Displays 87 (2025) 102970

Contents lists available at ScienceDirect

Displays

journal homepage: www.elsevier.com/locate/displa

Dual discriminator GANs with multi-focus label matching for image-aware layout generation☆

Chenchen Xu [a,b], Kaixin Han [b], Min Zhou [c], Weiwei Xu [b,*]

[a] Anhui Normal University, Wuhu, China
[b] State Key Lab of CAD&CG, Zhejiang University, Hangzhou, China
[c] Alibaba Group, Hangzhou, China

ARTICLE INFO

Keywords:
Image-aware layout generation
Dual discriminators
DD-GAN
Multi-focus label matching

ABSTRACT

Image-aware layout generation involves arranging graphic elements, including logo, text, underlay, and embellishment, at the appropriate position on the canvas, constituting a fundamental step in poster design. This task requires considering both the relationships among elements and the interaction between elements and images. However, existing layout generation models struggle to simultaneously satisfy explicit aesthetic principles like alignment and non-overlapping, along with implicit aesthetic principles related to the harmonious composition of images and elements. To overcome these challenges, this paper designs a GAN with dual discriminators, called DD-GAN, to generate graphic layouts according to image contents. In addition, we introduce a multi-focus label matching method to provide richer supervision and optimize model training. The incorporation of multi-focus label matching not only accelerates convergence during training but also enables the model to better capture both explicit and implicit aesthetic principles in image-aware layout generation. Quantitative and qualitative evaluations consistently demonstrate that DD-GAN, coupled with multi-focus label matching, achieves state-of-the-art performance, producing high-quality image-aware graphic layouts for advertising posters.

1. Introduction

Graphic layout generation, as shown in Fig. 1, involves arranging various classes of 2D elements, such as logos, texts, underlays, and other embellishment elements, in appropriate positions. This task is fundamental to the design of posters [1–3], magazines [4,5], and webpages [6–8]. Recently, layout generation based on deep generative models like Generative Adversarial Networks (GANs) has attracted increasing interest [9–11]. Conditional GANs provide precise control over the layout generation process by incorporating conditions such as image content and graphic element attributes (e.g., category, area, and aspect ratio) [1,12]. Notably, image content plays a crucial role in producing image-aware layouts for posters and magazines [2,4,13]. However, many current models define the layout generation problem simply as arranging graphic elements on a blank canvas [14,15]. Although some conditional methods [12,16] have been proposed to guide the layout generation process, the majority are based on graphic element attributes and rarely consider image content.

In poster layout (image-aware layout) design tasks, it is essential to consider both the geometric relationships among graphical elements, such as overlap and alignment, and the coordination between the image content and the graphical layout. This requires the model to simultaneously model the geometric relationships between graphical elements and the interactions between the graphical layout and the image content. This paper focuses on generating image-aware graphic layouts for advertising posters, considering both explicit aesthetic principles (such as alignment and non-overlap) and implicit aesthetic principles (such as the harmonious composition of images and elements).

For the image-aware layout generation task, Zhou et al. [1] annotated and utilized inpainting [17,18] to remove graphic elements from designed posters, constructing the CGL-Dataset with paired images and graphic layouts. Although the CGL-Dataset significantly benefits the training of image-aware networks, a domain gap [19] exists between inpainted posters (source domain data) and clean product images (target domain data). To bridge the domain gap, CGL-GAN [1] applies Gaussian blur on the inpainted poster to eliminate inpainted artifacts. Although this strategy effectively removes these artifacts, it may damage the color and texture details of images, leading to unpleasing results in implicit aesthetic principles (content-relevant metrics). Xu et al. [2]

☆ This paper was recommended for publication by Prof. Guangtao Zhai.

* Corresponding author at: State Key Lab of CAD&CG, Zhejiang University, Hangzhou, China.
E-mail addresses: xuchenchen@zju.edu.cn (C. Xu), hankx@zju.edu.cn (K. Han), yunqi.zm@alibaba-inc.com (M. Zhou), xww@cad.zju.edu.cn (W. Xu).

https://doi.org/10.1016/j.displa.2025.102970
Received 26 July 2024; Received in revised form 24 December 2024; Accepted 12 January 2025
Available online 31 January 2025
0141-9382/

combined unsupervised domain adaptation techniques [20] to propose PDA-GAN for generating image-aware layouts. To balance the influence of the two domains, PDA-GAN [2], trained with few annotated samples, produces unsatisfactory results in explicit principles (graphic metrics) due to a lack of rich supervisory signals. Our research aims to develop a new method that can fully harness the generative power of GAN-based models while considering both explicit and implicit aesthetic principles.

In this paper, we design a GAN with dual discriminators, called DD-GAN, to generate image-aware layouts for posters that adhere to both explicit and implicit aesthetic principles. One discriminator, similar to the pixel-level discriminator in PDA-GAN [2], is employed to bridge the domain gap and explore implicit aesthetic principles. Another global discriminator, structurally resembling the generator, is introduced to differentiate whether the images and layouts match and assess the coherence of the composition of graphic elements. The layout generator structure is designed based on the transformer architecture.

In the CGL-Dataset, individual images typically contain up to ten elements, leading to a query count of 10 for the generator transformer module. Given that the number of annotated elements for the majority of samples is less than 5, only a small subset of queries is supervised during each training step. To address this issue, we introduce multi-focus label matching (MLM) to offer richer supervision and enhance data efficiency. This strategy involves replicating annotated element information multiple times before conducting Hungarian matching [21]. The duplicated annotated element information is then matched with the model's output through Hungarian matching, computing a reconstruction loss that supervises the model. This approach effectively addresses the issue of limited annotated elements and significantly enriches the supervisory signals for the transformer module.

The experimental results demonstrate that DD-GAN, coupled with multi-focus label matching, achieves state-of-the-art (SOTA) performance on both graphic metrics (explicit aesthetic principles) and content-relevant metrics (implicit aesthetic principles). Compared to CGL-GAN with comparable performance in graphic metrics, our method demonstrates relative improvements in content-relevant metrics, including background complexity, occlusion subject degree, and occlusion product degree metrics, by 4.61%, 9.94%, and 12.77%, respectively. Similarly, in contrast to PDA-GAN with comparable performance in content-relevant metrics, our method shows relative improvements over graphic metrics, encompassing layout overlap, underlay overlap, and layout alignment degree metrics, by 71.95%, 2.75%, and 7.32%, respectively. Moreover, comprehensive ablation experiments reveal that multi-focus label matching not only improves the performance of our model, but also enhances the performance of other models. The training loss curves further validate that multi-focus label matching can accelerate model convergence.

In summary, this paper comprises the following contributions:

- We introduce a multi-focus label matching strategy to provide richer supervision for the model. This strategy can be easily combined with different models, accelerating training convergence and improving models' performances.
- We exploit the generative power of the GAN-based model to propose a DD-GAN, featuring dual discriminators, to generate image-aware layouts that simultaneously satisfy both explicit and implicit aesthetic principles of advertising posters.
- Both quantitative and qualitative evaluations demonstrate that our model, utilizing multi-focus label matching, achieves SOTA performance and outperforms in generating image-aware graphic layouts for posters.

2. Related work

Recently, the importance of layout in graphic design has driven significant research efforts in layout generation. We categorize existing works into two types based on their consideration of image content: image-agnostic layout generation methods [22,23], which solely study the relationship between graphic elements (explicit aesthetic principles), and image-aware layout generation methods [2,24–26], which simultaneously explore both the relationships among internal graphic elements and between layouts and images (explicit and implicit aesthetic principles).

Fig. 1. Examples of image-conditioned advertising posters graphic layouts generation. Our model generates graphic layouts (middle) with multiple elements conditioned on product images (left). The designer or even automatic rendering programs can utilize graphic layouts to render advertising posters (right).

Image-agnostic layout generation. Early works [27–31] embed design rules into manually crafted energy functions but struggled to generate complex and diverse layout results. LayoutGAN [10] is the first to introduce deep generative networks for layout generation tasks, promoting data-driven approaches to accomplish layout generation. LayoutVAE [14] and LayoutVTN [15] both utilize variational autoencoder (VAE) techniques and are autoregressive methods. As the field evolved, a prominent research trend involved imposing additional constraints [32–34] on models to achieve desired results. These constraints are in various forms, including element attributes, scene graphs, and partial layouts. However, in a nutshell, models with these constraints primarily concentrate on modeling the internal relationship between graphic elements and often neglect the relationship between layouts and images. Consequently, they generate graphic layouts that do not align with implicit aesthetic principles.

