TEXT2 海报：在检索到的图像上排布风格化文本

Chuhao Jin[1,2] Hongteng Xu[1,2⋆] Ruihua Song[1,2⋆] Zhiwu Lu[1,2]

[1]高瓴人工智能学院，中国人民大学，北京 100872，中国[2]北京大数据管理与分析方法重点实验室{jinchuhao, hongtengxu,rsong}@ruc.edu.cn

arXiv:2301.02363v1 [cs.MM] 6 Jan 2023

摘要

海报生成是一项重要任务，适用于广泛的应用场景，但通常耗时较长，且需要大量手动编辑和艺术经验。本文提出了一种新颖的数据驱动框架，名为 *Text2Poster*，用于根据文本信息自动生成视觉效果出色的海报。仿照人工编辑海报的流程，我们的框架利用大规模预训练视觉—文本模型，根据给定文本检索背景图像，通过级联自编码器在图像上迭代排布文本，最后采用基于匹配的方法为文本添加风格。我们通过弱监督和自监督学习策略训练框架中的各个模块，从而降低对标注数据的需求。客观实验和主观实验均表明，在生成海报的质量方面，我们的 Text2Poster 优于最先进的方法，包括学术研究成果和商业软件。

索引词—— 海报生成、图文检索、布局预测、弱监督与自监督学习

1. 引言

海报兼具艺术性与功能性，广泛应用于许多商业和非商业场景，用于宣传和传播信息。例如，电子商务平台会使用吸引人的海报来推广商品。会议等社交活动的网站通常会以精美且信息丰富的海报装饰。这些高质量海报通过将风格化文本嵌入合适的背景图像中生成，而这一过程依赖大量人工编辑和难以量化的艺术经验。然而，这种耗时且主观的流程无法满足现实应用中对精心设计海报日益增长的巨大需求，从而降低了信息传播效率，并导致推广效果欠佳。

在这项工作中，我们提出了一种新颖的数据驱动框架 *Text2Poster*，它能够实现性能优异的自动海报生成器。如图 1 所示，Text2Poster 首先利用大规模预训练视觉文本模型，根据输入文本检索合适的背景图像。随后，该框架通过从估计的布局分布中采样来初始化文本布局，并借助级联自动编码器迭代优化布局。最后，它从带有语义标签的字体和颜色集合中检索文本的颜色和字体。我们采用弱监督和自监督学习策略来训练该框架的各个模块。实验表明，我们的 Text2Poster 框架能够自动生成高质量的海报，在客观和主观评估方面均显著优于学术界和工业界的竞争者。

2. 相关工作与挑战 ES

如前所述，海报生成需要（i）根据给定的文本信息确定背景图像，以及（ii）优化图像上的文本布局。在第一步中，大多数现有方法要么未利用高层特征 [1] ，要么只考虑单一方面的特征 [2],，这往往会导致不理想的检索结果。近年来，大规模预训练视觉文本模型，如 Oscar [3], 、CLIP [4], 和我们的 BriVL [5], ，在根据文本检索图像方面取得了令人鼓舞的表现，为海报设计寻找背景图像提供了一种潜在的解决方案。值得注意的是，尽管一些方法尝试根据文本生成图像 [6, 7], ，但其生成图像的质量并不适用于海报生成。

对于布局预测，传统的基于规则的方法 [8, 9] 从数量有限的预定义布局模板中探索文本布局，其灵活性值得商榷。对于基于学习的方法， [10] 中的神经设计网络需要背景图像感兴趣区域（ROI）等辅助信息来生成文本布局。 [11] 中的方法按顺序为每个文本元素进行布局，却忽略了后续元素对前面元素布局的影响。 LayoutGAN [12], LayoutGAN++ [13] 和 SMARTTEXT [14] 是为杂志或广告设计而开发的。它们在排布文本时未充分考虑背景图像上的视觉对象，因此可能导致海报生成的结果并非最优。

⋆Hongteng Xu 和 Ruihua Song 为通讯作者。我们的代码可在 https://github.com/chuhaojin/ Text2Poster-ICASSP-22 获取。

