this paper, we propose an efficient content-aware learning-based framework to generate harmonious textual layout over natural image. Our method incorporates both semantic features and visual perception principles. First, we combine a semantic visual saliency detection network with diffusion equations and a text-region proposal algorithm to generate candidate text anchors with various positions and sizes. Second, we develop a deep scoring network to assess the aesthetic quality of the candidate results. We design multiple evaluations to compare our method with several baselines and a commercial poster design tool. The results demonstrate that our method can generate harmonious textual layout in various actual scenarios with better performance.

***Index Terms*—** Textual layout, saliency detection, image aesthetics, visual design, deep learning

1. INTRODUCTION

In the field of visual design, designers frequently invest a high amount of time fulfilling certain repetitive requirements, for example, designing a photograph, adjusting the text layout for different devices, modifying the foreground color, and evaluating visual effectiveness. In particular, the typography work of text and natural images is very cumbersome. Typography often requires high cost in terms of time, and it is difficult for an inexperienced designer to make higher-level creative designs. There are many advantages to utilizing an efficient typography design framework. First, our work can help designers avoid low-level and highly repetitive tasks while generating high-quality textual layout. Second, this framework is conducive to commercial advertising, allowing different users to view different advertising content. Third, with the increasing number of mobile devices and artificial intelligence technologies, the demand for creative graphic design is increasing. People without art knowledge can use a visual framework to achieve a high-level creative design.

Automatic typography is a challenging problem. On the one hand, numerous factors, such as image content, text layout and size, need to be considered. However, it is difficult to judge typography results because visual aesthetics rules are complex. In the past, there has been some work on automatic visual design for combining the text and image. However, in Text-to-Viz [1], there were many parameters that need to be adjusted for text layout and size selection. Additionally, data-driven methods [2] require amounts of data, and collecting and labeling high-quality data in the visual design area is costly. The evaluation of the design results is also not mature. Aiming at the above problems, our work focuses on the specific problem of graphic layout: textual layout over natural image. We take the text position and size into account and propose a content-aware learning-based framework for placing the text in the right location over the natural image as shown in Figure 1. In summary, the main contributions of our work are as follows:

(1) A learning-based model to optimize the text position and size can effectively generate a set of candidate text layout results.
(2) A deep scoring network with a smaller structure and fewer parameters is used to select the optimal textual layout result.
(3) Several reasonable evaluation approaches are applied to demonstrate the effectiveness of our method.

2. RELATED WORK

Our work is related to three aspects of the utilized techniques, namely, automatic graphic design, saliency detection, and image aesthetics assessment.

Automatic Graphic Design Automatic graphic design has been studied extensively. Previous studies have emphasized in automated layout design systems, such as a constraint-based recommendation system that can generate the design of magazine covers based on user preferences [3]. Yang et al. [4] revealed the effectiveness of optimization approach with aesthetic design principles. With the help of large database, some

*Corresponding author. This work was supported in part by the National Natural Science Foundation of China (No. 61802128, 61672237).