Image-aware layout generation. High-quality training datasets for image-aware layout generation are difficult to obtain because they

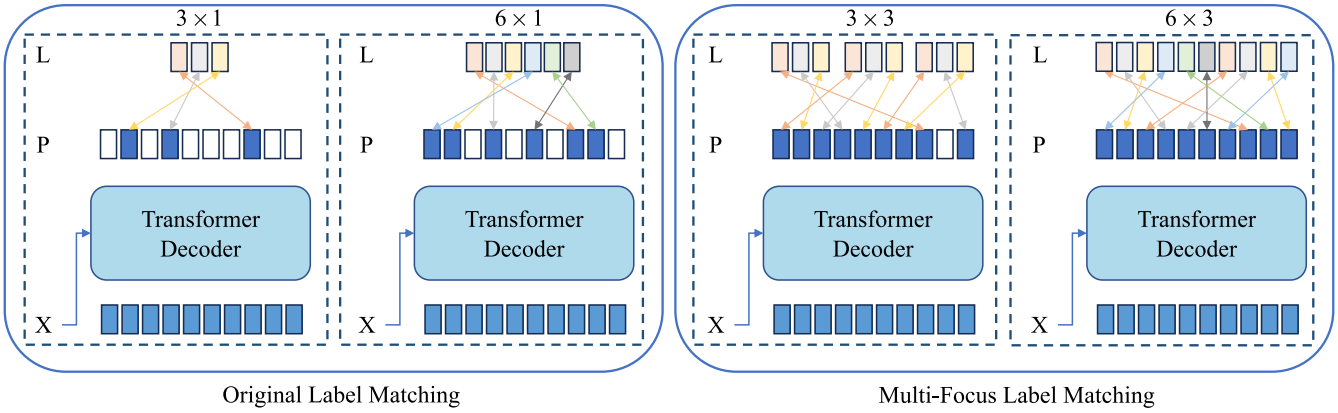

Fig. 2. Multi-focus label matching. The left side of this figure presents examples with three and six annotated elements, respectively. White queries indicate instances without matches and, consequently, lack corresponding losses for supervision. On the right side, the elements on the left are replicated three times according to multi-focus label matching, as detailed in Section 3.1.

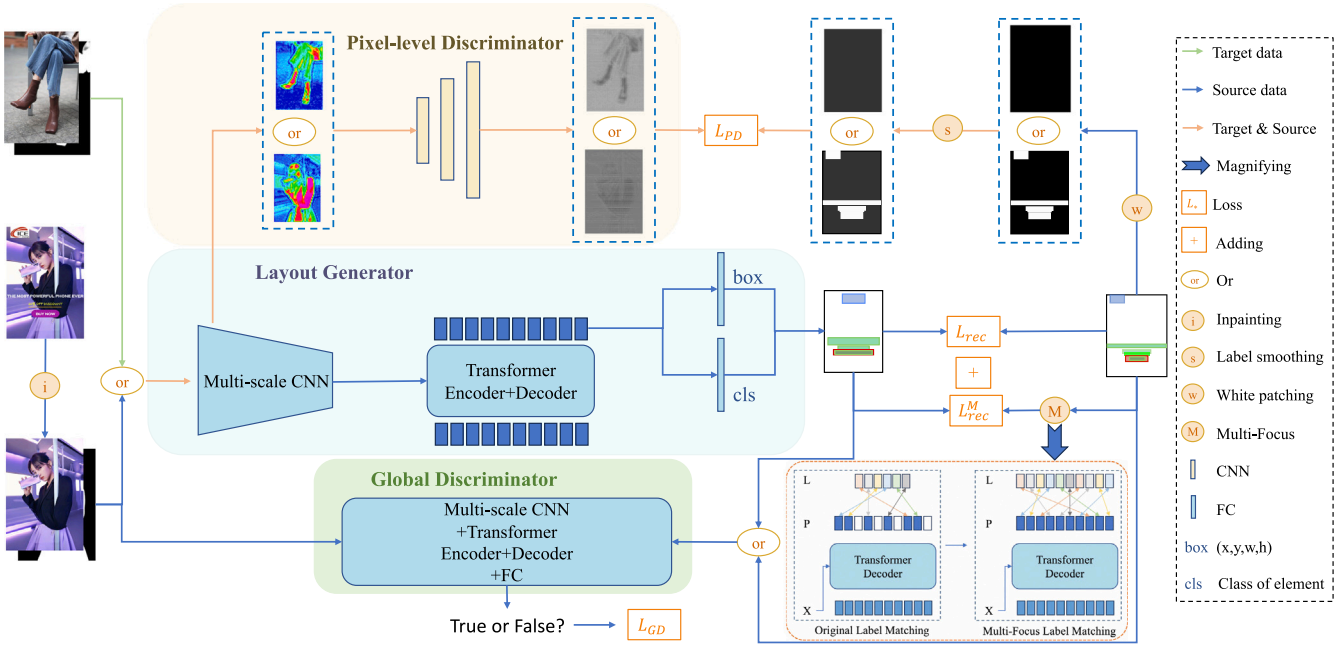

Fig. 3. The architecture of our network. Clean product images (target domain data) are solely processed by the multi-scale CNN of the layout generator and fed into the pixel-level discriminator since they lack annotated labels. Inpainted posters (source domain data) are processed by the layout generator, the pixel-level discriminator, and the global discriminator. The definition of MLM can be found in Section 3.1. The weight coefficients of the various loss functions are provided in Section 3.2. The global discriminator is structurally similar to the generator, while the pixel-level discriminator consists of three transposed convolutional layers with filter size 3 × 3 and stride 2. The annotations for the symbols in the figure are shown within the dashed box on the right.

require numerous professional stylists to design layouts that provide paired clean images and annotated layouts. ContentGAN [4] utilizes white patches to mask graphic elements on magazine pages and replaces clean images with processed pages for training. This approach is the first to propose modeling the relationships among internal graphic elements and between layouts and images. For this task, Zhou et al. [1] propose to collect designed poster images to construct a dataset with paired product images and graphic layouts. The graphic elements imposed on the poster images are removed through image inpainting [17] and annotated with their geometric arrangements in the posters, resulting in the state-of-the-art CGL-Dataset comprising 54,546 paired data items. Although the CGL-Dataset significantly benefits the training of image-aware networks, a domain gap [35–37] exists between inpainted posters (source domain data) and clean product images (target domain data). To narrow the domain gap, CGL-GAN [1] applies Gaussian blur to the entire poster to eliminate inpainted artifacts and generate image-aware layouts. However, Gaussian blur may damage the delicate color and texture details of images, leading to the generation of layouts that are unpleasing in implicit aesthetic principles (content-relevant metrics). Xu et al. [2] combined unsupervised domain adaptation techniques [19,38–41] to design a GAN with a pixel-level discriminator, named PDA-GAN [2], for generating image-aware layouts. To balance the influence of the two domains, PDA-GAN used a limited number of annotated samples during training. Although PDA-

GAN effectively models implicit aesthetic principles, it struggles with explicit aesthetic principles due to the lack of rich supervisory signals. Therefore, we introduce a multi-focus label matching strategy to provide richer supervision for image-aware layout generation models and design a DD-GAN, featuring dual discriminators, based on PDA-GAN. This approach simultaneously considers both explicit and implicit aesthetic principles.

3. Our method

This paper aims to develop a model with high learning capacity that can capture both explicit and implicit aesthetic principles to generate high-quality image-aware layouts. To achieve this, we introduce a multi-focus label matching strategy to provide richer supervision for the image-aware layout generation model and design a DD-GAN with dual discriminators based on PDA-GAN. This section provides detailed descriptions of the multi-focus label matching strategy and our network architecture.

