PosterLayout: A New Benchmark and Approach for Content-aware Visual-Textual Presentation Layout

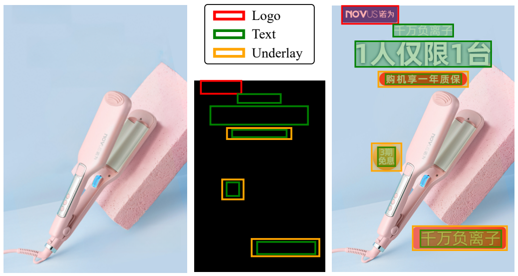
HsiaoYuan Hsu[1,2], Xiangteng He[1,2], Yuxin Peng[1,2*], Hao Kong[3] and Qing Zhang[3]
[1]Wangxuan Institute of Computer Technology, Peking University [2]National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University [3]Meituan
kslh99@stu.pku.edu.cn, {hexiangteng, pengyuxin, konghao}@pku.edu.cn, zhangqing31@meituan.com

Abstract

*Content-aware visual-textual presentation layout aims at arranging spatial space on the given canvas for pre-defined elements, including text, logo, and underlay, which is a key to automatic template-free creative graphic design. In practical applications, e.g., poster designs, the canvas is originally non-empty, and both inter-element relationships as well as inter-layer relationships should be concerned when generating a proper layout. A few recent works deal with them simultaneously, but they still suffer from poor graphic performance, such as a lack of layout variety or spatial non-alignment. Since content-aware visual-textual presentation layout is a novel task, we first construct a new dataset named **PKU PosterLayout**, which consists of 9,974 poster-layout pairs and 905 images, i.e., non-empty canvases. It is more challenging and useful for greater **layout variety**, **domain diversity**, and **content diversity**. Then, we propose design sequence formation (DSF) that reorganizes elements in layouts to imitate the design processes of human designers, and a novel CNN-LSTM-based conditional generative adversarial network (GAN) is presented to generate proper layouts. Specifically, the discriminator is design-sequence-aware and will supervise the "design" process of the generator. Experimental results verify the usefulness of the new benchmark and the effectiveness of the proposed approach, which achieves the best performance by generating suitable layouts for diverse canvases. The dataset and the source code are available at https://github.com/PKU-ICST-MIPL/PosterLayout-CVPR2023.*

1. Introduction

Nowadays, visual-textual presentation rendering informative and decorative elements on an image, i.e., canvas, is widely used to convey information, such as advertisement posters [5, 13, 16], magazines [20, 22], and so on [4, 10, 15].



(a) (b) (c)

Figure 1. Content-aware visual-textual presentation layout: (a) Non-empty canvas; (b) Content-aware layout; (c) An example of rendered presentation applying (b).

The basis of these creative works is the layout that indicates the spatial structure of the arranged elements, as shown in Fig. 1, which is also a key factor influencing their effectiveness and aesthetics. For their popularity and usefulness, not only experienced designers but also novice ones or "newbies" are commonly in need of creating them. People resort to pre-defined templates when they don't have enough prerequisites or need mass production. However, one can easily imagine that these templates harshly limit the flexibility and diversity of the presentations. These drawbacks of relying on templates hence highlight the importance and practicality of template-free creative graphic design, which can be preliminarily satisfied by automatically generating visual-textual presentation layouts.

With the advance in deep learning and big data, more and more data-driven approaches for visual-textual presentation layout have emerged in this decade. However, most of them have only been devoted to mining the relationship between elements and seldom concerned between layers, i.e., layout and canvas. Without proper constraints, elements are easily prone to cover the salient contents in the canvas, causing a severe occlusion problem. For example, in advertisement poster design, one of the most content-rich presentations, the product in the canvas shouldn't be over-occluded, which is no doubt. A few works [1, 23] deal with inter-element and

[*] Corresponding author.

inter-layer relationships simultaneously, but they still suffer from poor graphic performance, such as a lack of layout variety or spatial non-alignment. To this end, we propose a CNN-LSTM-based generative adversarial network (GAN) conditioned by the input canvases to generate layouts, which has a balanced performance on both graphic and content-aware metrics.

CNN-LSTM is proved effective in time series forecasting or behavior analysis tasks [6, 14]. To enable this time-sensitive model in layout generation, we propose design sequence formation (DSF) to generate design sequences that imitate the design processes of human designers. In particular, elements in layouts are reorganized to involve implicit temporal features, and less important ones can be discarded painlessly. It is in line with the logic of human-computer interaction logic [5] and has the potential to help train the LSTM model on a training set of size smaller than 20,000 [18]. GAN is a generative model that contains a discriminator and a generator gaming against each other to learn the distribution of training data. In the proposed design sequence GAN (DS-GAN), the discriminator is design-sequence-aware and will supervise the "design" process, i.e., generated layouts, of the generator under the constraints of the given canvas. As far as we know, this paper is the first adoption of CNN-LSTM in layout generation.

Since content-aware visual-textual presentation layout remains a novel task, there is only one public dataset in the field, and it has insufficient variety. In this paper, we first construct and release a new dataset and benchmark named **PKU PosterLayout**, which consists of 9,974 poster-layout pairs and 905 images, i.e., non-empty canvases. Each layout is represented by a set of elements labeled with class and bounding box. We collect data from multiple sources to guarantee diversity and variety in content, domain, and layout, supporting it as a challenging benchmark expected to encourage further research. Besides the dataset, we propose and clearly define new metrics to accompany the old ones, a total of eight graphic and content-aware metrics. They evaluate the layouts in terms of utilization, non-occlusion, and aesthetics. Both quantitative results and visualized results show that the proposed approach outperforms other approaches by generating proper layouts on diverse canvases.

