ISSN: 2661-3719(Print); 2661-3727(Online)

基于对抗生成网络的超分辨率海报生成模型

利锦轩　邓秀勤
广东工业大学　数学与统计学院　广东广州　510006

摘　要：随着计算机技术的不断进步，图像生成技术也在飞速发展。本文提出了一种以中文文本为输入的海报生成方法，将StackGan模型的第二阶段改进为超分辨率模型（GAN_SR3）。我们引入CLIP模型将中文文本映射到高维编码，提升文本与图像的匹配度。通过在StackGan的第一阶段输入编码和随机噪声，生成低分辨率图像，然后输入至改进的超分辨率模型SR3，获得高分辨率海报。在自建数据集上与其他模型进行比较，实验结果显示GAN_SR3在电影海报和网络海报数据集上获得更高的Inception Score和更小的FID。

关键词：文本图像生成；对抗生成网络；深度学习；图像生成

引言：

随着互联网技术的飞速发展，信息传播的速度也呈现出日益加快的趋势。在这种背景下，海报作为一种生动、直观、吸引人的宣传工具，广泛应用于商业、文化等各个领域。传统的海报设计通常需要设计师在电脑上的特定应用程序中使用电子画板进行创作，这需要大量的时间和精力。同时，随着对个性化服务的需求不断增加，对于需要大量或频繁制作海报的设计任务，手工制作显然不够高效。因此，实现多样化的设计风格和提高海报制作效率，已经成为海报设计领域亟待解决的难题。

另一方面，传统的海报设计往往依赖于设计师的经验和直觉，而基于对抗生成模型的海报生成模型则依托数学规律和计算机技术，不受传统设计师的个人经验和限制。在一定程度上，这种模型可以应对较为复杂的设计任务。此外，通过利用人工智能等技术来制作海报，我们可以借鉴大量的海报数据，运用深度学习等技术来不断优化算法，从而实现更高效的海报生成过程。这种方法为海报设计领域带来了更多可能性，并且有望提高制作海报的效率和多样性。在计算机技术不断进步的背景下，图像生成模型已受到了广泛的关注和研究，并已应用于若干领域。生成对抗网络（Generative Adversarial Networks，简称GANs）[1]是当下备受关注的图像生成技术之一，GANs融合生成器和鉴别器以实现各种相关任务，例如LIU等人就提出了一种基于Gans的图像恢复算法[2]，该算法能够有效的修复图像中的缺陷和噪声。而KINGMA等人提出的变分自编码器（Variational self-encoder，简称VAE）[3]代表了一种基于神经网络的图像生成模型，其核心特点在于从低维空间中采样生成图像。在国际领域，KARRAS等人提出的StyleGAN2模型[4]已经成功实现了高质量人脸图像的生成。为了实现文本信息与图像信息的对齐，RADFORD等人提出了CLIP模型[5]，随后RAMESH等人在此基础上提出的DALL-E2模型[6]能够生成各种形状和场景的图像，这两个模型在图像生成领域取得了显著的成就。

文本生成图像（text to image）是一个跨领域的综合性问题，需要计算机既能理解文本内容，又能生成相应的图像。该问题可分为两个子问题：首先，对文本信息进行编码以生成计算机可识别的内容，并去除不相关的文本信息；其次，生成模型需要基于处理后的文本向量生成与文本信息匹配的图像样本。因此，文本生成图像涉及两个领域：自然语言处理和计算机视觉。目前，大多数关于文本生成图像的研究论文都是基于生成对抗网络（GANs）[1]和扩散模型（Denoising Diffusion Probabilistic Models，简称DDPM）[7]来实现的。例如，2016年REED等人[8]提出了一种基于GANs文本描述生成图像的方法，该方法可以从文本描述中提取有效的语义信息，生成更符合文本内容的图像。随后，该团队提出了生成对抗网络（Generative Adversarial What-Where Network，简称GAWWN）[9]，这是一种使用文本信息、位置和姿势作为条件输入的架构，丰富了图像生成的细节，但使训练过程更加复杂。ZHANG等人提出了堆叠生成对抗网络（Stacked Generative Adversarial Network，简称StackGAN）架构[10]，该架构使用两个生成对抗网络

基金项目：国家自然科学基金资助项目（12101136），2021年广东省研究生教育创新计划项目（2021SFKC030）
作者简介：利锦轩(1997-)，男，硕士研究生，主要研究方向为机器学习，数据挖掘。
通信作者：邓秀勤（1966-），女，本科，教授，主要研究方向为数据挖掘，机器学习。