3.1. Multi-focus label matching

Due to the exceptional representational power of transformers in modeling information and learning complex feature representations, our model is designed based on transformer modules [42]. The inherent power of transformers lies in their ability to capture intricate patterns and dependencies within data, making them well-suited for various tasks. However, in scenarios where real-world samples (target domain data) are scarce, the performance of models built on transformers tends to suffer. This is particularly attributed to transformers defaulting to a one-to-one matching strategy for label assignment, implying that only a limited number of transformer queries receive positive supervision signals in each training iteration [43,44]. We perform the original label matching between the predictions $\mathbf{P}$, labels $\mathbf{L}$ pair and compute the reconstruction loss as follows:

$$L_{rec} = \mathcal{L}_{Hungarian}(\mathbf{P}, \mathbf{L}) \tag{1}$$

where $\mathbf{P} = \{p_1, p_2, \ldots, p_{10}\}$ represent the fixed set of ten queries from the transformer's output, $\mathbf{L} = \{l_1, \ldots, l_i\}$ $(0 < i \leq 10)$ indicate the annotated labels information. $\mathcal{L}_{Hungarian}$ following DETR [45]. As a result, the transformer modules heavily rely on obtaining adequate supervision from an extensive dataset or through prolonged training epochs to generalize effectively and enhance performance. Therefore, exploring strategies to alleviate the impact of limited positive supervision is crucial to ensure the model's robustness and generalization capabilities in the face of insufficient datasets and limited computing devices.

In tasks such as object recognition and detection, image data augmentation techniques are commonly employed. Techniques like cropping, scaling, and rotation increase the diversity of training samples, thereby enhancing the robustness and generalization capabilities of the models. However, these methods are not suitable for the image-aware graphic layout generation task. The reason is that changes in image content lead to significant alterations in the corresponding graphic layout, including the categories, quantities, and positional relationships of elements within the layout. Such variations mean that previously annotated layout information based on specific content will no longer be applicable, rendering it ineffective for continued supervised training. Therefore, traditional image data augmentation methods do not meet the specific requirements of the image-aware graphic layout generation task.

In this paper, we propose a label augmentation strategy called multi-focus label matching (MLM) to enhance the supervision signal for transformers by repeating ground truth (GT) labels before Hungarian matching. As depicted in Fig. 2, the method involves replicating the annotated labels $\mathbf{L}$ m times while keeping $\mathbf{P}$ unchanged. For example, the first sample in Fig. 2 includes three annotated elements, each replicated three times ($m = 3$), resulting in a total of nine GT labels $\tilde{\mathbf{L}}$. The Hungarian matching is then performed between the fixed set of ten queries $\mathbf{P}$ from the transformer output and the replicated labels $\tilde{\mathbf{L}}$. In contrast to the original label matching, where only three queries were supervised, the current MLM with nine queries provides effective supervision signals. When the number of replicated elements exceeds 10, the first 10 GT labels are retained, and the rest are discarded. For instance, in the second scenario shown in Fig. 2 with six annotated elements, the original six GT labels are retained and the first four replicated GT labels are added. The reconstruction loss of MLM can be computed similarly to Eq. (1):

$$L^{M}_{rec} = \mathcal{L}_{Hungarian}(\mathbf{P}, \tilde{\mathbf{L}}) \tag{2}$$

where $\tilde{\mathbf{L}} = \{\mathbf{L^1}, \ldots, \mathbf{L^m}\} = \{l^1_1, \ldots, l^1_i, \ldots, l^m_1, \ldots, l^m_i, \}$. $\mathbf{L^1} = \cdots = \mathbf{L^m}$, similar $l^1_i = \cdots = l^m_i$. If $i \times m$ exceeds ten, the first ten labels are retained and the rest are discarded. MLM offers a more enriched supervisory signal to the transformers, leading to enhanced model performance and convergence, especially in scenarios with a scarcity of annotated elements. Additionally, these richer and more stable layout reconstruction supervisory signals can significantly improve the stability of the early-stage training of the generative adversarial network. It is worth noting that MLM may prompt the model to generate multiple similar bounding boxes at the same location. Consequently, our model needs to incorporate non-maximum suppression (NMS) [46] to filter out duplicate predictions.

3.2. DD-GAN

We explore the generative power of a GAN-based model to design a DD-GAN, incorporating dual discriminators, to generate image-aware layouts that conform to the explicit and implicit aesthetic principles of advertising posters. Fig. 3 illustrates our network architecture, comprising three main components: a layout generator, a pixel-level discriminator (PD) [2,39], and a global discriminator (GD) [47,48]. Please note that unlabeled target domain data exclusively undergo processing by the multi-scale CNN [49,50] of the generator and are subsequently fed into PD to assist the generator in bridging the domain gap and studying implicit aesthetic principles. These data will not be propagated to other modules of the generator or GD as they lack annotated labels.

The architecture of the layout generator network follows the DETR [45] principle, consisting of three modules: a multi-scale convolutional neural network (CNN), a transformer encoder–decoder, and two fully connected layers (FCs). The multi-scale CNN takes the concatenation of the inpainted poster $\boldsymbol{x}_{inp}$ with its saliency map [51] $\boldsymbol{x}^{sal}_{inp}$ (or the clean product image $\boldsymbol{x}_{img}$ with its saliency map $\boldsymbol{x}^{sal}_{img}$) as input and extracts image features. The encoder utilizes the standard transformer architecture to further refine the image features. The decoder employs cross-attentions to learn the relationship between image content and graphic layout. Both the encoder and decoder consist of six transformer blocks, each equipped with ten queries. The decoder features of each query are passed through two FCs to predict the corresponding class and box coordinates. The predicted class and box information $\mathbf{P}$ are obtained using the softmax and sigmoid functions, respectively. These predictions are then used to compute L_{rec} and L^{M}_{rec} as specified in Eqs. (1) and (2).

The first discriminator, PD, consists of three transposed convolutional layers designed to bridge the domain gap between source and target domain data. PD is connected to the shallow-level feature map and computes the GAN loss for each input-image pixel. Since inpainted areas occupy a small proportion of the input image [2], we apply label smoothing [9,52] to pixels not in the inpainted area (those pixels with value 0 in the white patch map), which we refer to as one-target label smoothing. Specifically, we adjust the value of 0 to 0.2 in the ground truth white patch map to enhance the generalization ability of the trained model. The loss L_{PD} (or L^{G}_{PD}), with label smoothing applied for updating PD (or generator), can be calculated similarly to [2].

Table 1
Comparison with content-aware methods. Bold and underlined numbers denote the best and second best respectively. ↓ (or ↑) means the smaller (or bigger) value, the better.

Model	R_{com} ↓	R_{shm} ↓	R_{sub} ↓	R_{ove} ↓	R_{und} ↑	R_{ali} ↓	R_{occ} ↑
ContentGAN [4]	45.59	17.08	1.143	0.0397	0.8626	0.0071	93.4
CGL-GAN [1]	34.11	15.41	0.783	0.0413	0.9400	0.0098	99.7
PDA-GAN [2]	**32.07**	13.56	0.727	0.0353	0.9205	0.0109	99.7
IUC-Layout [26]	33.06	15.93	0.826	0.0174	0.9221	**0.0055**	99.9
Ours	32.54	**13.44**	**0.705**	**0.0099**	**0.9458**	0.0101	**100.0**

Table 2
User study. P_e (P_b) represents the percentage of eligible-selected (best-selected) layouts. The symbol * denotes the professional group.

Model	P_e ↑	P_b ↑	P_e^* ↑	P_b^* ↑
CGL-GAN [1]	19.91	15.62	18.79	17.15
PDA-GAN [2]	24.31	27.35	18.91	28.88
IUC-Layout [26]	20.22	27.17	23.62	22.45
DD-GAN (Ours)	**35.56**	**29.86**	**38.68**	**31.52**

Table 3
Comparison with content-agnostic methods. LT and VTN represent LayoutTransformer [22] and LayoutVTN [15], respectively.