We summarize the contribution of this paper as follows:

- A new and more challenging dataset and benchmark for content-aware visual-textual presentation layout, **PKU PosterLayout**, consists of 9,974 poster-layout pairs and 905 images, with greater diversity and variety in content, domain, and layout.
- An algorithm for design sequence formation (**DSF**) converts plain layout data into design sequences involving temporal features by imitating the design process of human designers.
- A CNN-LSTM-based GAN, design sequence GAN (**DS-GAN**), is conditioned by images and learns the distribution of design sequences to generate content-aware visual-textual presentation layouts. It makes a good trade-off between graphic and content-aware metrics, which outperforms the other approaches.

2. Related Work

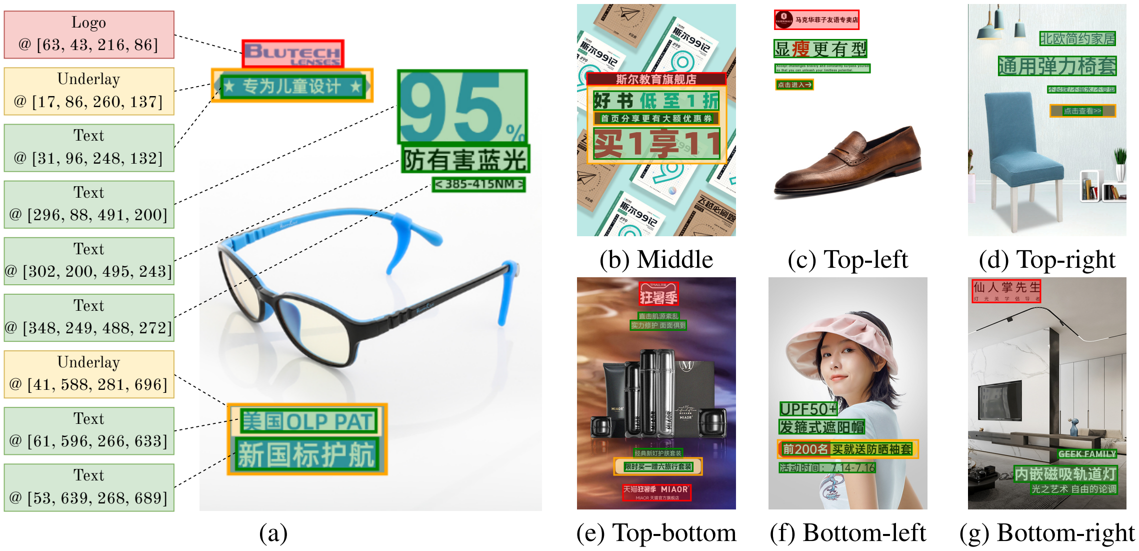
Research on content-agnostic visual-textual presentation has developed for a relatively long time, assuming the given canvas is empty. O'Donovan et al. [15] proposed an energy-based model that penalizes the part of layouts that violates pre-defined, complex design principles and thus could obtain a more desirable one after non-linear inverse optimization. The authors further presented a system [16] adopting this model with simpler principles, such as the size of elements and pair alignment, to alleviate time-consuming problem in heuristics.

Li et al. proposed LayoutGAN [12], taking a big step forward in data-driven approaches by introducing GANs in layout tasks. It adopted a differentiable wireframe rendering layer flattening layouts and canvases into wireframe images, remaining the discrimination process an image classification problem. In contrast, it differed from a conventional GAN in starting from a random initial layout that is primitively valid and modulating it into an eligible one instead of synthesizing layouts from fully random noise. The authors further presented an attribute-conditioned LayoutGAN [13] that guides the layout with the given element attributes, such as minimum size, fixed aspect ratio, and reading order of elements. Moreover, it accompanied elements dropout in the discrimination process, forcing the discriminator to be aware of the local pattern of layouts, which is helpful in visual-textual presentation layout. Besides the element attributes, Zheng et al. [22] demonstrated the efficiency of concerning the visual and textual semantics of the elements and presentation topics. They proposed an embedding network fusing cross-modal features to condition the GAN.

Kikuchi et al. proposed LayoutGAN++ [9] demonstrating an improvement in handling user-specific constraints by optimizing layout in latent space. It got rid of using wireframe images with respect to the findings that the rendering layer is unstable with a dataset of a limited size. Similarly, Lee et al. [10] were concerned with user-specific constraints and dealt with them using a graph neural network modeling elements as nodes and their relationships as edges. Clarification is needed that these user-specific constraints are merely inter-layout and insufficient for the task interested in this paper. Specifically, content-aware visual-textual presentation layout concerns both inter-layout and inter-layer relationships, i.e., layout and canvas, which is driven by canvas with no mandatory constraints attached. However, the ideas behind these content-agnostic approaches are still

	Status	# data		Layout		Canvas	Content category
		Layout	Canvas	Element types	Complex?		
NDN [10]	Private	500	NaN	Text, logo, RoI, image, brand name, button	No	Empty	Car
ICVT [1]	Private	117,624	166	Text, logo, underlay	Yes	Non-empty	Not given
CGL-GAN [23]	Public	60,548	1,000	Text, logo, underlay, embellishment	No	Non-empty	Cosmetics, electronics, clothing, etc.
PKU PosterLayout (Ours)	Public	9,974	905	Text, logo, underlay	Yes	Non-empty	Cosmetics, electronics, clothing, delicatessen, toys/instruments, etc.

Table 1. Comparison of properties of benchmark for visual-textual presentation layout.