978-1-7281-1331-9/20/$31.00

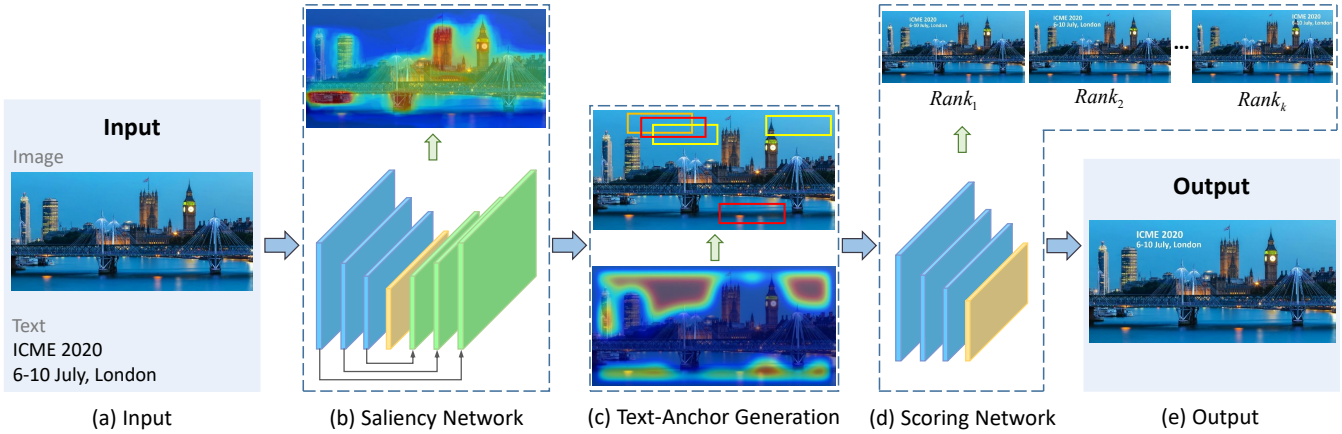

Fig. 1. The structure of our SmartText model. We receive a natural image and text as input and output a harmonious textual layout result. Our model has 3 main components. Saliency Network (b): a network to predict the semantic visual saliency map. Text-Region Proposal (c): an approach with two algorithms to generate several text anchors with different positions and sizes. Scoring Network (d): a network to assess the aesthetic quality of the candidate results and output the optimal one.

learning-based methods are proposed. Zhao et al. [5] designed a deep learning framework to score graphic designs with different personalities.

Saliency Detection Researchers have explored many efficient methods of saliency detection. Hou et al. [6] redefined the saliency of the image, after which saliency detection is mainly based on regional detection. Recently, approaches to saliency detection have been moved to deep learning. Relevant work includes a CNN-based visual attention prediction model [7] and an LSTM-based saliency attentive model [8].

Image Aesthetics Assessment Early approaches assessed image aesthetics by extracting handcrafted features and classifying the images based on these features [9]. With the development of deep learning methods for image classification, later work achieved significant improvement compared with the traditional approaches. Lu et al. [10] divided images into a global view and local view and used double-column CNNs based on the architecture of AlexNet for classification. Ma et al. [11] proposed a multi-patch aggregation network in which patch selection was based on saliency detection.

3. METHODS

3.1. Overview

Our goal is to generate a harmonious textual layout over a natural image that, given a natural image M_i and text T_i, outputs a harmonious textual layout result MT_o with the optimal text position T_p and size T_s. Unlike many other methods in general graphic design, which require original vector data, such as image categorization and element attributes, our models take input images in bitmap form. Our approach has 3 main components as shown in Figure 1:

- *Saliency Network*: a network that receives M_i as input and outputs the semantic visual saliency map S_M.
- *Text-Region Proposal*: an approach with two algorithms. One is the diffusion equation, which receives S_M as input and outputs the probabilistic density map PD_M. The other is text-anchor generation, which receives PD_M as input and outputs several anchors with different positions and sizes. We can obtain a set of candidate results $M_{o_{set}} = \{M_{o_1}, M_{o_2}, ..., M_{o_k}\}$ based on the region proposals above.
- *Scoring Network*: a network that assesses the aesthetic quality of the candidate results, which receives $M_{o_{set}}$ as input and outputs the score of each candidate result. Thus, we can obtain the optimal textual layout result $MT_o \in M_{o_{set}}$.

Before introducing our methods to each section, we define some basic aesthetic rules of the visual design layout.

- *Rule 1* The "text-region" should be a rectangular area, which is better not to overlap a complete significant design element, or to cross a strong background image or create discontinuities to offer more harmonious visual effects [12].
- *Rule 2* The "text-region" is better in the center of the sub-area and should look symmetric to associated design elements [4].
- *Rule 3* The layout result should gain a higher score in the aesthetic quality assessment.

3.2. Saliency Map Detection

The saliency map can help us understand the visual importance of different elements in a natural image or graphic design. Inspired by some state-of-the-art saliency detection

models, our architecture is based on an encoder-decoder network with a residual refinement module. We refer the reader to BASNet [13] for more details.