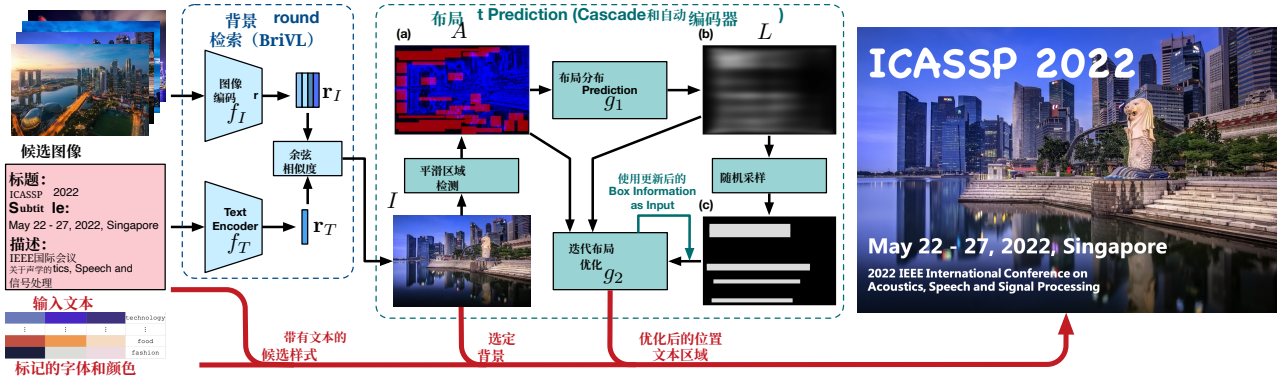


图1。我们所提 Text2Poster 框架示意图。**请注意，该海报由我们的方法生成。**

3. 所提方法

将输入文本记为 $\mathcal{T}=\{(T_i,\ \alpha_i)\}_{i=1}^{N_T}$，其中 T_i 表示第 i 个文本信息（$e.g.$，一个句子或短语），α_i 是 T_i（$e.g.$，“标题”或“副标题”）的属性。我们通过以下三个步骤，根据 $\mathcal{T}$ 生成海报：(i) 从一组候选图像中检索背景图像 I，该图像集合记为 $\mathcal{I}=\{I_j\}_{j=1}^{N_I}$。($ii$) 估计每个 T_i 的布局，记为 $\boldsymbol{P}=[\boldsymbol{p}_i]\in\mathbb{R}^{2\times N_T}$，其中 $\boldsymbol{p}_i\in\mathbb{R}^2$ 表示 T_i 在背景图像 I 上的归一化左上角坐标。(iii) 确定每个 T_i 的字体和颜色。

3.1. 通过预训练视觉-文本模型进行图像检索

在为海报生成检索背景图像时，我们希望探索与文本“弱相关”的图像。例如，当通过短语“Bob 和 Alice 的婚礼”检索图像时，我们更倾向于找到一些包含爱情隐喻的图像，$e.g.$例如一张蓝天下白色教堂的图片。为实现这一目标，我们利用在 [5], 中提出的 BriVL——一种 SOTA 预训练视觉-文本模型——从文本中检索背景图像。如图1所示，BriVL 由图像编码器f_T 和文本编码器 f_I组成。 f_T 采用 RoBERTa-Large [15] 的编码器作为其文本主干网络。 f_I 采用预训练的 Faster R-CNN [16] 和 EfficientNet [17] 作为其视觉主干网络。给定上述主干模型的输出，BriVL 堆叠四个 Transformer [18] 层，以提取2048维的视觉和文本特征。BriVL 在来自互联网的3000万对弱相关图文对上进行训练，因此适用于我们的任务；它采用 InfoNCE损失 [19]来对齐文本特征与图像特征。有关其更多实现细节，请参阅 [5]。

我们从 unsplash. com 收集了 284,781 张高质量图像，构建图像检索库。基于 BriVL，我们提取输入文本和候选图像的潜在编码，$i.e.$, $\boldsymbol{r}_{\mathcal{T}}:=f_T(\cup_i T_i)$, and $\{\boldsymbol{r}_{I_j}:=f_I(I_j)\}_{j=1}^{N_I}$。据此，我们计算 $\boldsymbol{r}_{\mathcal{T}}$ 与每个 $\boldsymbol{r}_{I_j}$ 之间的余弦相似度，并检索相似度最高的背景图像 similarity, $i.e.$, $I=\arg\max_{I_j\in\mathcal{I}}\frac{\boldsymbol{r}_{\mathcal{T}}^{\top}\boldsymbol{r}_{I_j}}{\|\boldsymbol{r}_{\mathcal{T}}\|_2\|\boldsymbol{r}_{I_j}\|_2}$.