Logo @ [63, 43, 216, 86]
Underlay @ [17, 86, 260, 137]
Text @ [31, 96, 248, 132]
Text @ [296, 88, 491, 200]
Text @ [302, 200, 495, 243]
Text @ [348, 249, 488, 272]
Underlay @ [41, 588, 281, 696]
Text @ [61, 596, 266, 633]
Text @ [53, 639, 268, 689]
BLUTECH LENSES
★ 专为儿童设计 ★
95%
防有害蓝光
<385-415NM>
美国OLP PAT
新国标护航

(a)

(b) Middle

(c) Top-left

(d) Top-right

(e) Top-bottom

(f) Bottom-left

(g) Bottom-right

Figure 2. Examples of poster-layout pairs in PKU PosterLayout.

worthy of exploring to enhance the performance of content-aware ones.

For content-aware visual-textual presentation layout, Zhuo et al. proposed CGL-GAN [23] utilizing the standard transformer block, which is the first and most relevant work in the disciplines. The encoder and decoder had the visual features of canvas and the embedding of layout as input, respectively. Therefore, the self-attention and cross-attention in the decoder can simultaneously model the inter-layout and inter-layer relationships. Although experimental results demonstrated its usefulness in improving content-aware metrics, it had a relatively poor performance in graphic metrics, especially the spatial non-alignment. Following almost the same ideas as CGL-GAN, most recently, Cao et al. [1] proposed an image-conditioned variational transformer with the proposed geometry-aligned fusion formula applied in the cross-attention layer. Not surprisingly, it suffered from a similar problem, especially the undesired overlap between elements. What these approaches encounter encourages us to research and propose a novel approach making a trade-off between two types of metrics to generate the most proper layout.

3. A New Benchmark: PKU PosterLayout

A few datasets related to visual-textual presentation layouts for posters have been presented in previous works. Tab. 1 shows the properties of these datasets. In detail, the pre-defined element types, target canvas, scale, i.e., amounts of layout and canvas data, and diversity, i.e., categories of posters, are compared. Since some datasets are still private to this day, statistics are from their source papers. NDN [10] presented a banner-layout dataset composed of 500 car advertisements for validating its content-agnostic approach, which doesn't completely align with the target task and suffers from scale problems and a lack of diversity. ICVT [1] presented a large-scale 117k poster-layout dataset, strongly supporting training content-aware approaches. However, its usefulness for validation is doubted since the excessively small testing set, i.e., 166 canvases, and the unknown diversity. CGL-GAN [23] presented a dataset with 60k poster-layout pairs and 1k canvases. It seems to be overall satisfactory; however, it collects all data from a single source, covers only a few categories with a disproportionate proportion, and does not in-

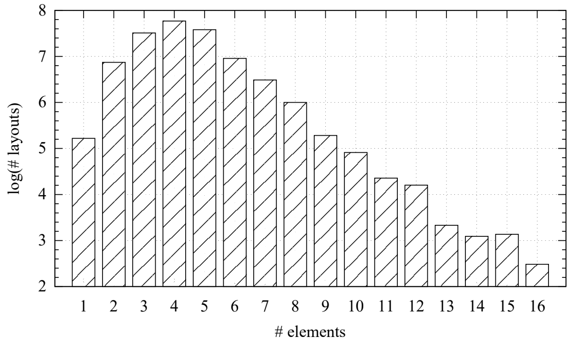


Figure 3. Statistics on layout variety in PKU PosterLayout.

volve complex layouts containing more than 10 elements. These drawbacks limit its generality.

To this end, a new dataset and benchmark, PKU PosterLayout, is constructed in this paper. It shows advantages in domain diversity, content diversity, and layout variety. We first adopt posters from a subset of an e-commerce posters dataset [7] and define element types. Each poster is paired with a layout composed of a variable-length set n of elements, $L = \{e_i \mid i = 0, 1, ..., n-1\}$. Each element e is represented with its type c and bounding box $b = [x_1, y_1, x_2, y_2]$, standing for the top-left and bottom-down coordinates. Without loss of generality, three element types are defined, including text, logo, and underlay, as shown in Fig. 2. Logos are graphic elements that indicate brand names or promotion activities, and underlays are decorations below or around any elements. While underlay-below-underlay is allowed, a strict rule to follow when labeling is that every underlay must decorate at least one text or logo independently, as shown in Fig. 2(b). Finally, texts are all the other informative elements that are not logos. Manually labeling from scratch is inefficient and infeasible, and thus an iterative scheme is applied with the help of an object detection model [3]. All poster-layout pairs are refined and examined to be correct by human annotators, and then the rendered elements on posters are erased using a Fourier-convolution-based inpainting method [17]. Fig. 3 shows statistics on layout variety observed briefly from the number of elements in layouts, indicating a broad distribution. It is emphasized that there are several complex layouts with more than 10 elements in PKU PosterLayout. Therefore, it can be sufficient for layout tasks under complicated settings, such as [8], while [23] cannot.

Afterward, we collect background and product images via searches to create canvases of various qualities while carefully keeping the numbers of ones even in each category. There are totally nine categories, including food/drinks, cosmetics/accessories, electronics/office supplies, toys/instruments, clothing, sports/transportation, groceries, appliances/decor, and fresh produce. Eventually, 9,974 poster-layout pairs and 905 canvases constitute PKU PosterLayout, a median-scale dataset and benchmark with guaranteed layout variety, domain diversity, and content diversity.

4. Proposed Approach

After analyzing the weaknesses of previous works, a novel generative model that makes a good trade-off between graphic and content-aware performance is proposed in this paper. This section will start with the basic ideas of design sequences and the algorithm used to form them automatically. Afterward, the proposed CNN-LSTM-based GAN empowered by these sequences will be expounded. An overview of the proposed approach is shown in Fig. 4.