The original BASNet is used to detect salient objects, while we aim to obtain a real-valued visual saliency map considering some aesthetic criteria such as *Rule 1*. Human annotations tend to reflect similar relative importance being attributed to a whole design element [14]. Accordingly, to obtain more accurate regional segmentation and clear boundaries, we use a hybrid training loss:

$$L(\Theta) = L_B(\Theta) + L_S(\Theta) \tag{1}$$

where $L_B(\Theta)$ and $L_S(\Theta)$ denote BCEWithLogits loss [15] and SSIM loss [16], Θ is the BASNet model parameters.

Given the ground truth saliency map at each pixel p, $G_{M_p} \in [0, 1]$, over all pixels $p = 1, ..., N$, the BCEWithLogits loss is defined as:

$$L_B(\Theta) = -\frac{1}{N}\sum_{p=1}^{N}(G_{M_p}\log S_{M_p} + (1 - G_{M_p})\log(1 - S_{M_p})) \tag{2}$$

where S_{M_p} is the prediction of the saliency network.

Let $\mathrm{x} = \{x_p | p = 1, 2, ..., N\}$ and $\mathrm{y} = \{y_p | p = 1, 2, ..., N\}$ be two corresponding image patches extracted from G_{M_p} and S_{M_p}, and let μ_x, μ_y and σ_x^2, σ_y^2 be the mean and variance of x and y, σ_{xy} be their covariance. SSIM loss is defined as:

$$L_S(\Theta) = 1 - \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{3}$$

where $C_1 = 0.01^2$ and $C_2 = 0.03^2$ are scalar constants.

BCEWithLogits loss is used for semantic segmentation on all pixels, and it helps with accurate regional segmentation. SSIM loss can capture the structural information of the element in an image; hence, it helps clarify boundaries as their weights increase. Our predicted saliency results from the comparison with VisImportance [14] is shown in Figure 5 (b) and (c). The examples demonstrate that our model can predict a more accurate saliency map and clearer boundaries under various challenging scenarios.

3.3. Text-Region Proposal

After obtaining the original saliency map, to find the optimal text anchor, a straightforward solution is to list all possible text sizes and positions in the unimportant areas and rank the results. Unfortunately, this solution may cost substantial time on account of the large search space and it is hard to select the optimal result, since there are many unimportant areas with the same minor saliency values, as Figure 2(a) shows. To handle the problem above, we first develop a diffusion equation model to generate a text-driven probability density map. Then, we adopt a text-anchor generation algorithm to find the potential text anchors.

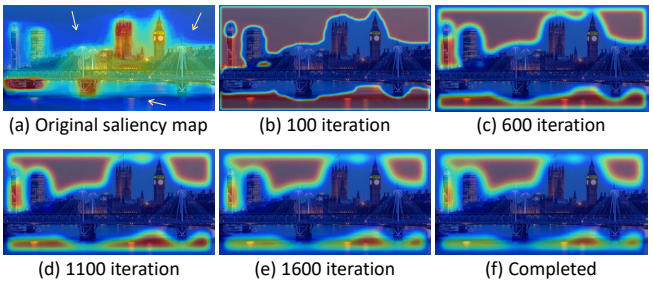

Fig. 2. The iterative process of the diffusion equation. Given the original saliency map, there are many unimportant areas with the same minor saliency values, such as areas marked by yellow arrows (a). (b)-(f) show the process of increasing the number of iterations.

Diffusion Equation

The diffusion equation is widely used in image denoising and scale space image analysis [17]. We utilize the method to generate a probability density map that incorporates the aforementioned *Rule 2*, and the diffusion equation can be described as Equation 4:

$$\begin{cases} PD_{M+1} = PD_M + \lambda(dX + dY) \\ dX = c_X\nabla_X(PD_M),\ dY = c_Y\nabla_Y(PD_M) \end{cases} \tag{4}$$

where ∇_X, ∇_Y are functions to calculate the gradient of each pixel from horizontal and vertical direction. c_X, c_Y represent the diffusion coefficient of the two directions.