Model	R_{com} ↓	R_{shm} ↓	R_{sub} ↓	R_{ove} ↓	R_{und} ↑	R_{ali} ↓	R_{occ} ↑
LT	40.92	21.08	1.310	0.0156	0.9516	0.0049	**100.0**
VTN	41.77	22.21	1.323	0.0130	**0.9698**	**0.0047**	99.9
Ours	**32.54**	**13.44**	**0.705**	**0.0099**	0.9458	0.0101	**100.0**

The second discriminator, GD, has a structure similar to the above generator, taking concatenated $\boldsymbol{x}_{inp}$ with $\boldsymbol{x}_{inp}^{sal}$ and replicated labels $\tilde{\mathbf{L}}$ or predicting layouts $\mathbf{P}$ as input. GD, with only one fully connected layer, judges whether the input image-layout pairs are true or false, resulting in the loss L_{GD} (or L_{GD}^{G}) for updating GD (or Generator).

Therefore, the training loss for the layout generator network is as follows:

$$L_G = L_{rec} + \alpha * L_{rec}^{M} + \beta * L_{PD}^{G} + \gamma * L_{GD}^{G}, \tag{3}$$

where the weight coefficients α, β, and γ in this work are set to 1, 6, and 8, respectively. By incorporating richer reconstruction supervisory signals through L_{rec}^{M} and jointly training with adversarial losses, the training of the generative adversarial network becomes more stable. This enhanced stability enables the model to generate layouts that simultaneously satisfy both explicit and implicit aesthetic principles.

4. Experiments

In this section, we primarily compare our model with SOTA layout generation methods and present ablation studies for MLM.

4.1. Implementation details

We implemented our model in PyTorch 1.7.1 and utilized the Adam optimizer [55] for training. The initial learning rates are set to 10^{-5} for the generator backbone and 10^{-4} for the remaining part of the model. The model is trained for 300 epochs with a batch size of 128, and all learning rates are reduced by a factor of 10 after 200 epochs. To ensure fair experimental comparisons, we employed the CGL-Dataset for both training and test datasets, resizing the inpainted posters and clean product images to 240 × 350 following the procedures of CGL-GAN and PDA-GAN. The total training time is approximately 9 h using 16 NVIDIA V100 GPUs. To encourage more researchers to participate in the research and exploration of graphic layout generation tasks, we will release our model code to this community.

Table 4
Ablations for different discriminators. GD and PD represent the global discriminator and pixel-level discriminator, respectively. The symbol × denotes DD-GAN without the inclusion of the corresponding discriminator, while the symbol ✓ signifies DD-GAN with the discriminator incorporated. Bold and underlined numbers denote the best and second best respectively.

GD	PD	R_{com} ↓	R_{shm} ↓	R_{sub} ↓	R_{ove} ↓	R_{und} ↑	R_{ali} ↓	R_{occ} ↑
×	×	34.12	14.55	0.806	0.0247	0.9321	0.0089	99.1
✓	×	**32.53**	15.66	0.776	0.0211	0.9306	**0.0081**	99.9
×	✓	32.95	**12.88**	**0.628**	0.0125	**0.9517**	0.0109	99.9
✓	✓	32.54	13.44	0.705	**0.0099**	0.9458	0.0101	**100.0**

Table 5
Different configurations of the weights β and γ in Eq. (3). The weights of L_{rec}, L_{rec}^{M}, and L_{PD}^{G} are fixed at 1, 1, and 6, respectively. The effect of different weights for the two discriminators in DD-GAN is evaluated by adjusting the weight of L_{GD}^{G}.

β	γ	R_{com} ↓	R_{shm} ↓	R_{sub} ↓	R_{ove} ↓	R_{und} ↑	R_{ali} ↓	R_{occ} ↑
6	4	32.55	13.11	**0.667**	0.0093	0.9321	0.0127	99.9
6	6	40.60	**9.36**	0.678	**0.0077**	0.7586	**0.0094**	99.7
6	8	**32.54**	13.44	0.705	0.0099	0.9458	0.0101	**100.0**
6	10	33.12	13.37	0.686	0.0107	**0.9568**	0.0118	**100.0**
6	12	33.36	14.64	0.757	0.0139	0.9544	0.0130	**100.0**

Table 6
Different weight configurations for L_{rec} and L_{rec}^{M} in Eq. (3). θ represents the weight of L_{rec}. The values of β and γ are fixed at 6 and 8, respectively. The effect of different weights for L_{rec} and L_{rec}^{M} in DD-GAN is evaluated.

θ	α	R_{com} ↓	R_{shm} ↓	R_{sub} ↓	R_{ove} ↓	R_{und} ↑	R_{ali} ↓	R_{occ} ↑
1	1	32.54	13.44	0.705	0.0099	0.9458	0.0101	100.0
1	2	33.13	13.99	0.691	0.0138	0.9361	0.0113	100.0
2	1	31.42	12.56	0.647	0.0074	0.9505	0.0113	99.6
1	3	32.15	14.27	0.697	0.0156	0.9550	0.0108	99.5
3	1	34.49	13.13	0.678	0.0066	0.9015	0.0090	98.0
1	4	32.99	18.57	0.813	0.0092	0.9374	0.0079	99.6
4	1	34.37	16.01	0.801	0.0140	0.9430	0.0093	99.9

4.2. Evaluation metrics

For quantitative assessments, we adhere to the methodology outlined in [1,2] and categorize layout metrics into composition-relevant and graphic metrics based on explicit and implicit aesthetic principles of advertising posters. Graphic metrics associated with explicit aesthetic principles encompass R_{ove}, R_{und}, and R_{ali}, measuring layout overlap, underlay overlap, and layout alignment degree, respectively. Composition-relevant metrics linked to implicit aesthetic principles include R_{com}, R_{shm}, and R_{sub}, which quantify background complexity, occlusion subject degree, and occlusion product degree, respectively. Additionally, we employ the metric R_{occ} to denote the ratio of non-empty layouts predicted by models. All the aforementioned metrics will be used to assess different models, validating the effectiveness of our method. The formal definitions of these metrics are presented in [1,2,10,12]. In addition to the aforementioned general quantitative metrics, a user study based on subjective evaluations was conducted to compare various methods.

4.3. Comparison with state-of-the-art methods

Image-aware layout generation methods. We begin by conducting experiments to compare DD-GAN with ContentGAN [4], CGL-GAN [1], PDA-GAN [2], and IUC-Layout [26], all capable of generating image-aware layouts. Quantitative results are presented in Table 1. Our model has demonstrated robust performance across most content-relevant and graphic metrics, showcasing the effectiveness of DD-GAN with MLM in addressing both implicit and explicit aesthetic principles. For example, compared to the SOTA method, our model achieves superior performance on all evaluation metrics except R_{com}. Similarly, compared to the other two other image-aware layout generation methods, ContentGAN and CGL-GAN, our model outperforms them on all

Fig. 4. Qualitative evaluation for image-aware layout generation methods. Layouts in a column are conditioned with the same image. And those in a row are from the same model. This figure qualitatively compares and analyzes different models from explicit and implicit aesthetic principles.

metrics except R_{ali}. Similarly, in terms of graphical metrics, DD-GAN achieves the best performance across all graphical metrics except R_{ali}. Notably, it is also the only model to reach a score of 100.0 on the metric R_{occ}.

To provide a comprehensive evaluation beyond general quantitative metrics, a user study was conducted, as detailed in Table 2. We randomly selected 80 test samples. Each sample includes one product image and four corresponding predicted layouts (by CGL-GAN, PDA-GAN, IUC-Layout, and DD-GAN). Participants were divided into two groups (10 professional designers and 30 novice designers) and asked to select the eligible and best layouts from the four predictions. The eligible-selected (P_e) and best-selected (P_b) layout percentages, calculated as the ratio of this model's vote count to the total votes across all methods. The results demonstrate that our model outperforms other methods, particularly in terms of the significantly higher proportion of eligible layouts compared to the others.

In the left part of Fig. 4, our model effectively prevents overlapping of text boxes, contrasting with models ContentGAN, CGL-GAN, and PDA-GAN. Specifically, when generating an underlay element, our model generates a corresponding text box at the respective position to complement it. These experimental cases illustrate the model's proficiency in learning explicit aesthetic principles of graphic layouts. Analyzing the rightmost two columns in Fig. 4, text boxes generated by our model are often positioned in areas with simple backgrounds, enhancing text readability. Moreover, in the first two columns on the right side of Fig. 4, it is evident that when the entire background is complex, the model simultaneously generates an underlay box to replace the intricate background, ensuring readability of text information. Combining the right side of Fig. 4 with Fig. 5, the layouts generated by our model effectively avoid subject regions of the products, facilitating a comprehensive presentation of product information. These qualitative analyses of these cases provide compelling evidence that our model effectively learns implicit aesthetic principles of image-aware layouts.