4.1. Design Sequence Formation

Referring to Guo et al.'s work studying artistic creation in human-computer interaction [5], modeling human designers' behaviors can be a promising way toward content-aware visual-textual presentation layout. Specifically, behaviors are represented by design sequences, which indicate the order human designers place elements on canvases. Since the layout data is plain and short of this information, an algorithm for DSF is presented, shown in Algorithm 1. The main principle of DSF is putting the most informative, significant elements at the front and vice versa, which is a basic understanding of the design process. Following a rule of thumb, the conspicuousness of logos is influenced by reading order, e.g., from top-left to bottom-right [13], and thus their top-left coordinates are chosen as a criterion. For

Algorithm 1 Algorithm for design sequence formation

```
Require: Layout $L = \{e_i \mid i = 0, 1, ..., n-1\}$
Ensure: $R$ = design sequence formed from $L$
$R \leftarrow \{\}$
$l \leftarrow \{e_i \mid c_i$ is *logo*$\}$
$t \leftarrow \{e_i \mid c_i$ is *text*$\}$
$u \leftarrow \{e_i \mid c_i$ is *underlay*$\}$
Sort $(l, (y_{top}, x_{left}))$ in ascending order
Sort $(t, area)$ in descending order
$G \leftarrow$ groups of $l$ and $t$ with commonly overlaid $u$
                                                ▷ Stable merging
Queue $Inst \leftarrow Concat(l, t)$
while $Inst \neq \emptyset$ do
    $Inst' \leftarrow pop(Inst)$
    if $Inst' \notin R$ then
        $G' \leftarrow G \supset Inst'$
        $u' \leftarrow u$ overlaid $G'$
        $push(R, G')$
        $push(R, u')$
    end if
end while
```

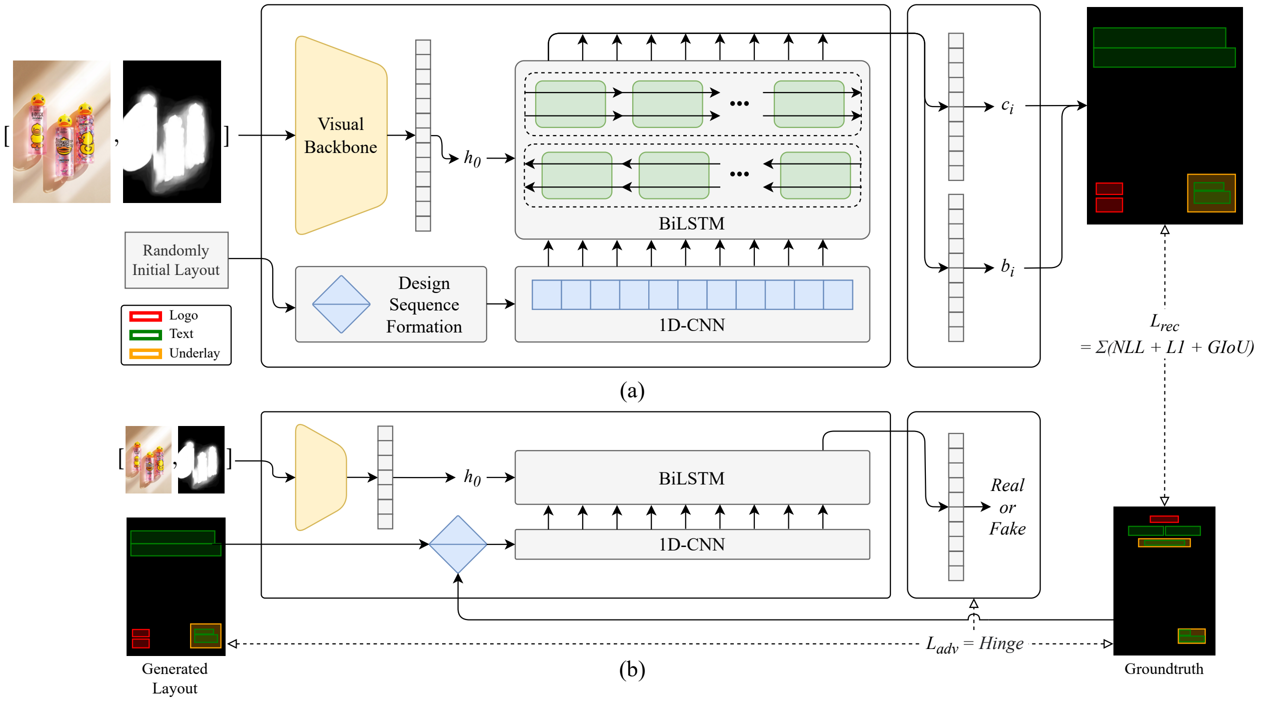


Figure 4. An overview of the proposed approach: (a) generator of DS-GAN; (b) discriminator of DS-GAN.

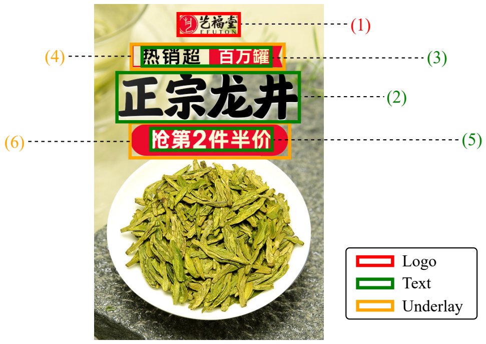


Figure 5. A design sequence formed by the proposed DSF.

texts, their areas are chosen, which is obvious. Underlays, as the name implies, must be below others and thus won't be put into the sequence until all elements overlaying it get arranged. Fig. 5 gives an example of a design sequence formed by DSF. By maintaining the descending order of element importance, design sequences will not be severely disrupted even if the last few elements are discarded. This additional benefit enhances the CNN-LSTM model adopted later, for it takes a fixed-length input.