3.2. 通过级联自动编码 排序器

给定所选图像 I，我们通过级联自动编码器预测输入文本 $\mathcal{T}$、$i.e.$的布局，即 $\boldsymbol{P}=[\boldsymbol{p}_i]$。

平滑区域检测： 受 Faster R-CNN [16], 启发，我们首先在 I中生成大小不同且相互重叠的 N_A 个区域，记为 $\{\mathcal{A}_i\}_{i=1}^{N_A}$。应用光谱残差方法 [20], ，我们生成背景图像 I的显著性图，记为 S。对于每个区域 $\mathcal{A}_i$，我们通过计算该区域内显著性图的平均值，并结合尺寸敏感的偏移量，为其赋予一个值：$i.e.$、 $v_i=\frac{1}{|\mathcal{A}_i|}(\lambda+\sum_{\boldsymbol{p}\in\mathcal{A}_i}S(\boldsymbol{p}))$，其中 $S(\boldsymbol{p})$表示 $\boldsymbol{p}$ 处的显著性， $|\mathcal{A}_i|$ 表示 $\mathcal{A}_i$中的像素数量。原则上，当 $\mathcal{A}_i$ 为平滑区域时，$\frac{1}{|\mathcal{A}_i|}\sum_{\boldsymbol{p}\in\mathcal{A}_i}S(\boldsymbol{p})$较小；对于尺寸较大的$\mathcal{A}_i$，偏移量 $\frac{\lambda}{|\mathcal{A}_i|}$ 也较小。因此，我们设定阈值 $v_{\max}$，并选择值较小的大区域$i.e.$、 $\{\mathcal{A}_i|v_i<v_{\max},\forall i=1,...,N_A\}$。对于每幅图像，我们将 $v_{\max}$ 自适应地设为 $1.4\times$ 均值$\{v_i\}$，并应用非极大值抑制（NMS）方法 [16]，以确保所选区域互不重叠。所选区域生成一个二值图，表示 I的平滑区域，记为A。图1中的图像（a）以蓝色显示显著性图 S，以红色显示平滑区域图 A。

布局分布预测： 给定映射 A，我们使用生成器 g_1 来预测布局分布，记为 L。对于每个像素 $\boldsymbol{p}$，提出 $L(\boldsymbol{p})\in[0,1]$，用于表示 $\boldsymbol{p}$ 属于文本框的概率。这里，g_1 采用自编码架构，其编码器 f_1 由多个卷积神经网络堆叠而成，解码器 h_1 由多个转置卷积神经网络堆叠而成。遵循文献 [21], 中的方法，我们将编码器的输出与可学习的位置嵌入图（记为 E）拼接，以此构造解码器的输入。因此，我们有 $L=g_1(A)=h_1(\text{Concat}(f_1(A)\ E))$。图 1 中的图像 (b) 展示了布局分布 L。

迭代布局优化： 我们将 L 视为目标布局的先验，并通过以下方式初始化布局 $\boldsymbol{P}^{(0)}=[\boldsymbol{p}_i^{(0)}]$

从 L中采样每个 $\boldsymbol{p}_i^{(0)}$。将未归一化的 $\boldsymbol{p}_i^{(0)}$ 作为第 i 个文本框的左上角坐标，我们据此初始化文本框，其大小由文本 T_i 的长度及其属性 α_i决定。图 1 中的图像 (c) 展示了这些文本框。

我们利用一个记为 g_2 的自动编码器，以自回归的方式优化布局信息，其中

$$\boldsymbol{P}^{(k+1)} = g_2(\text{Concat}(A, L), \boldsymbol{P}^{(k)}),\ k = 0, ..., K-1 \quad (1)$$

K 为迭代次数。对于 g_2，其编码器由堆叠的卷积神经网络构成，解码器则是一个 2 层双向长短期记忆网络。

用于布局分布预测的自动编码器与用于布局优化的自动编码器，共同构成了一个采用级联自动编码架构的布局预测器，该架构模拟了手动图像编辑的过程。实际上，海报设计师在排布文本时总会避开醒目且信息丰富的区域。他们通常先粗略确定文本的位置，然后反复调整。从这个角度来看，我们的布局预测器在某种程度上实现了相同的过程。