Table 7
Different models with multi-focus label matching. *CGL* and *PDA* represent CGL-GAN [1] and PDA-GAN [2], respectively. ✓ (or ×) indicates the model with (or without) multi-focus label matching.

Model	*MLM*	$R_{com}\downarrow$	$R_{shm}\downarrow$	$R_{sub}\downarrow$	$R_{ove}\downarrow$	$R_{und}\uparrow$	$R_{ali}\downarrow$	$R_{occ}\uparrow$
CGL	×	34.11	**15.41**	0.783	0.0413	**0.9400**	0.0098	99.7
CGL	✓	**32.53**	15.66	**0.776**	**0.0211**	0.9306	**0.0081**	**99.9**
PDA	×	**32.07**	13.56	0.727	0.0353	0.9205	**0.0109**	99.7
PDA	✓	32.95	**12.88**	**0.628**	**0.0125**	**0.9517**	**0.0109**	**99.9**
Ours	×	32.62	13.51	0.756	0.0300	0.9323	0.0116	99.8
Ours	✓	**32.54**	**13.44**	**0.705**	**0.0099**	**0.9458**	**0.0101**	**100.0**

Image-agnostic layout generation methods. We also compare our model with image-agnostic methods, namely LayoutTransformer [22] and LayoutVTN [15]. As depicted in Table 3, these image-agnostic methods perform well on graphic metrics, focusing solely on explicit aesthetic principles while neglecting implicit aesthetic principles. Consequently, regarding content-relevant metrics, our model significantly outperforms them. Specifically, DD-GAN surpasses LayoutTransformer by 20.5%, 36.2%, and 46.2% in terms of R_{com}, R_{com}, and R_{com}, respectively. Similarly, DD-GAN exceeds LayoutVTN by 22.1%, 39.5%, and 46.7% with respect to R_{com}, R_{com}, and R_{com}, respectively. Compared to VTN, DD-GAN improves the score of R_{occ} from 99.9 to 100.0, generating eligible layouts for all test images.

4.4. Ablations

4.4.1. Ablations for different discriminators

Ablation experiments evaluating DD-GAN with various discriminator configurations are detailed in Table 4. When DD-GAN is configured without any discriminator, as indicated in the first row of Table 4, it exhibits the lowest performance across multiple metrics, including

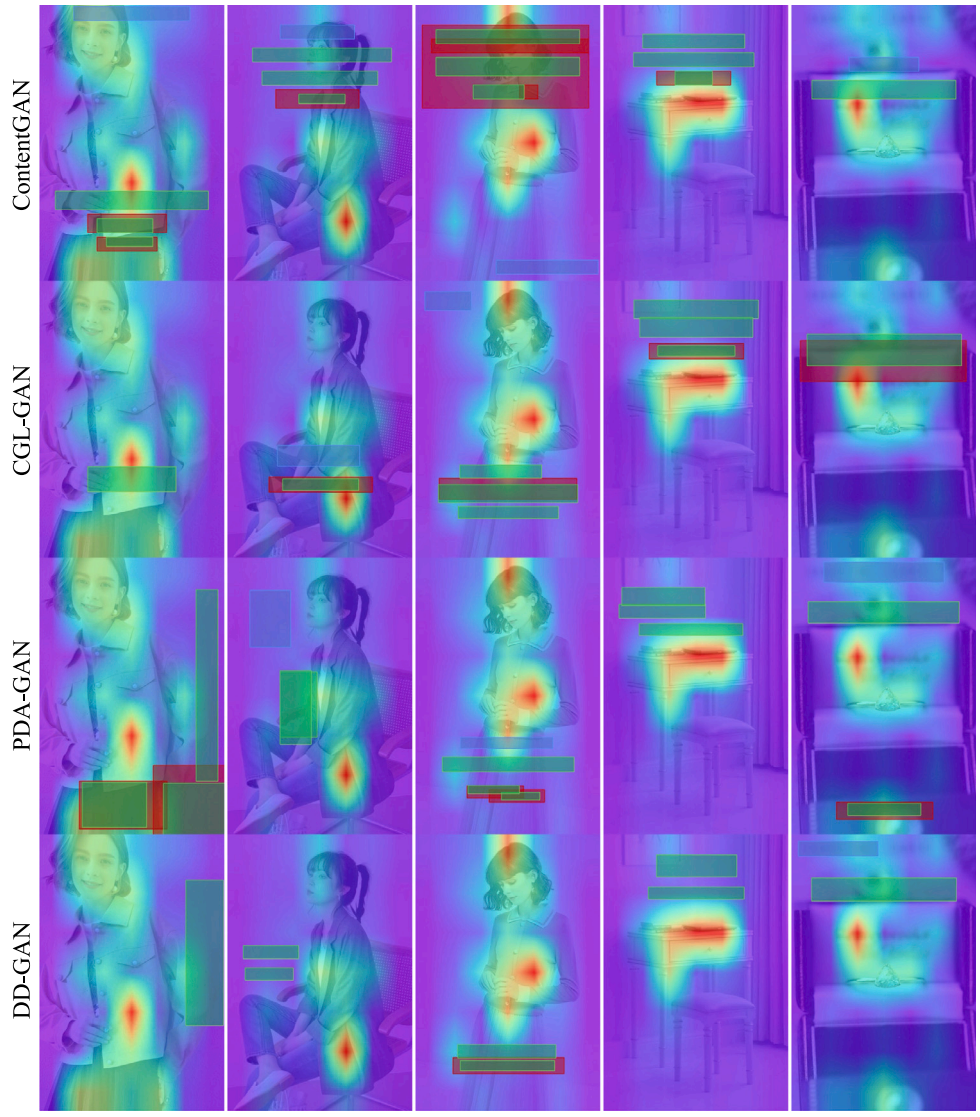

Fig. 5. Qualitative evaluation for different models in product attention maps. The color transition from blue to red in heatmaps reflects the variation of product attention values from low to high [1,53,54]. A well-designed layout can strategically avoid areas with high heat values, ensuring it does not impact the display of products.

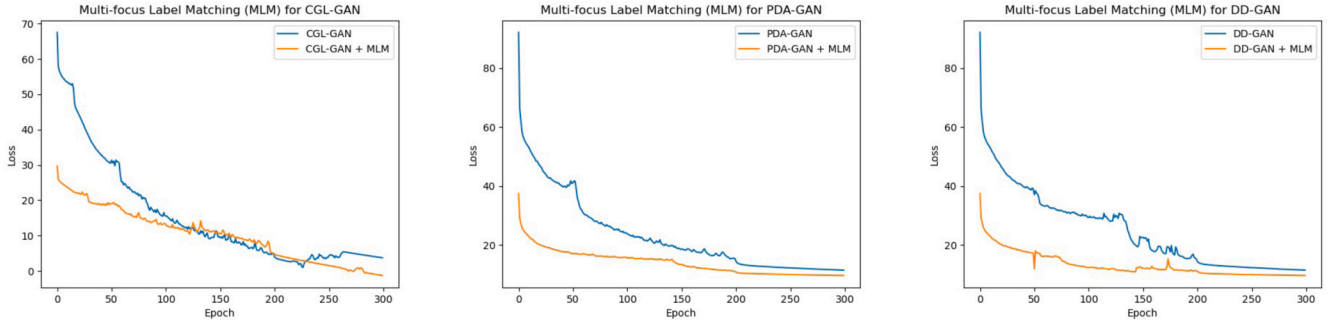

Fig. 6. Training loss curves for different models with (without) the incorporation of multi-focus label matching.

R_{com}, R_{sub}, R_{ove}, and R_{occ}. Although DD-GAN with only the PD achieves the best results on metrics R_{shm}, R_{sub}, and R_{und}, it performs slightly worse than DD-GAN with both GD and PD on the other four metrics. Overall, as shown in Table 4, DD-GAN equipped with both discriminators consistently achieves first or second-best performance on most metrics. This improvement arises from the complementary effects of the two discriminators: PD addresses domain discrepancy issues, while GD enhances the coordination between the generated layout and the input image.