Our diffusion equation aims to generate a text-driven probability density map instead of edge preserving filtering. With this consideration, we therefore set diffusion coefficients c_X, c_Y both to 1 in our implementation. Figure 2 shows an illustrative result.

Text-Anchor Generation

In the object detection area, there are many approaches used to generate bounding boxes, such as selective search [18] and region proposal network (RPN) [19]. However, these methods require rich features or ground-truth boxes, while collecting labels is expensive and impractical. In our specific task, we develop a new text-region proposal algorithm based on the probability density map instead of the original image.

Given the probability density map PD_M, we find that candidate regions with high probability values are more likely to be good text regions. Hence, we should generate more candidate anchors in high probability areas. Our algorithm assumes the text region has an aspect ratio $Ratio_T$. First, we divide the important parts of the probability density map into several connected regions and find the peak candidate text anchor of each connected region R_p, with the top-left corner (x_1, y_1) and the bottom-right corner (x_2, y_2). Then, we can generate other candidate anchors based on R_p, and the new top-left corner can be described as:

$$\{(x_{new_1}, y_{new_1}) | x_1 - \Delta x \le x_{new_1} \le x_1 + \Delta x, \\ y_1 - \Delta y \le y_{new_1} \le y_1 + \Delta y\} \tag{5}$$

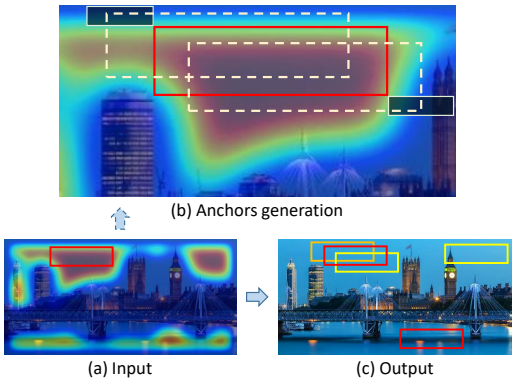

Fig. 3. Illustration of anchor generation. In (b), the light yellow anchors are generated based on the red anchor R_p, and the black areas show the adopted range of candidate anchors.

where $\Delta x = \delta |x_2 - x_1|, \Delta y = Ratio_T \Delta x$ means the adopted range of anchors, and $\delta \propto \sum_{i=1}^{|x_2-x_1|} \sum_{j=1}^{|y_2-y_1|} R_p(i,j)$ defines the deviation coefficient. If R_p is in a higher probability area, the adopted range Δx is larger; that is, there are more candidate anchors generated. R_p can be transformed into multiple scales with the same aspect ratio $Ratio_T$. Figure 3 shows an illustrative result.

3.4. Deep Scoring Network

After generating text-region proposals, we need to estimate the aesthetics score of each result. To solve this problem, we utilize a data-driven model to capture the perceptual differences between the good and bad textual layout results. Similar to some popular image aesthetics assessment methods, we build a binary classifier and use the class probabilities as the aesthetics scores. Our model architecture is based on ResNet101 [20], a residual learning framework. We create a textual layout dataset containing good results and bad results (Section 4.1). To obtain aesthetics scores, we modify the output layer and reserve the probabilities of *good* class only.

First, we take one of $M_{o_{set}} = \{M_{o_1}, M_{o_2}, ..., M_{o_k}\}$ as the input image. The convolutional layers of ResNet101 are used to extract features from M_{o_i}. Following is an FC layer, which divides images into two classes, one being good and the other being bad. A Softmax function is used to calculate the class probabilities. Figure 4 shows some predicted scores using our deep scoring network.