4.2. Design Sequence GAN

GAN is a generative model that contains two submodels, a discriminator and a generator. The former is devoted to creating fake samples that sufficiently cheat the latter, and the latter is devoted to making a conscious distinction between the real and the fake samples. As time goes by, they keep gaming against each other and can learn the distribution of training data. In this paper, a CNN-LSTM-based GAN is proposed. CNN-LSTM is a hybrid model well-known for its potential to handle time-series forecasting and behavior analysis [6, 14]. With the help of the DSF presented above, layout generation is transferred to be a novel behavior modeling problem and thus becomes solvable for CNN-LSTM.

Even more sensible, another reason to use LSTM is that it is triggered by an initial hidden state, exactly where the conditions can be attached. Specifically, instead of roughly concatenating visual and layout embedding, like what conditional GANs do to make models image-conditioned, DS-GAN perceives visual content by explicitly initializing its hidden state h_0 with visual features F, as

$$\begin{aligned} F &= \text{ResNet-FPN}([I, max(S_{\text{PFPN}}, S_{\text{BASNet'}})]) \\ h_0 &= Linear(F), \end{aligned} \tag{1}$$

where I represents an input image, i.e., repaired poster or canvas, S_i represents a saliency map constructed by different domain salient detection methods [19, 21], $[f_1, f_2]$ represents feature concatenation along channel-axis, operator ResNet-FPN refers to that presented in [23], operator max represents a pixel-wise maximum operation, and operator $Linear$ represents a fully connected layer. Then, design sequences formed from randomly initialized layouts [12] (in the generator) or real/fake samples (in the discriminator) are directly input to the CNN-LSTM model, with no extra embedding layers needed.

	Target	$Val\uparrow$	$Ove\downarrow$	$Ali\downarrow$	$Und_l\uparrow$	$Und_s\uparrow$	$Uti\uparrow$	$Occ\downarrow$	$Rea\downarrow$
SmartText [11]	T	-	-	-	-	-	0.0849	**0.0912**	**0.1528**
CGL-GAN [23]	V-T	0.7066	0.0605	0.0062	**0.8624**	0.4043	0.2257	0.1546	0.1715
DS-GAN (Ours)	V-T	**0.8788**	**0.0220**	**0.0046**	0.8315	**0.4320**	**0.2541**	0.2088	0.1874

Table 2. Comparison of quantitative results.

Both generator and discriminator are built with the structure expounded above, i.e., a CNN-LSTM model that takes visual features as the initial hidden state and design sequences as input. Though, of course, they have different ending parts for their respective target. In the generator, two fully connected layers are further cascaded to decode the output of the CNN-LSTM model into the type and bounding box of each element in generated design sequence, turning valid elements into a layout. Besides the adversarial loss L_{adv} discussed later, a reconstruction loss L_{rec} [2] aggregating negative log-likelihood (NLL) loss, L1 loss, and generalized intersection of union (IoU) loss helps to train the generator. By contrast, the discriminator ends up with only one fully connected layer that classifies whether input design sequences are real samples. The adversarial loss that guides the discriminator and, thus, the generator is the hinge loss.

5. Experiment

For validating the proposed approach, quantitative indicators of content-aware visual-textual presentation layout are elaborated. Then, several experiments based on the proposed benchmark, PKU PosterLayout, are conducted, enabling the comparison between existing approaches and the proposed one. Results show that DS-GAN with DSF achieves the best performance by generating the most proper layouts on diverse canvases while making a good trade-off between the two aspects, i.e., graphic and content-aware metrics.

5.1. Evaluation Metrics

Eight metrics in two aspects are defined as follows. Some of them are newly proposed or clearly defined for the first time, including Val ↑, Und ↑, and Occ ↓. The up arrow indicates a higher value is better, and vice versa.

Graphic metrics Validity, annotated as $Val\uparrow$, is the ratio of valid elements to all elements in the layout, where the area within the canvas of a valid element must be greater than 0.1% of the canvas. Note that all remaining metrics consider only valid elements. Overlay, annotated as Ove ↓, is the average IoU of all pairs of elements except for underlay. *Non*-alignment, annotated as Ali ↓, is the extent of spatial non-alignment between elements, referring to [13]. Underlay effectiveness, annotated as $Und\uparrow$, is the ratio of valid underlay elements to total underlay elements, where a valid one u must truly decorate at least one non-underlay element $Inst$. Subscript l means loose, calculating $\frac{area(u \cap Inst)}{area(Inst)}$ for each pair, and Und_l takes the maximum value. Subscript s means strict, scoring the underlay as 1 if there is a non-underlay element completely inside, otherwise, 0, and Und_s takes the average score.

Content-aware metrics Utility, annotated as Uti ↑, is the utilization rate of space suitable for arranging elements, implemented by the negative image S' of the compounded saliency map S, as mentioned in Sec. 4.2. In particular, the denominator is the total pixel values of S', and the numerator is that of S' masked all areas without elements. In opposite, occlusion, annotated as Occ ↓, is the average pixel value of areas covered by elements in S. *Un*readability, annotated as Rea ↓, is the non-flatness of regions that text elements are solely put on, referring to CGL-GAN.