4.4.2. Ablations for different configurations

PDA-GAN demonstrated strong performance in content-related metrics. Based on this, we adopted a similar setup, with the weight ratio between the reconstruction loss and the pixel-level discriminator loss

Table 8
Comparison with multi-iteration training. The first six rows present the test results of PDA-GAN [2] at intervals of 300 epochs, ranging up to 1800 epochs during training. The last row shows the test results of the PDA model, which incorporates our designed MLM and undergoes training for 300 epochs.

MLM	$Epoch$	$R_{com}\downarrow$	$R_{shm}\downarrow$	$R_{sub}\downarrow$	$R_{ove}\downarrow$	$R_{und}\uparrow$	$R_{ali}\downarrow$	$R_{occ}\uparrow$
×	300	**32.07**	13.56	0.727	0.0353	0.9205	**0.0109**	99.7
×	600	32.15	13.33	0.705	0.0373	0.9113	0.0114	99.4
×	900	32.43	13.44	0.704	0.0377	0.9134	0.0115	99.4
×	1200	32.44	13.44	0.703	0.0377	0.9139	0.0115	99.4
×	1500	32.44	13.43	0.703	0.0377	0.9139	0.0115	99.4
×	1800	32.44	13.43	0.703	0.0377	0.9139	0.0115	99.4
✓	300	32.95	**12.88**	**0.628**	**0.0125**	**0.9517**	**0.0109**	**99.9**

Table 9
MLM with different repeated times. The first row presents the test results of PDA-GAN without MLM. The second to fifth rows represent the introduction of MLM into PDA-GAN with repetition 2 to 5 times, respectively.

$times$	$R_{com}\downarrow$	$R_{shm}\downarrow$	$R_{sub}\downarrow$	$R_{ove}\downarrow$	$R_{und}\uparrow$	$R_{ali}\downarrow$	$R_{occ}\uparrow$
1	32.07	13.56	0.727	0.0353	0.9205	0.0109	99.7
2	**31.41**	12.93	0.698	0.0109	0.9306	0.0097	98.4
3	32.95	12.88	0.628	0.0125	0.9517	0.0109	**99.9**
4	32.22	11.58	0.594	0.0099	**0.9690**	**0.0076**	98.8
5	31.67	**9.768**	**0.523**	**0.0095**	0.9228	0.0091	99.3

Table 10
Different configurations of L_{rec} and L_{rec}^{M}. Bold and underlined numbers denote the best and second best respectively. The first row presents the test results of PDA-GAN without L_{rec}^{M}. The second row presents the test results of PDA-GAN without L_{rec}. The third row represents the model with only L_{rec}^{M} in the first 150 epoch and only L_{rec} in the second 150 epoch. The fourth row represents the model with both L_{rec} and L_{rec}^{M} throughout the training process.

L_{rec}	L_{rec}^{M}	$R_{com}\downarrow$	$R_{shm}\downarrow$	$R_{sub}\downarrow$	$R_{ove}\downarrow$	$R_{und}\uparrow$	$R_{ali}\downarrow$	$R_{occ}\uparrow$
✓	×	**32.07**	13.56	0.727	0.0353	0.9205	0.0109	99.7
×	✓	34.91	19.48	0.927	0.0166	0.9232	**0.0055**	**100.0**
–	–	35.20	20.00	0.975	**0.0070**	0.8961	0.0101	99.9
✓	✓	32.95	**12.88**	**0.628**	0.0125	**0.9517**	0.0109	99.9

set to 1:6. Other training details, such as the number of epochs and learning rates, are provided in Section 4.1. With this configuration, we performed ablation studies on the weight parameters in Eq. (3), including the weight ratio between the two discriminators and the two reconstruction losses, to determine the optimal settings for DD-GAN.

Different weight configurations for L_{PD}^{G} and L_{GD}^{G} As mentioned earlier, based on the weight ratio of 1:1:6 for L_{rec}, L_{rec}^{M}, and L_{PD}^{G}, we conducted ablation experiments on different weight configurations for L_{GD}^{G}, as shown in Table 5. It can be observed that when both β and γ are set to 6, the model achieves optimal results across most metrics. However, in this case, the R_{occ} metric, which measures the ratio of non-empty layouts predicted by the model, shows that three images fail to generate any layout elements. When the weight of γ is set to 8 or higher, the model successfully generates graphical elements for all test images, resulting in an R_{occ} value of 100.0. Therefore, we set β to 6 and γ to 8.

Different weight configurations for L_{rec} and L_{rec}^{M} We also conducted ablation experiments with different weight configurations for L_{rec} and L_{rec}^{M}, as shown in Table 6. It can be seen that the model's overall performance is not significantly affected by different weight configurations. This stability benefits from our model, which considers both the explicit aesthetic of the graphical layout geometry and the implicit aesthetic of the coordination between the image and the graphical layout. Since the model achieves stable performance across all metrics with an R_{occ} value of 100.0 when the weight ratio of L_{rec} to L_{rec}^{M} is 1:1, we ultimately set both weight coefficients to 1.

4.4.3. Ablations for MLM

To fully demonstrate the effectiveness and seamless integration of the proposed MLM, we incorporated MLM into the CGL-GAN, PDA-GAN, and DD-GAN in the first ablation study, comparing the performance changes before and after the integration. In subsequent ablation studies, we further demonstrated the straightforward integration of MLM by using PDA-GAN as the base model. We conducted comparative experiments involving multi-iteration training, ablation studies with varying repetition times, and configuration ablation experiments related to L_{rec} and L_{rec}^{M}.

Different models with MLM. Table 7 presents the performance results of integrating MLM into various models for comparative evaluation. In comparison to the original PDA-GAN [2] model, the inclusion of MLM resulted in improvements across all measurement metrics except for R_{com}. Particularly within our model, the introduction of MLM led to improvements in all measurement metrics, affirming that MLM imparts stronger supervisory signals. These supervisory signals enable the model to gain a more profound understanding of both explicit and implicit aesthetic principles in image-aware graphic layouts. As shown in Fig. 6, the model incorporating MLM demonstrates rapid convergence in the initial training phase. Although the CGL-GAN with MLM experiences loss fluctuations during training, the final loss value stabilizes at a low level after training. Furthermore, for PDA-GAN and DD-GAN with MLM, the loss values remain consistently low throughout the training process. From the two subfigures on the right side of Fig. 6, the model without MLM exhibits oscillations during training. In contrast, the model with MLM demonstrates greater stability during training, benefiting from the richer reconstruction information provided by MLM for supervision. This clearly illustrates the convenience and effectiveness of integrating MLM into other models, providing more effective supervisory signals.

Comparison of MLM and Multi-Iteration Training. To validate the ease of integrating MLM into other models and its provision of more effective supervisory signals, we incorporated MLM into PDA-GAN [2] and compared it with the original PDA-GAN through multiple training iterations, as depicted in Table 8. As the training iterations increased, the original PDA-GAN exhibited marginal improvements in R_{com}, R_{shm}, and R_{sub}. In contrast, the PDA-GAN that incorporates MLM achieved optimal results across all measurement metrics after only training 300 epochs, except R_{com}. Particularly noteworthy is the significant improvement in R_{shm}, R_{sub}, R_{ove}, and R_{und}, where the PDA-GAN with MLM, trained for 300 epochs, outperformed the PDA-GAN without MLM trained for 1800 epochs by 4.1%, 10.7%, 66.8%, and 4.1%, respectively. This clearly demonstrates that MLM provides more comprehensive supervisory signals for the model compared to multi-iteration training, enabling the model to better learn both explicit and implicit aesthetic principles of image-aware graphic layouts.

MLM with different repeated times. Table 9 demonstrates that the model achieves enhanced performance with the inclusion of MLM under various repeated times. Moreover, a comparison with Table 8 in the paper reveals that incorporating MLM into the model outperforms multi-iteration training. The primary reason is that multi-iteration training does not guarantee effective supervision for each query, whereas MLM ensures that every query receives consistently effective supervisory signals.