学习级联自动编码器：我们分别训练这两个自动解码器。具体而言，我们从 huaban.com 收集了 154,013 张海报图像，涵盖 16 个类别。huaban.com 是一个鼓励用户收藏精美图片的网站。对于数据集中的每张海报，我们首先使用光学字符识别工具 [22] 检测其中的文本 $\hat{\mathcal{T}}$ 及相应的文本框 $\boldsymbol{P} = [\hat{\boldsymbol{p}}]$。这些文本框生成一张二值布局图，记为 L。利用布局图遮挡图像，并采用 [23], 中的图像修复方法填充遮罩区域，我们就能获得海报的背景图像 I。应用平滑区域检测器后，即可相应地得到平滑区域图 A。因此，我们将数据集表示为 $\mathcal{D} = \{(\hat{\mathcal{T}}_n, \hat{\boldsymbol{P}}_n, \hat{L}_n, \hat{A}_n)\}_{n=1}^{N}$。相应地，我们分别训练这两个自动编码器，方法是

$$\min_{g_1} \frac{1}{N} \sum_{n=1}^{N} \|g_1(\hat{A}_n) - \hat{L}_n\|_2^2. \quad (2)$$

$$\min_{g_2} \frac{1}{N} \sum_{n=1}^{N} \|g_2(\text{Concat}(\hat{A}_n, \hat{L}_n), \boldsymbol{P}_n^{(0)}) - \hat{\boldsymbol{P}}_n\|_F^2 \quad (3)$$

其中，$|\boldsymbol{P}_n|$ 表示第 n 个海报的框数量。这里，我们采用自监督学习策略来训练g_2，通过 $\boldsymbol{p}_{i,n}^{(0)} \sim \text{Uniform}(\hat{\boldsymbol{p}}_{i,n} - \Delta, \hat{\boldsymbol{p}}_{i,n} + \Delta)$ 对初始位置 $\boldsymbol{P}_n^{(0)} = [\boldsymbol{p}_{i,n}^{(0)}]$ 进行采样，$\forall \hat{\boldsymbol{p}}_{i,n} \in \boldsymbol{P}_n$。扰动 $\Delta = [0.1, 0.1]^T$ 控制初始位置与目标位置之间的方差。$\boldsymbol{P}_n^{(0)}$根据真实标注 $\boldsymbol{P}_n$进行采样，从而实现自监督机制。

实现细节： g_1 中使用的每个卷积层包含 16 个大小为 9×9 的卷积核。g_1 的编码器最终输出一个 64 维特征向量。对于编码器

其中 g_2的每个卷积层包含 64 个大小为 5×5 的卷积核。对于g_2的 2 层双向长短期记忆网络（解码器），其隐藏层维度设为 200。训练这两个自动编码器时，我们将数据集 $\mathcal{D}$ 划分为 138,013 张训练海报和 16,000 张验证海报。每张海报都调整为 300×400。我们使用 Adam 算法 [24] 优化模型，学习率为 0.05