实现文本到图像的生成，将任务分为两个阶段，从而提高了生成样本的质量。随后ZHANG等人进一步提出了StackGAN++模型[11]，该模型设计了生成式对抗网络，以树状结构进行并行训练，有效地减少了模型的训练时间。2018年，XU等人提出了生成式对抗网络的注意机制（AttnGAN）[12]，在Stack-GAN++中增加了一种注意机制，不仅使用文本内容作为条件输入的全局约束，还通过注意机制提取文本中的词汇向量作为网络架构输入的局部约束，从而生成与文本内容更匹配的图像。ARJOVSKY等人提出WGAN模型[13]缓解梯度消失的同时，提升生成图像的多样性。此外，RAMESH等人在DALL-E模型[6]运用dVAE+CLIP+Transformer的技术，实现文本特征与图像特征对齐，使得生成的图片更加贴合文本提示。随后，DALL-E2模型[6]将CLIP和DDPM结合，进一步提高生成的大尺寸图片的质量。

根据KANG等人的研究[14]显示，相比于同等量级别的DDPM模型，GANs推理速度更快，可以合成更高分辨率的图像，以及支持各种潜空间编辑。因此，受StackGan++模型[11]和DALL-E2[6]模型的启发，本文提出GAN_SR3模型。该模型使用CLIP模型[5]对文本信息进行编码，采用GANs模型结构对输入的文本进行解析，生成满足要求的小尺寸的图像。接着，使用由三个DDPM组成的模型对图像的清晰度进一步完善。实验结果表明，在相同的文本信息下，GAN_SR3相比其他模型可以生成更加稳定，更加清晰，更加契合文本信息的图片。

一、相关工作

1.生成对抗网络

Goodfellow等人在2014年提出生成式对抗网络GANs[1]。GANs作为无监督学习领域的一个里程碑，在数据生成领域开辟了新的研究途径。在GANs模型中，一个神经网络充当生成器（G），将噪声作为输入，通过一系列的卷积、池化等操作生成目标图像，随后与真实图像一起交给下一个神经网络，该神经网络充当判别器（D），负责辨别输入进来的图像哪个是真实的图像，哪个是生成的图像。

在原始的生成对抗网络中，生成器和判别器都使用多层感知器作为网络框架。为了从真实数据X_{real}中学习生成模型的分布R_p，从先验随机分布中采样生成一个随机噪声向量z，然后将随机噪声向量z输入到生成器模型中生成X_{fake}。生成器模型优化过程的目标函数如（1）所示：

$$\min_G V(D,G)=E_{z-pz}[\log(1-D(G(z)))] \quad (1)$$

而判别器的任务是判断生成器生成的数据的真实性，并输出一个表示该数据为真实的概率值。通常，鉴别器经过训练，成为一个二值分类器，接收两种类型的数据：真实数据和生成器生成的伪造数据。对于每个输入数据，鉴别器产生一个介于0到1之间的概率值，表明该输入数据是真实数据的可能性。当鉴别器的输出错误时，其权重会被更新以提高其准确性。通过不断优化权重，鉴别器逐渐学习如何更好地区分真实数据和伪造数据，从而协助生成器生成更逼真的虚拟数据。这一过程是生成对抗网络中的核心机制之一，用于推动生成器不断提升生成数据的真实度。

优化后的判别模型的目标函数（2）式所示：

$$\max_D V(D,G)=E_{x-pdata(x)}[logD(x)]+E_{z-p(z)}[log[(1-D(G(z)))] \quad (2)$$

2.去噪扩散概率模型（Denoising Diffusion Probabilistic Models）

2020年，Ho等研究人员提出了Denoising Diffusion Probabilistic Models（DDPM）模型[7]，以解决生成网络GANs在训练过程中的不稳定性和无法一次性生成高分辨率图像的问题。DDPM分为前向扩散过程和后向逆扩散过程，其中前向过程是向真实图像数据中逐渐添加服从高斯分布的噪声直到数据完全变成噪声的过程。具体步骤是，首先，在真实图像分布$q(x_0)$中采样出x_0，其次在区间（1，...，T）中随机生成扩散次数t，随后在标准正态分布中采样一个随机噪声ϵ，根据（3）式得到样本x_t

$$x_t=\sqrt{\bar{\alpha}}x_0+\sqrt{1-\bar{\alpha}_t}\epsilon \quad (3)$$

其中$\sqrt{\bar{\alpha}_t}$是超参数。

后向逆扩散过程推导如下：

$$x_{t-1}=\frac{1}{\sqrt{\alpha_t}}\left(x_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t,t)\right) \quad (4)$$