5.2. Implementation Details

Considering the complexity of respective tasks, the generator of DS-GAN is with ResNet50 backbone and 4-layer CNN-BiLSTM, while the discriminator is with ResNet18 backbone and 2-layer CNN-BiLSTM. When training the network, layout data c is in one-hot vector form, b is in $[x_c, y_c, w, h]$ form, standing for center coordinates and the width, height of the bounding box, and the batch size is 128. The weights in reconstruction loss are 2, 5, and 2 for NLL, L1, and generalized IoU loss, respectively. The weight of reconstruction is constantly 1, and that of adversarial loss increases linearly from 0 to 1 in a warm-up of 100 epochs. The entire network is trained for 300 epochs. Adam optimizers are used, of which learning rates are initialized as: 10^{-4}, 10^{-5} for the generator and its visual backbone, and 10^{-3}, 10^{-4} for discriminator and its backbone. All experiments are carried out with Pytorch framework and four NVIDIA A100-SXM4-80GB GPUs.

5.3. Comparison with State-of-the-arts

The proposed approach is compared with SmartText [11] and CGL-GAN [23]. While the latter was mentioned in Sec. 2, the former was not, for it was for content-aware textual presentation layout. Besides the lack of comparable existing approaches, we choose SmartText for two reasons. First, based on a saliency-aware region proposal, it intrinsically experts in avoiding unreadability and occlusion, so a comparison is worthy to see how long the target task still has to go. Second, however, SmartText suffers from a fatal problem in that it puts all elements into the selected anchor

	$Val \uparrow$	$Ove \downarrow$	$Ali \downarrow$	$Und_l \uparrow$	$Und_s \uparrow$	$Uti \uparrow$	$Occ \downarrow$	$Rea \downarrow$
without CNN-LSTM	0.6765	0.0888	0.0112	0.0106	0.0000	0.2155	0.2804	0.2015
with CNN-LSTM (DS-GAN)	**0.8788**	**0.0220**	**0.0046**	**0.8315**	**0.4320**	**0.2541**	**0.2088**	**0.1874**

Table 3. Ablation study on CNN-LSTM model.

	$Val \uparrow$	$Ove \downarrow$	$Ali \downarrow$	$Und_l \uparrow$	$Und_s \uparrow$	$Uti \uparrow$	$Occ \downarrow$	$Rea \downarrow$	$AE \downarrow$
Random	**1.0000** (+0.1454)	0.0881 (+0.0666)	0.0062 (+0.0007)	0.7417 (-0.1380)	0.3243 (-0.1499)	0.2240 (-0.0328)	0.2475 (+0.0361)	**0.1909** (+0.0035)	0.5730
Geometric	0.9667 (+0.1215)	**0.0261** (+0.0026)	0.0050 (+0.0004)	0.7849 (-0.0824)	0.4433 (-0.0757)	0.2439 (-0.0170)	0.2482 (+0.0438)	0.1937 (+0.0052)	0.3486
DSF-based (DS-GAN-8)	0.9572 (+0.0784)	0.0362 (+0.0142)	**0.0043** (-0.0003)	**0.8850** (+0.0535)	**0.5824** (+0.1504)	**0.2526** (-0.0015)	**0.2341** (+0.0253)	0.1910 (+0.0036)	**0.3272**

Table 4. Ablation study on design sequence formation.

box in a rigid way, making graphic metrics purposeless. It may be sufficient for textual presentation but not for visual-textual one, showing the difference between them and the importance of the novel task discussed in this paper.

In the experiments, 905 canvases from nine categories in PKU PosterLayout are used for validation. For fairness, in all approaches, the length of each design sequence (or just plain layout) is the maximum number of elements in all layout data, which means non-objects are padded if necessary. We leave the effects of discarding less significant elements in the ablation study. Tab. 2 shows the quantitative results of the proposed DS-GAN and compared approaches. Values in the first column denote the target presentation of the corresponding approach, where T means textual, and V means visual. It is observed that the proposed DS-GAN achieves the best performance in almost all graphic metrics as expected, for it benefits from CNN-LSTM, which helps recognize a pattern of sequences of geometric parameters, i.e., bounding boxes. First of all, most elements in its generated layouts are valid, which is the basis. Impressively, it does an excellent job of avoiding undesired overlap between elements, obtaining $Ove \downarrow$ only 37% of CGL-GAN's. It also has a good result on avoiding spatial non-alignment, reducing $Ali \downarrow$ by more than 25%. In terms of loose underlay effectiveness, DS-GAN is a little behind, but on the opposite, the leading in strict underlay effectiveness exactly proves it has not abused underlay elements. As for content-aware metrics, both CGL-GAN and ours fall behind SmartText, showing there is indeed a long way for research to go toward content-aware visual-textual presentation layout. Nevertheless, DS-GAN can achieve nearly comparable performance with CGL-GAN on avoiding unreadability.

Fig. 6 shows several layouts generated from these approaches. Visualized results verify the high $Uti \uparrow$ of DS-GAN, while cases like the one shown in column (b) fundamentally explain why it conversely gets behind in terms of $Occ \downarrow$ and $Rea \downarrow$. That is, for its ability and inclination toward exploiting all available regions. These cases tell that slight occlusion sometimes actually brings more appealing layouts. Moreover, DS-GAN's abilities to avoid overlay or non-alignment between elements are observed, and it can handle diverse canvases that annoy others. The proposal-based method, i.e., SmartText, is problematic with cases shown in columns (c) and (d), where the only salient object almost fills the canvas. CGL-GAN is prone to lose directions when confronting a canvas with a circle-outline object or numerous salient objects, as shown in columns (f) and (g). Moreover, column (h) shows that DS-GAN actively generates more *complex* layouts than others, even though some unpleasant overlay is witnessed. Since DS-GAN already does best on $Ove \downarrow$, further improvements can be challenging. We claim it as a promising direction, especially using our PKU PosterLayout dataset, which contains complex layouts. Overall, the proposed approach generates more appealing layouts for diverse canvases by balancing its performance on graphics and content awareness. Both quantitative and visualized results validate this conclusion.