图 2。通过不同文本查询检索到的图像。

批量大小为 512。在四块 V100 GPU 上，我们分别训练 g_1 和g_2，耗时 4 小时和 48 小时。

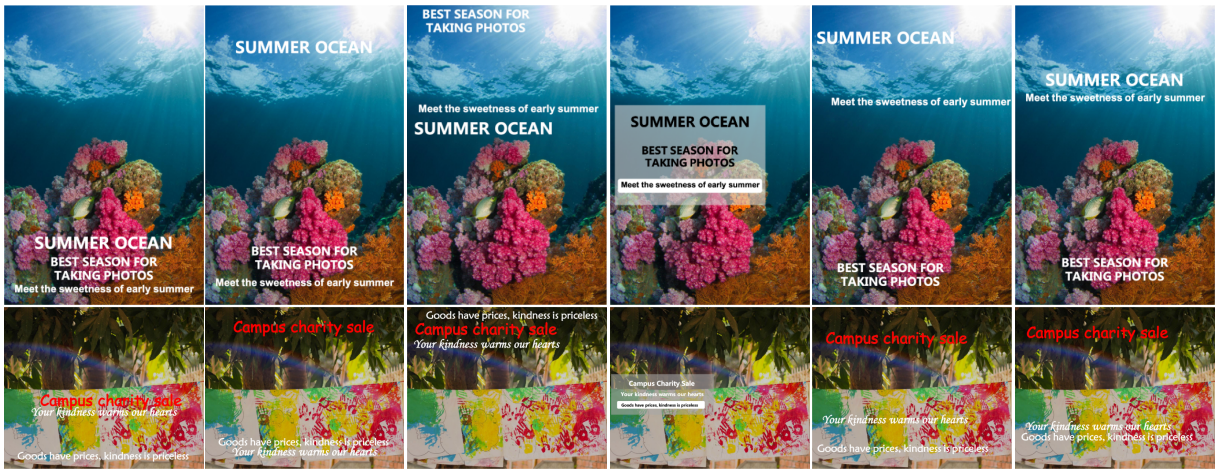
3.3. 文本风格化

给定数据集 $\mathcal{D}$，我们构建一个元组集合，记作 $\{(\hat{\boldsymbol{r}}_{i,n}, c_{i,n}^I, c_{i,n}^T, z_{i,n})\}$，其中 $\hat{\boldsymbol{r}}_{i,n} = f_T(\hat{T}_{i,n})$ 是 $\hat{T}_n$ 中第 i 个文本的特征，$c_{i,n}^I$ 是其背景颜色，$c_{i,n}^T$ 是其颜色，$z_{i,n}$ 是其字体。给定这些元组，我们根据文本的词频解析出 53 个语义标签，并应用 K-means 等传统聚类方法，找出与这些标签对应的一组典型文本风格，记作 $\mathcal{F} = \{(\boldsymbol{r}_m, c_m^I, c_m^T, z_m)\}_{m=1}^{53}$。因此，对于每个输入文本 $T \in \mathcal{T}$，我们提取其特征 $\boldsymbol{r} = f_T(T)$，并获取其背景颜色 $c^I = I(\boldsymbol{p})$，其中 $\boldsymbol{p}$ 是 T 的预测布局。基于 $(\boldsymbol{r}, c^T)$，我们通过余弦相似度从 $\mathcal{F}$ 中找出匹配的聚类中心，并据此确定 T 的颜色和字体。

4. 实验

背景图像检索：除了基于 BriVL 的图像检索方法外，我们还考虑（a）采用 unsplash.com 的搜索引擎，以及（b）将输入文本与图像检索库中图像的标签进行匹配。图 2 展示了不同方法获得的代表性检索结果。我们的方法检索到的图像确实包含与输入文本相对应的隐喻。给定文本“校园慈善义卖”，基线方法往往会找到包含“义卖”和“校园”等明确概念的图像，而我们的方法则找到了一张包含茁壮成长的树木、彩虹和彩色手印的图像，其内容体现了隐含但恰当的语义。即使面对“ 一起看世界”和“梦想永不止步”这类具有挑战性的抽象描述，我们的方法仍能找到合适的图像。

在主观评估中，给定 50 个文本查询，我们使用不同方法为每个查询检索排名前 5 的图像。我们邀请了三位志愿者对检索图像的质量进行评分，分数从 0（非常差）到 4（非常好）。Unsplash 搜索引擎、基于标签的匹配方法和我们基于 BriVL 的方法的平均分和标准差分别为 2.17 ± 0.10 1.64 ± 0.16 和 $\mathbf{2.38} \pm 0.13$，这进一步证明了我们方法的优越性。



LayoutGAN++ IUI DeSal LUBAN Ours(K=1) Ours(K=30)

图 3。各种布局预测方法生成的海报。

布局预测： 我们从定量和定性两方面评估布局预测器，并将其与以下基线方法进行比较：(*i*) SOTA 基于学习的方法 *Lay-outGAN++* [13]；(*ii*) SOTA 基于规则的方法*IUI* [8] 和 *DeSal* [25]；(*iii*) 商用海报生成器*LUBAN*，网址为 https://luban.aliyun.com。为展示迭代布局优化策略的实用性，我们分别将布局预测器的 K 设为 1、5 和 30。