4. EXPERIMENT

4.1. Datasets

Saliency Map Detection: To find the best configuration for our SmartText model, we conduct experiments on two different datasets, SALICON (a large-scale dataset collected with the crowdsourcing paradigm) [21] and GDI (visual importance annotations for graphic designs) [14]. The experimental results in Table 1 show a comparison of our saliency network (Ours) with VisImportance [14]. We use the same evaluation metrics as VisImportance for GDI datasets. Higher CC, lower $RMSE$ and higher R^2 are better. Our saliency network improves performance in GDI datasets. The textual layout results are shown in Table 2. We observe that GDI outperforms SALICON, since the annotations in the GDI dataset are better aligned with the boundaries of the element. Therefore, we select GDI dataset to train our saliency detection model.

Fig. 4. Ranking textual layout results with different text positions and sizes. Predicted scores are shown below each image.

Table 1. Saliency Comparison of Ours with VisImportance

Method	Dataset	$CC \uparrow$	$RMSE \downarrow$	$R^2 \uparrow$
VisImportance [14]	GDI	0.811	0.181	0.617
Ours	GDI	**0.879**	**0.149**	**0.748**

Deep Scoring Network: There are many datasets for image aesthetics assessment, such as the AVA dataset [22], which contains a score distribution of approximately 255,000 images. However, most of the datasets are curated for the assessment of the composition and aesthetic quality of photographs, which are not applicable to our task of evaluating textual layout results. Hence, we construct a dataset of good and bad results (each class has 120 training images and 30 testing images). For good results, we use the keywords "good", "beautiful" and "poster" in the Google image search engine and select results satisfying the criteria of having text over natural images. For bad results, we generate a synthetic dataset automatically.

4.2. Results

Text with multiple sizes: The acceptable range of lines for users' input text is from one to five lines. Hence, it is necessary to assign multiple sizes to different lines. For example, subtitles are often smaller than the main title in posters. Our method can incorporate the user-specified input as a constraint. If users input n lines of texts, we consider all lines as

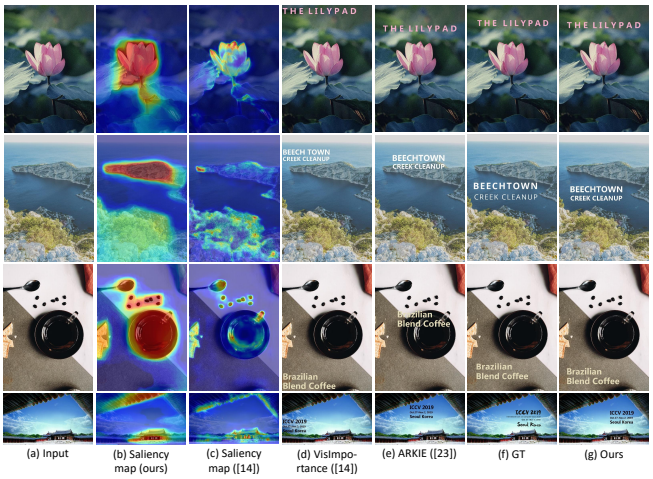

Fig. 5. Textual layout comparison of our SmartText model with VisImportance [14], ARKIE [23] and the ground truth. Our SmartText model is able to generate harmonious textual layout in various actual scenarios with better performance.

a whole grid, whose ratio is defined as:

$$Ratio_T = \frac{\max(Len(text))}{1 + \sum_{i=2}^{n} \sigma_i} \tag{6}$$

where $\max(Len(text))$ is the maximum length of all text, and σ_i is the ratio of the main title to other lines. The results are shown in Figure 5.

Text with mask: In natural images, designers tend to put text in well-defined regions with uniform color and texture. However, in some cases, the background images have a strong color contrast and complex texture. Therefore, we propose a method to judge whether it is necessary to apply a mask behind the text, which helps obtain clearer visual effects. We calculate the maximum number of pixels in each connected important region $np_{\max}$ in the probabilistic density map PD_M and use the following formula to judge:

$$np_{\max} < \frac{W \times H}{\mu_{\max}} \tag{7}$$

where $\mu_{\max}$ is the acceptable maximum scaling coefficient. $W \times H$ is the resolution of the given natural image M_i. The results are shown in Figure 6.