拟合噪声ε_θ，使得图片经过t次去噪声后，得到原始图像。

二、海报生成模型

在借鉴了StackGAN模型[10]和DALL-E2模型[6]的灵感之后，通过对StackGAN生成模型的第二阶段进行了改良，诞生了GAN_SR3模型。模型流程如图1所示：

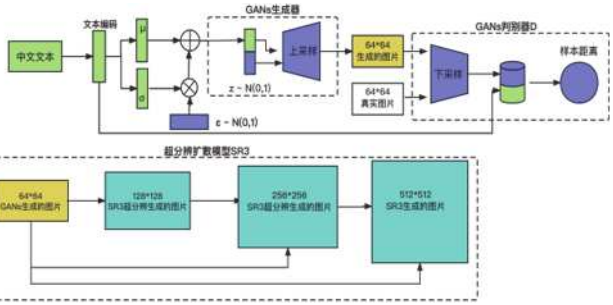

图1　GAN_SR3模型流程图

Fig.1 The flow chart of GAN_SR3 model

首先，采用预训练的Chinese-CLIP模型[15]，进行了文本嵌入，通过将提示文本输入到该模型中，获得了文本向量$T_\theta = (T_1,\ldots,T_N)$.

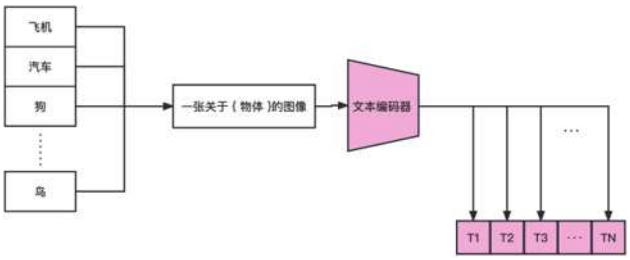

图2　CLIP模型文本编码过程

Fig.2 The text encoding process of CLIP model

随后，将随机生成的一维高斯噪声拼接在得到的文本向量末尾，并且输入给GANs模型的生成器G，并且进入第二阶段。

生成器G通过上采样（Up-sample）操作和卷积生成64*64的图像，并且与真实图像一起输入给判别器D。在判别器D中，借鉴了WGAN模型[16]的思想，将生成的模拟样本分布P_g与原始样本分布P_r相结合，将它们看作所有可能联合分布的集合。然后，计算它们之间的距离以及距离的期望值。通过模型的训练，生成器朝着期望值下界的方向进行优化，以更好地模拟所有可能联合分布。在这一过程中，使用Wasserstein距离来度量两个样本之间的距离，即：

$$|f(x_1)-f(x_2)|\le k|x_1-x_2| \quad (5)$$

其中k是函数f（x）的Lipschitz常数。

依据式子（5），两者的分布的距离公式可整理为：

$$L=k\,[D(real)-D(G(x))] \quad (6)$$

由于生成器G希望生成的结果更加接近原始样本的分布，因此生成器G的损失函数可以简写为：

$$G(loss)=-D(G(x)) \quad (7)$$

而判别器的任务是将两者区分开，因此损失函数为：

$$D(loss)=D(G(x))-D(real) \quad (8)$$

通过不断的迭代训练，当生成器G已经可以生成让判别器D无法分辨真假的图片的时候，训练进入第三阶段。

在第三阶段，采用通过重复细化实现超分辨率的SR3模型[17]。SR3模型如图3所示，将去噪扩散模型应用于条件图像生成，通过不断迭代细化，对每次迭代所需去除的噪声进行预测，最终实现从低分辨率图像到高分辨率图像的生成。

在SR3模型中对噪声的预测采用如图4所示的U-Net网络[18]进行学习，其中真实图像与双三次插值上采样得到的目标分辨率图像在通道维度上进行级联作为U-net的输入。