5.4. Ablation Study

Since the CNN-LSTM model is the key to DS-GAN, an ablation experiment is conducted by remaining only the last fully connected layers to evaluate its effectiveness. As shown in Tab. 3, the monotonically decreasing performance strongly demonstrates the necessity of the model. It is expected since CNN-LSTM helps behavior analysis, which is the main working logic of DS-GAN.

To gain insight into the effect of DSF, an ablation study is carried out. Remember that an important capability of DSF is maintaining a descending order of element importance in the design sequence, which indicates that discarding the least important elements should be trivial to the final performance. Therefore, the dependent variable is the length of the design sequences. We set it as (a) the maximum number of elements in all layout data or (b) 8, i.e., DS-GAN-8, and verify the effects on three different formation strategies, including (1) *random* order, (2) ascending order

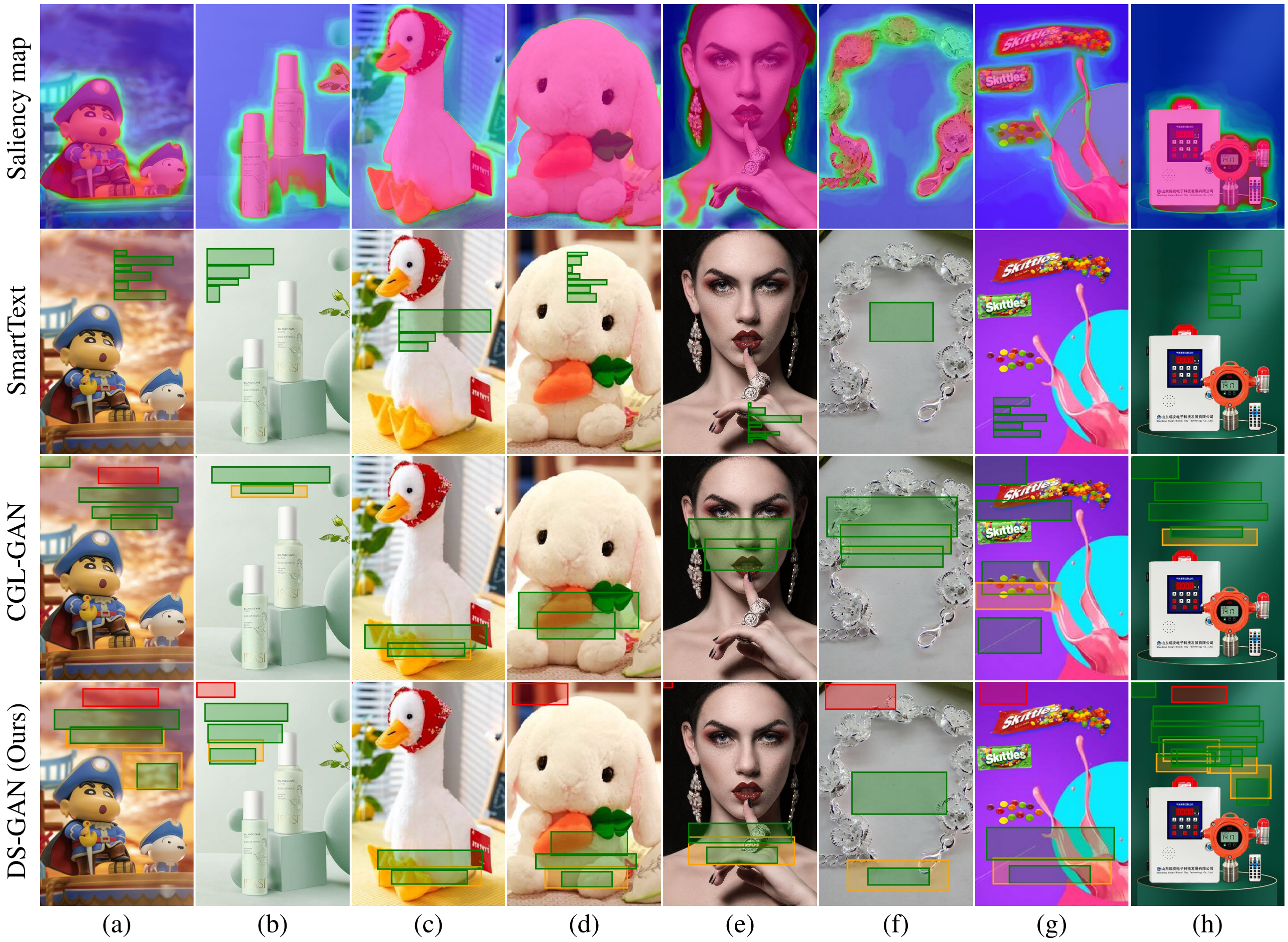


Figure 6. Comparison of layouts generated by different approaches.

of top-left coordinates, i.e., *geometric*, and (3) the proposed *DSF-based* order. Experimental results are shown in Tab. 4. Except for the last column, the values represent the metrics under (b), and those in the parentheses are the difference between metrics under (a) and (b), calculated by (b) - (a) and thus show what happens if almost half of the elements are discarded. Observing the results, we find the DSF-based gets the best performance as expected, for that random and geometric strategies may discard essential elements, especially the random one. Moreover, by aggregating the total absolute difference, annotated as $AE \downarrow$, we verify that the perturbation brought by the number of elements is the most trivial when adopting the proposed DSF algorithm.