5. EVALUATION AND DISCUSSION

To evaluate the effectiveness of our SmartText model, we design several evaluation methods, including qualitative and quantitative analysis. We collect a new set of good textual layout results (50 images) and remove the existing text from these images. Thus, the background images obtained can be used for evaluation. Those removed texts are regarded as the ground truth. We compare our method to VisImportance [14],

Fig. 6. Text with mask. The mask can help to obtain clearer visual effects when the background image has a strong color contrast and complex texture.

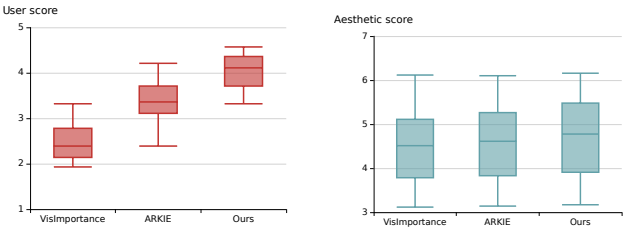

Fig. 7. Results of the visual effect analysis. Left is user scores, and right is aesthetic scores. The results indicate that our method can gain higher scores in both user study and image aesthetics assessment.

in which only visual saliency is considered, and an existing smart poster design tool ARKIE [23].

User Study To evaluate the global visual effects of the results, we select 20 images from the textual layout evaluation dataset and recruit 30 users to give a score for the 3 methods of generating results: VisImportance [14], ARKIE [23] and ours. Scores range from 1 (worst) to 5 (best).

Image Aesthetics Assessment In natural images, designers tend to put text in well-defined regions, such as regions with uniform color and texture, and good composition. Hence, the text-level image regions should gain high aesthetic scores. With this consideration, we use the NIMA model proposed by Talebi et al. [24] to assess the image aesthetics. NIMA contains a convolutional neural network that can predict the distribution of human opinion aesthetic scores. We crop the text-level image region and feed it into NIMA. Then we can obtain the predicted aesthetic scores for local visual effects.

Benchmark Evaluation With the good textual layout results as the ground-truth, we use the root-mean-square-error (RMSE) as the quantitative evaluation metric defined as follows:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (S_i - S_i^{gt})^2} \tag{8}$$

where N is the number of evaluation results, S_i is the generated results of our SmartText model, which can represent two metrics: text position and size. For position, S_i indicates top-left corner coordinates. S_i^{gt} is the corresponding ground truth. A smaller $RMSE$ value indicates a smaller error.

Table 2. Benchmark Evaluation

Method		Position	Size
VisImportance [14]		0.682	0.584
ARKIE [23]		0.461	0.480
Ours	SALICON	0.334	0.466
	GDI (no Sco.)	0.316	0.458
	GDI (Sco.)	**0.288**	**0.426**

Discussion Figure 7 shows the average user scores in user study and aesthetic scores in image aesthetics assessment. Our method is rated higher than others, which indicates that our textual layout results can get more harmonious visual effects in both global and local views.

In Figure 5, we find that ARKIE may simply use templates and lacks consideration of image content. Moreover, if we simply consider visual saliency and put the text on the most unimportant area, the text may be placed too close to the image edge in contrast to some aesthetic rules (Figure 5(d)). Thus, our content-aware learning-based framework is effective, as *Saliency Network* incorporates semantic features and *Text-Region Proposal* helps with visual perception principles. Table 2 shows a comparison of our method, VisImportance [14] and ARKIE [23] in benchmark evaluation. We can find that our method outperforms others in terms of both text position and size measures (2nd, 3rd and last rows). Since it is hard to decide the accurate text size without *Scoring Network*, we first generated candidate results with multiple text sizes and select randomly (2nd, 3rd and 5th rows). Compared 5th row with 6th row, the results imply that *Scoring Network* can help with the optimation of both text position and size.

6. CONCLUSIONS

In this paper, we present a new content-aware textual layout approach for visual text-image design. We design a feasible framework to place the text in a suitable location according to the requirements of the human vision system. We also offer some reasonable evaluations to compare our method with several baselines and a commercial poster design tool. The presented work can be applied to poster and advertisement design. Another potential direction of exploration is to generate watermarks for images.