Additionally, as shown in Table 9, the model's performance varies with different repetition times, exhibiting a nonlinear relationship. This indicates that performance does not simply improve or decline monotonically with the number of repetitions. For example, when the repeated time exceeds 3, excessive supervised learning can lead to overfitting, resulting in decreased test set performance of R_{occ}. Conversely, when the repeated time is less than 3, insufficient supervision leads to underfitting, also decreasing R_{occ}. Considering the performance in terms of the ratio of non-empty layouts (R_{occ}), we ultimately choose to replicate three times as the final setting for the model.

Configurations of L_{rec} and L_{rec}^{M}. In our ablation study, we also explored the effects of supervising our model using only L_{rec}, only L_{rec}^{M}, or both L_{rec} and L_{rec}^{M}. Furthermore, we experimented with using L_{rec}^{M} during the initial unstable training phase and switching to L_{rec} after 150 epochs for another 150 epochs. As shown in Table 10, when the model was supervised only by L_{rec}, it performed well on

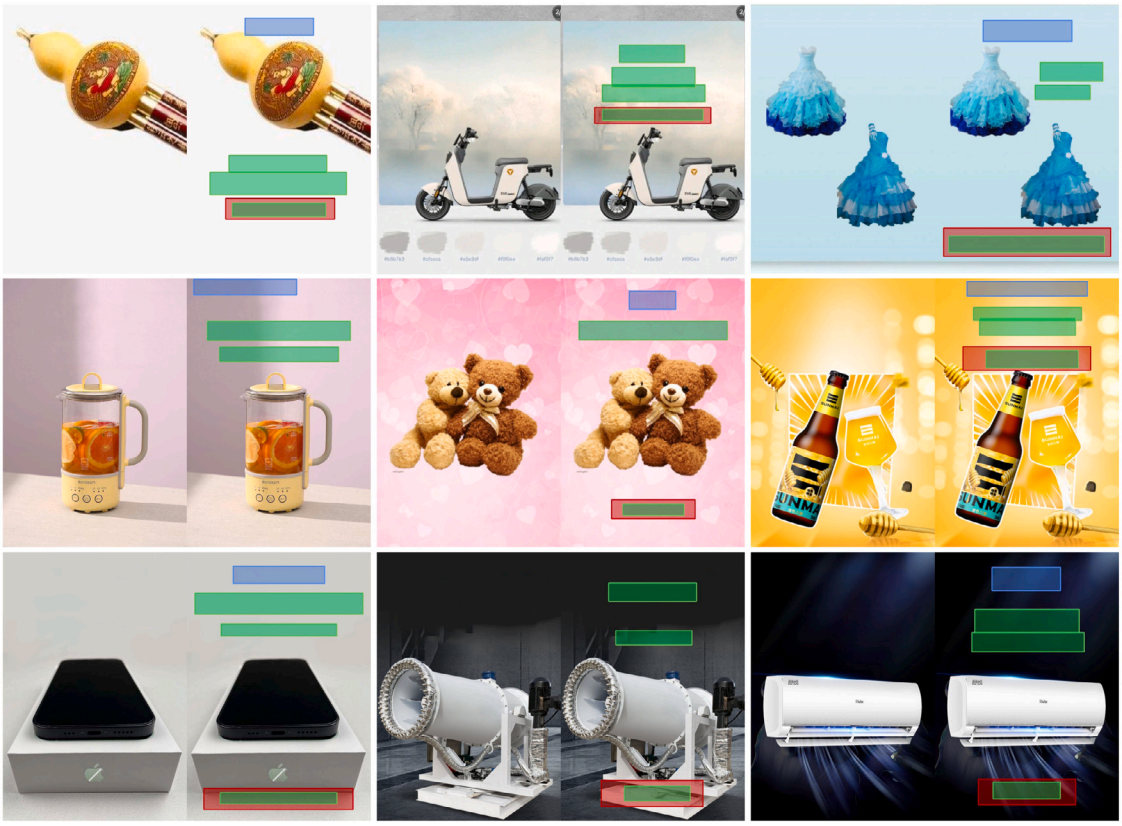

Fig. 7. The demonstration of DD-GAN's performance when transferred to other datasets.

content-relevant metrics but poorly on graphic metrics. Conversely, when supervised only by L_{rec}^{M}, the model showed improvement in graphic metrics but a significant decline in content-relevant metrics. When both L_{rec} and L_{rec}^{M} were used simultaneously for training, as indicated in the fourth row of Table 10, the model achieved the best or second best performance in all metrics, particularly excelling in metrics of R_{shm}, R_{sub}, and R_{und}. Based on these results, we ultimately decided to use both L_{rec} and L_{rec}^{M} for supervision to generate content-aware graphic layouts that balance both explicit and implicit aesthetics.

4.5. Testing DD-GAN on other datasets

In the previous experiments, we primarily focused on evaluating the performance of various models on the CGL-Dataset. Here, we assess the performance of DD-GAN on the PKU-Dataset [13], as shown in Fig. 7. From the first row of Fig. 7, it can be observed that DD-GAN effectively avoids the main subject areas in the PKU-Dataset, ensuring that both the primary content of the original image and the generated layout elements are well presented. In the second row of Fig. 7, text elements tend to be placed in smooth background regions, improving the readability of the text. The third row shows that when the background of text elements is complex, DD-GAN often generates an underlay element to replace the intricate background, thereby enhancing the readability of the textual information. In summary, these results demonstrate that DD-GAN can be effectively applied to other datasets, showcasing strong robustness and adaptability.

5. Conclusion

To address the challenges of capturing both explicit and implicit aesthetic principles in existing layout generation models, this paper introduces multi-focus label matching and proposes DD-GAN. The multi-focus label matching accelerates convergence and provides richer supervision for models during training. Both quantitative and qualitative evaluations demonstrate that DD-GAN, coupled with multi-focus label matching, can generate high-quality image-aware graphic layouts for advertising posters, which simultaneously satisfy explicit aesthetic principles. In the future, we will further explore video layout generation tasks, focusing on acquiring video datasets and ensuring smooth transitions in continuous frame layouts.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank the reviewers for their constructive comments. Weiwei Xu is partially supported by "Pioneer" and "Leading Goose" R&D Program of Zhejiang (No. 2023C01181). This paper is supported by Information Technology Center and State Key Lab of CAD&CG, Zhejiang University.

Data availability

The data that has been used is confidential.

References

[1] M. Zhou, C. Xu, Y. Ma, T. Ge, Y. Jiang, W. Xu, Composition-aware graphic layout GAN for visual-textual presentation designs, in: IJCAI, ijcai.org, 2022, pp. 4995–5001.

[2] C. Xu, M. Zhou, T. Ge, Y. Jiang, W. Xu, Unsupervised domain adaption with pixel-level discriminator for image-aware layout generation, in: CVPR, IEEE, 2023, pp. 10114–10123.

[3] S. Cheng, D. Sheng, J. Yao, Z. Shen, Poster graphic design with your eyes: An approach to automatic textual layout design based on visual perception, Displays 79 (2023) 102458.

[4] X. Zheng, X. Qiao, Y. Cao, R.W.H. Lau, Content-aware generative modeling of graphic design layouts, ACM Trans. Graph. 38 (4) (2019) 133:1–133:15.

[5] H. Lee, L. Jiang, I. Essa, P.B. Le, H. Gong, M. Yang, W. Yang, Neural design network: Graphic layout generation with constraints, in: ECCV (3), in: Lecture Notes in Computer Science, vol. 12348, Springer, 2020, pp. 491–506.

[6] R. Kumar, J.O. Talton, S. Ahmad, S.R. Klemmer, Bricolage: example-based retargeting for web design, in: CHI, ACM, 2011, pp. 2197–2206.

[7] X. Pang, Y. Cao, R.W.H. Lau, A.B. Chan, Directing user attention via visual flow on web designs, ACM Trans. Graph. 35 (6) (2016) 240:1–240:11.

[8] S. Xiao, Y. Chen, Y. Song, L. Chen, L. Sun, Y. Zhen, Y. Chang, T. Zhou, UI semantic component group detection: Grouping UI elements with similar semantics in mobile graphical user interface, Displays 83 (2024) 102679.

[9] I.J. Goodfellow, NIPS 2016 tutorial: Generative adversarial networks, 2017, CoRR abs/1701.00160.