我们从 huaban.com 收集了 16,000 张海报，构建了一个参考数据集，并准备了三个背景图像集：Unsplash2K、Unsplash10K 和 PSD1.6K。Unsplash2K 和 Unsplash10K 分别包含来自 unsplash.com 的 2,000 张和 10,000 张背景图像。PSD1.6K 包含从 PSD 格式的海报文件中提取的 1,637 张背景图像。对于每个图像集，我们使用各种方法在背景图像上排版输入文本并生成海报。我们遵循 [13] 的工作，计算生成的海报与参考数据集之间的 Fr´echet Inception Distance（FID）。FID 广泛用于基于 GAN 的图像生成任务，是一种评估指标，它使用 Inception-v3 衡量两个数据集之间的分布距离。表1中的结果表明，我们的方法始终优于基线方法，并且随着 K 的增加，其性能也有所提升，这验证了我们迭代细化策略的合理性。

此外，我们人工选取了50组文本，每组包含一个标题以及若干副标题或描述。对于每种布局方法，我们首先使用 BriVL 模型为每组文本检索五张背景图像，并据此生成250张海报。我们邀请13名志愿者对生成的海报进行评分，评分范围为0（很差）至4（很好），以评估布局的美学效果。表1展示了每种方法的平均分和标准差。在主观实验中，我们的方法同样取得了最佳结果。图3展示了一些生成的海报，进一步验证了我们的 Text2Poster 方法的有效性。

方法	FID			美学得分
	PSD1.6K	Unsplash2K	Unsplash10K	
LayoutGAN++	55.46	75.63	58.11	$1.63_{\pm 0.22}$
IUI	44.39	68.85	53.24	$2.25_{\pm 0.24}$
DeSal	43.47	70.30	55.25	$2.28_{\pm 0.32}$
我们的方法（$K=1$）	41.83	71.64	55.20	$2.14_{\pm 0.25}$
我们的方法（$K=5$）	39.28	68.77	52.78	$2.36_{\pm 0.21}$
我们的方法（$K=30$）	**39.09**	**67.94**	**52.75**	**$2.39_{\pm 0.22}$**

表1。各种版面预测方法的客观和主观评估。*LUBAN*仅提供收费服务，因此无法进行大规模数值测试。

5. 结论

本文提出了一种新颖的框架，可自动根据输入文本生成海报，并取得了最先进的性能。我们利用预训练的视觉文本模型进行图像检索，并使用级联自编码器进行布局预测。该框架的各个模块通过前沿的弱监督和自监督学习策略进行训练。未来，我们计划通过微调预训练的 BriVL 模型并采用端到端学习策略来改进该框架。

6. 致谢

本研究得到北京优秀青年科学家计划（项目编号：BJJWZYJH012019100020098）、北京人工智能研究院大规模预训练计划468、北京大数据管理与分析方法重点实验室，以及中国人民大学“双一流”建设重大创新与规划跨学科平台智能社会治理平台的支持。我们还要感谢中国人民大学公共政策与决策研究实验室提供的支持。

7. 参考文献

[1] Wei Huang、Yan Gao 和 Kap Luk Chan，《基于区域的图像检索综述》，*Journal ofSignal Processing Systems*，第59卷第2期，第143–161页，2010年。

[2] Tian-qiang Peng 和 Fang Li，《基于深度卷积神经网络和二进制哈希学习的图像检索》，发表于*ICASSP*，2017年，页码 1742–1746。

[3] Xiujun Li、Xi Yin、Chunyuan Li、Pengchuan Zhang、Xiaowei Hu、Lei Zhang、Lijuan Wang、Houdong Hu 等，《Oscar：面向视觉语言任务的对象语义对齐预训练》，发表于*ECCV*，2020年，页码 121–137。