6. Conclusion

In this paper, we construct a new dataset and benchmark for content-aware visual-textual presentation layouts, named **PKU PosterLayout**. With satisfactory layout variety, domain diversity, and content diversity, it is more challenging and expected to encourage further research. We also propose a generative approach, **DS-GAN**, with the DSF algorithm to treat layout generation as a behavior modeling problem. The DSF algorithm can form plain layout data into design sequences and help DS-GAN learn the pattern better. Several experiments are conducted to verify (1) the usefulness of the proposed benchmark and (2) the effectiveness of the proposed approach that generates suitable layouts for diverse canvases.

The future works mainly lie in two aspects: (1) Further improving content-aware performance without violating graphic performance, which may be done by replacing the off-the-shelf saliency detection method with a dedicated one or involving it in the end-to-end training process. (2) Devoted to high-quality complex layout generation, which is promising and achievable utilizing the first public dataset containing complex layouts– PKU PosterLayout.

Acknowledgements

This work was supported by grants from the National Natural Science Foundation of China (61925201, 62132001, 62272013, U22B2048) and Meituan.

References

[1] Yunning Cao, Ye Ma, Min Zhou, Chuanbin Liu, Hongtao Xie, Tiezheng Ge, and Yuning Jiang. Geometry aligned variational transformer for image-conditioned layout generation. In *Proceedings of the ACM International Conference on Multimedia*, pages 1561–1571, 2022. 1, 3

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*, pages 213–229, 2020. 6

[3] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 4

[4] Mengxi Guo, Danqing Huang, and Xiaodong Xie. The layout generation algorithm of graphic design based on transformer-CVAE. In *Proceedings of the International Conference on Signal Processing and Machine Learning*, pages 219–224, 2021. 1

[5] Shunan Guo, Zhuochen Jin, Fuling Sun, Jingwen Li, Zhaorui Li, Yang Shi, and Nan Cao. Vinci: an intelligent graphic design system for generating advertising posters. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2021. 1, 2, 4

[6] Hsiao-Yuan Hsu, Nai-Hsin Cheng, and Chun-Wei Tsai. A deep learning-based integrated algorithm for misbehavior detection system in VANETs. In *Proceedings of the ACM International Conference on Intelligent Computing and its Emerging Applications*, pages 53–58, 2021. 2, 5

[7] Gangwei Jiang, Shiyao Wang, Tiezheng Ge, Yuning Jiang, Ying Wei, and Defu Lian. Self-supervised text erasing with controllable image synthesis. In *Proceedings of the ACM International Conference on Multimedia*, page 1973–1983, 2022. 4

[8] Zhaoyun Jiang, Shizhao Sun, Jihua Zhu, Jian-Guang Lou, and Dongmei Zhang. Coarse-to-fine generative modeling for graphic layouts. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1096–1103, 2022. 4

[9] Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, and Kota Yamaguchi. Constrained graphic layout generation via latent optimization. In *Proceedings of the ACM International Conference on Multimedia*, pages 88–96, 2021. 2

[10] Hsin-Ying Lee, Lu Jiang, Irfan Essa, Phuong B Le, Haifeng Gong, Ming-Hsuan Yang, and Weilong Yang. Neural design network: Graphic layout generation with constraints. In *Proceedings of the European Conference on Computer Vision*, pages 491–506, 2020. 1, 2, 3

[11] Chenhui Li, Peiying Zhang, and Changbo Wang. Harmonious textual layout generation over natural images via deep aesthetics learning. *IEEE Transactions on Multimedia*, 2021. 6

[12] Jianan Li, Jimei Yang, Aaron Hertzmann, Jianming Zhang, and Tingfa Xu. LayoutGAN: Synthesizing graphic layouts with vector-wireframe adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7):2388–2399, 2020. 2, 5

[13] Jianan Li, Jimei Yang, Jianming Zhang, Chang Liu, Christina Wang, and Tingfa Xu. Attribute-conditioned layout GAN for automatic graphic design. *IEEE Transactions on Visualization and Computer Graphics*, 27(10):4039–4048, 2020. 1, 2, 4, 6

[14] Ronald Mutegeki and Dong Seog Han. A CNN-LSTM approach to human activity recognition. In *Proceedings of the International Conference on Artificial Intelligence in Information and Communication*, pages 362–366. IEEE, 2020. 2, 5

[15] Peter O'Donovan, Aseem Agarwala, and Aaron Hertzmann. Learning layouts for single-pagegraphic designs. *IEEE Transactions on Visualization and Computer Graphics*, 20(8):1200–1213, 2014. 1, 2

[16] Peter O'Donovan, Aseem Agarwala, and Aaron Hertzmann. Designscape: Design with interactive layout suggestions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1221–1224, 2015. 1, 2

[17] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2149–2159, 2022. 4

[18] Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. Order matters: Sequence to sequence for sets. In *Proceedings of the International Conference on Machine Learning*, pages 1–11, 2016. 2

[19] Bo Wang, Quan Chen, Min Zhou, Zhiqiang Zhang, Xiaogang Jin, and Kun Gai. Progressive feature polishing network for salient object detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12128–12135, 2020. 5

[20] 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*, 12(2):1–22, 2016. 1

[21] Peiying Zhang, Chenhui Li, and Changbo Wang. SmartText: Learning to generate harmonious textual layout over natural image. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 1–6, 2020. 5

[22] Xinru Zheng, Xiaotian Qiao, Ying Cao, and Rynson WH Lau. Content-aware generative modeling of graphic design layouts. *ACM Transactions on Graphics*, 38(4):1–15, 2019. 1, 2

[23] Min Zhou, Chenchen Xu, Ye Ma, Tiezheng Ge, Yuning Jiang, and Weiwei Xu. Composition-aware graphic layout GAN for visual-textual presentation designs. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 4995–5001, 2022. 1, 3, 4, 5, 6