[10] J. Li, J. Yang, A. Hertzmann, J. Zhang, T. Xu, LayoutGAN: Synthesizing graphic layouts with vector-wireframe adversarial networks, IEEE Trans. Pattern Anal. Mach. Intell. 43 (7) (2021) 2388–2399.

[11] S. Shahriar, GAN computers generate arts? A survey on visual arts, music, and literary text generation using generative adversarial network, Displays 73 (2022) 102237.

[12] J. Li, J. Yang, J. Zhang, C. Liu, C. Wang, T. Xu, Attribute-conditioned layout GAN for automatic graphic design, IEEE Trans. Vis. Comput. Graph. 27 (10) (2021) 4039–4048.

[13] H. Hsu, X. He, Y. Peng, H. Kong, Q. Zhang, PosterLayout: A new benchmark and approach for content-aware visual-textual presentation layout, in: CVPR, IEEE, 2023, pp. 6018–6026.

[14] A.A. Jyothi, T. Durand, J. He, L. Sigal, G. Mori, LayoutVAE: Stochastic scene layout generation from a label set, in: ICCV, IEEE, 2019, pp. 9894–9903.

[15] D.M. Arroyo, J. Postels, F. Tombari, Variational transformer networks for layout generation, in: CVPR, Computer Vision Foundation / IEEE, 2021, pp. 13642–13652.

[16] Z. Jiang, J. Guo, S. Sun, H. Deng, Z. Wu, V. Mijovic, Z.J. Yang, J. Lou, D. Zhang, LayoutFormer++: Conditional graphic layout generation via constraint serialization and decoding space restriction, in: CVPR, IEEE, 2023, pp. 18403–18412.

[17] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, V. Lempitsky, Resolution-robust large mask inpainting with Fourier convolutions, in: WACV, IEEE, 2022, pp. 3172–3182.

[18] W. Quan, J. Chen, Y. Liu, D.-M. Yan, P. Wonka, Deep learning-based image and video inpainting: A survey, Int. J. Comput. Vis. 132 (7) (2024) 2367–2400.

[19] A. Farahani, S. Voghoei, K. Rasheed, H.R. Arabnia, A brief review of domain adaptation, 2020, CoRR abs/2010.03978.

[20] J. Li, Z. Yu, Z. Du, L. Zhu, H.T. Shen, A comprehensive survey on source-free domain adaptation, IEEE Trans. Pattern Anal. Mach. Intell. (2024).

[21] H.W. Kuhn, The hungarian method for the assignment problem, Nav. Res. Logist. Q. 2 (1–2) (1955) 83–97.

[22] K. Gupta, J. Lazarow, A. Achille, L. Davis, V. Mahadevan, A. Shrivastava, LayoutTransformer: Layout generation and completion with self-attention, in: ICCV, IEEE, 2021, pp. 984–994.

[23] J. Zhang, J. Guo, S. Sun, J.-G. Lou, D. Zhang, LayoutDiffusion: Improving graphic layout generation by discrete diffusion probabilistic models, 2023, arXiv preprint arXiv:2303.11589.

[24] S. Chai, L. Zhuang, F. Yan, Z. Zhou, Two-stage content-aware layout generation for poster designs, in: ACM Multimedia, ACM, 2023, pp. 8415–8423.

[25] D. Horita, N. Inoue, K. Kikuchi, K. Yamaguchi, K. Aizawa, Retrieval-augmented layout transformer for content-aware layout generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 67–76.

[26] C. Xu, K. Han, W. Xu, Image-aware layout generation with user constraints for poster design, Vis. Comput. (2024) 1–14.

[27] D.C.L. Ngo, L.S. Teo, J.G. Byrne, Formalising guidelines for the design of screen layouts, Displays 21 (1) (2000) 3–15.

[28] C.E. Jacobs, W. Li, E. Schrier, D. Bargeron, D. Salesin, Adaptive grid-based document layout, ACM Trans. Graph. 22 (3) (2003) 838–847.

[29] Y. Cao, A.B. Chan, R.W.H. Lau, Automatic stylistic manga layout, ACM Trans. Graph. 31 (6) (2012) 141:1–141:10.

[30] P. O'Donovan, A. Agarwala, A. Hertzmann, Learning layouts for single-PageGraphic designs, IEEE Trans. Vis. Comput. Graph. 20 (8) (2014) 1200–1213.

[31] S. Guo, Z. Jin, F. Sun, J. Li, Z. Li, Y. Shi, N. Cao, Vinci: An intelligent graphic design system for generating advertising posters, in: CHI, ACM, 2021, pp. 577:1–577:17.

[32] C. Yang, W. Fan, F. Yang, Y.F. Wang, LayoutTransformer: Scene layout generation with conceptual and spatial diversity, in: CVPR, Computer Vision Foundation / IEEE, 2021, pp. 3732–3741.

[33] K. Kikuchi, E. Simo-Serra, M. Otani, K. Yamaguchi, Constrained graphic layout generation via latent optimization, in: ACM Multimedia, ACM, 2021, pp. 88–96.

[34] X. Yang, F. Hu, L. Ye, Z. Chang, J. Li, A system of configurable 3D indoor scene synthesis via semantic relation learning, Displays 74 (2022) 102168.

[35] D. Majumdar, V.P. Namboodiri, Unsupervised domain adaptation of deep object detectors, in: ESANN, 2018.

[36] Y. Zhang, B.D. Davison, Domain adaptation for object recognition using subspace sampling demons, Multimed. Tools Appl. 80 (15) (2021) 23255–23274.

[37] J. Zhang, J. Huang, Z. Tian, S. Lu, Spectral unsupervised domain adaptation for visual recognition, in: CVPR, IEEE, 2022, pp. 9819–9830.

[38] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V.S. Lempitsky, Domain-adversarial training of neural networks, in: Domain Adaptation in Computer Vision Applications, in: Advances in Computer Vision and Pattern Recognition, Springer, 2017, pp. 189–209.

[39] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, D. Krishnan, Unsupervised pixel-level domain adaptation with generative adversarial networks, in: CVPR, IEEE Computer Society, 2017, pp. 95–104.

[40] Z. Pei, Z. Cao, M. Long, J. Wang, Multi-adversarial domain adaptation, in: AAAI, AAAI Press, 2018, pp. 3934–3941.

[41] C. Ren, Y.H. Liu, X. Zhang, K. Huang, Multi-source unsupervised domain adaptation via pseudo target domain, IEEE Trans. Imag. Process. 31 (2022) 2122–2135.

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: NIPS, 2017, pp. 5998–6008.

[43] D. Jia, Y. Yuan, H. He, X. Wu, H. Yu, W. Lin, L. Sun, C. Zhang, H. Hu, DETRs with hybrid matching, in: CVPR, IEEE, 2023, pp. 19702–19712.

[44] Q. Chen, X. Chen, G. Zeng, J. Wang, Group DETR: fast training convergence with decoupled one-to-many label assignment, 2022, CoRR abs/2207.13085.

[45] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: ECCV (1), in: Lecture Notes in Computer Science, vol. 12346, Springer, 2020, pp. 213–229.

[46] J. Hosang, R. Benenson, B. Schiele, Learning non-maximum suppression, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4507–4515.

[47] P. Isola, J. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: CVPR, IEEE Computer Society, 2017, pp. 5967–5976.

[48] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: CVPR, Computer Vision Foundation / IEEE, 2019, pp. 4401–4410.

[49] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, IEEE Computer Society, 2016, pp. 770–778.

[50] T. Lin, P. Dollár, R.B. Girshick, K. He, B. Hariharan, S.J. Belongie, Feature pyramid networks for object detection, in: CVPR, IEEE Computer Society, 2017, pp. 936–944.

[51] B. Wang, Q. Chen, M. Zhou, Z. Zhang, X. Jin, K. Gai, Progressive feature polishing network for salient object detection, in: AAAI, AAAI Press, 2020, pp. 12128–12135.

[52] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: CVPR, IEEE Computer Society, 2016, pp. 2818–2826.

[53] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: ICML, in: Proceedings of Machine Learning Research, vol. 139, PMLR, 2021, pp. 8748–8763.

[54] H. Chefer, S. Gur, L. Wolf, Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers, in: ICCV, IEEE, 2021, pp. 387–396.

[55] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: ICLR (Poster), 2015.