7. REFERENCES

[1] Weiwei Cui, Xiaoyu Zhang, Yun Wang, He Huang, Bei Chen, Lei Fang, Haidong Zhang, Jian-Guan Lou, and Dongmei Zhang, "Text-to-viz: Automatic generation of infographics from proportion-related natural language statements," *IEEE TVCG*, vol. 26, no. 1, pp. 906–916, 2019.

[2] Xinru Zheng, Xiaotian Qiao, Ying Cao, and Rynson WH Lau, "Content-aware generative modeling of graphic design layouts," *ACM TOG*, vol. 38, no. 4, pp. 1–15, 2019.

[3] Ali Jahanian, Jerry Liu, Qian Lin, Daniel Tretter, Eamonn O'Brien-Strain, Seungyon Claire Lee, Nic Lyons, and Jan Allebach, "Recommendation system for automatic design of magazine covers," in *ACM IUI*, 2013.

[4] Xuyong Yang, Tao Mei, Ying-Qing Xu, Yong Rui, and Shipeng Li, "Automatic generation of visual-textual presentation layout," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 12, no. 2, pp. 1–22, 2016.

[5] Nanxuan Zhao, Ying Cao, and Rynson WH Lau, "What characterizes personalities of graphic designs?," *ACM TOG*, vol. 37, no. 4, pp. 1–15, 2018.

[6] Xiaodi Hou and Liqing Zhang, "Saliency detection: A spectral residual approach," in *IEEE CVPR*, 2007.

[7] Wenguan Wang and Jianbing Shen, "Deep visual attention prediction," *IEEE TIP*, vol. 27, no. 5, pp. 2368–2378, 2017.

[8] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara, "Predicting human eye fixations via an lstm-based saliency attentive model," *IEEE TIP*, vol. 27, no. 10, pp. 5142–5154, 2018.

[9] Yan Ke, Xiaoou Tang, and Feng Jing, "The design of high-level features for photo quality assessment," in *IEEE CVPR*, 2006.

[10] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z Wang, "Rapid: Rating pictorial aesthetics using deep learning," in *ACM MM*, 2014.

[11] Shuang Ma, Jing Liu, and Chang Wen Chen, "A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment," in *IEEE CVPR*, 2017.

[12] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman, "Synthetic data for text localisation in natural images," in *IEEE CVPR*, 2016.

[13] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand, "Basnet: Boundary-aware salient object detection," in *IEEE CVPR*, 2019.

[14] Zoya Bylinskii, Nam Wook Kim, Peter O'Donovan, Sami Alsheikh, Spandan Madan, Hanspeter Pfister, Fredo Durand, Bryan Russell, and Aaron Hertzmann, "Learning visual importance for graphic designs and data visualizations," in *ACM UIST*, 2017.

[15] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein, "A tutorial on the cross-entropy method," *Annals of operations research*, vol. 134, no. 1, pp. 19–67, 2005.

[16] Zhou Wang, Eero P Simoncelli, and Alan C Bovik, "Multiscale structural similarity for image quality assessment," in *IEEE Asilomar Conference on Signals, Systems & Computers*, 2003.

[17] Pietro Perona and Jitendra Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE TPAMI*, vol. 12, no. 7, pp. 629–639, 1990.

[18] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE CVPR*, 2014.

[19] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NeurIPS*, 2015.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, 2016.

[21] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao, "Salicon: Saliency in context," in *IEEE CVPR*, 2015.

[22] Naila Murray, Luca Marchesotti, and Florent Perronnin, "Ava: A large-scale database for aesthetic visual analysis," in *IEEE CVPR*, 2012.

[23] ARKIE, "Arkie," `https://www.arkie.cn/`, 2017.

[24] Hossein Talebi and Peyman Milanfar, "Nima: Neural image assessment," *IEEE TIP*, vol. 27, no. 8, pp. 3998–4011, 2018.