[4] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sas- try、Amanda Askell、Pamela Mishkin 等，《从自然语言监督中学习可迁移视觉模型》，*arXiv* 预印本 *arXiv:2103.00020*，2021年。[5] Yuqi Huo、Manli Zhang、Guangzhen Liu、Haoyu Lu、Yizhao Gao、Guoxing Yang、Jingyuan Wen、Heng Zhang、Baogui Xu 等，《Wenlan：通过大规模多模态预训练连接视觉与语言》，*arXiv* 预印本 *arXiv:2103.06561*，2021年。[6] Xiaofeng Mao、Yuefeng Chen、Yuhong Li、Tao Xiong 等，《基于语言的图像编辑中的双线性表示：采用条件生成对抗网络》，发表于*ICASSP*，2019年，页码 2047–2051。[7] Aditya Ramesh、Mikhail Pavlov、Gabriel Goh、Scott Gray、Chelsea Voss、Alec Radford、Mark Chen 和 Ilya Sutskever，《零样本文本到图像生成》，*arXiv* 预印本 *arXiv:2102.12092*，2021年。[8] Ali Jahanian、Jerry Liu、Qian Lin、Daniel Tretter、Eamonn O’Brien-Strain、Seungyon Claire Lee、Nic Lyons 和 Jan Allebach，《用于自动设计杂志封面的推荐系统》，发表于*IUI*，2013年，页码 95–106。[9] Xuyong Yang、Tao Mei、Ying-Qing Xu、Yong Rui 和 Shipeng Li，《视觉文本演示布局的自动生成》，*TOMM*，第 12 卷，第 2 期，页码 1–22，2016年。

[10] Hsin-Ying Lee、Lu Jiang、Irfan Essa、Phuong B Le、Haifeng Gong、Ming-Hsuan Yang 和 Weilong Yang，《神经设计网络：带约束的图形布局生成》，载于*ECCV*，2020年，第491–506页。

[11] Jianan Li、Jimei Yang、Jianming Zhang、Chang Liu、Christina Wang 和 Tingfa Xu，《用于自动图形设计的属性条件布局生成对抗网络》，*TVCG*，2020年。

[12]Jianan Li、Jimei Yang、Aaron Hertzmann、Jianming Zhang 和 Tingfa Xu，《LayoutGAN：利用线框判别器生成图形布局》，发表于*ICLR*，2018年。

[13] Kotaro Kikuchi、Edgar Simo-Serra、Mayu Otani 和 Kota Yamaguchi，《通过潜变量优化生成受约束的图形布局》，*arXiv* 预印本 *arXiv:2108.00871*，2021年。[14]Peiying Zhang、Chenhui Li 和 Changbo Wang，《SMARTTEXT：学习在自然图像上生成和谐的文本布局》，发表于*ICME*，2020年，第1–6页。[15] Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Mandar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov，《RoBERTa：一种鲁棒优化的 BERT 预训练方法》，*arXiv* 预印本 *arXiv:1907.11692*，2019年。[16] Shaoqing Ren、Kaiming He、Ross Girshick 和 Jian Sun，《Faster R-CNN：借助区域提议网络实现实时目标检测》，发表于*NeurIPS*，2015年，第91–99页。[17] Mingxing Tan 和 Quoc Le，《EfficientNet：重新思考卷积神经网络的模型缩放》，发表于*ICML*，2019年，第6105–6114页。[18] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin，《Attention is All You Need》，发表于*NeurIPS*，2017年，第5998–6008页。[19] Aaron van den Oord、Yazhe Li 和 Oriol Vinyals，《利用对比预测编码进行表征学习》，*arXiv* 预印本 *arXiv:1807.03748*，2018年。[20] Xiaodi Hou 和 Liqing Zhang，《显著性检测：一种谱残差方法》，发表于*CVPR*，2007年，第1–8页。[21] Gen Li、Nan Duan、Yuejian Fang、Ming Gong 和 Daxin Jiang，《UniCoder-VL：通过跨模态预训练实现视觉与语言的通用编码器》，发表于*AAAI*，2020年，第34卷，第11336–11344页。[22] Yuning Du、Chenxia Li、Ruoyu Guo、Xiaoting Yin、Weiwei Liu、Jun Zhou、Yifan Bai、Zilin Yu、Yehua Yang 等，《PP-OCR：一种实用的超轻量级光学字符识别系统》，*arXiv* 预印本 *arXiv:2009.09941*，2020年。[23] Alexandru Telea，《一种基于快速行进法的图像修复技术》，《图形工具期刊》，第9卷，第1期，第23–34页，2004年。[24] Diederik P Kingma 和 Jimmy Ba，《Adam：一种随机优化方法》，发表于*ICLR*，2015年。[25] Zoya Bylinskii、Nam Wook Kim、Peter O’Donovan、Sami Alsheikh、Spandan Madan、Hanspeter Pfister、Fredo Durand、Bryan Russell 和 Aaron Hertzmann，《学习图形设计和数据可视化中的视觉重要性》，发表于*ACM UIST*，2017年，第57–69页。