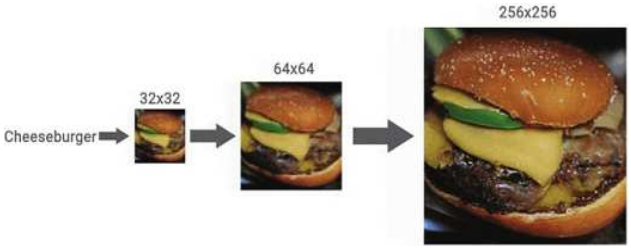

图3　SR3模型图像超分辨的过程

Fig.3 The process of SR3 model image super-resolution

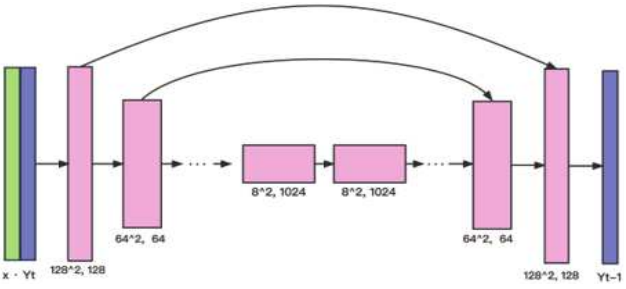

图4　U-net模型的结构图

Fig.4 Structure diagram of the U_net model

SR3模型训练步骤如算法1所示，

算法1　训练去噪模型f_θ

1）输入：输入输出图像对数据集$D=\{x_i,y_i\}_i^N$，超参数γ，其中x为低分辨率图像，y为高分辨率图像

2）从图像分布p（x，y）中采样得到图像对（x，y）

3）从标准高斯分布采样噪声ϵ

4）用梯度下降法拟合

5）$\partial_\theta \parallel f_\theta(x,\sqrt{\gamma}y_0+\sqrt{1-\gamma}\,\epsilon,\gamma)\parallel_p^p$

6）重复步骤2—步骤4，直到收敛。

迭代去噪的伪代码算法2所示

算法2　第T次迭代去噪的伪代码

1）从高斯分布中采样y_T

2）t=T

3）$\mathbf{z}\sim N(0,1)$ if t > 1,*else* $\mathbf{z}=0$

4）$y_{t-1}=\frac{1}{\sqrt{n}}(y_t-\frac{1-\alpha_t}{\sqrt{1-y_t}}f_\theta(x,y_t,\gamma_t)+\sqrt{1-\alpha_t}z$

5）t=t-1

6）直至t=1

7）输出y_0

三、实验

1.数据集

在训练阶段，选择两种不同类型的训练数据集，分别为电影海报和网络海报。这些数据集是通过网络爬虫从设计网站、社交媒体平台以及设计作品数据库中获取的。

电影海报数据集包括了共计39，515张电影海报图片。每张电影海报图片都伴随着10个不同的描述文本，这些描述文本涵盖了各种关于电影海报的信息和特点。另外，网络海报数据集则包含了共计11，788张海报图片。与电影海报类似，每张网络海报图片也附带了10个不同的描述文本，这些描述文本用于描述不同类型的海报内容和特征。

电影海报数据数量分布如图5所示：

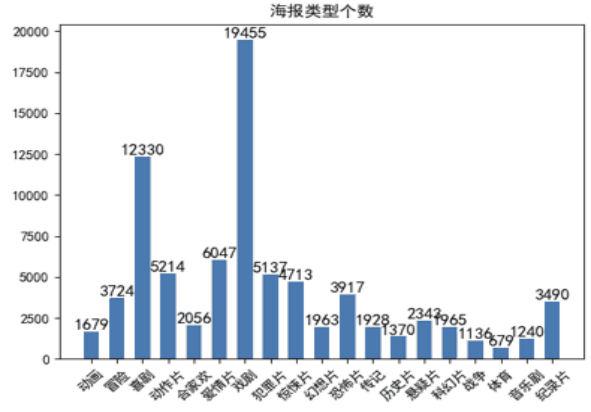

图5　电影海报数量分布

Fig.5 Distribution of the number of movie posters

在网络海报数据集中，海报的类型包括学术海报、音乐会海报、艺术展海报、体育赛事海报、演讲海报、学术会议海报、旅游推广海报、酒店推广海报、餐厅菜单海报、健身活动海报、慈善活动海报、科技展览海报、校园招聘海报、婚礼请柬海报、生日派对海报、儿童活动海报海报、时装秀海报、剧场演出海报社区活动海报、健康养生海报等。这些海报类型涵盖了不同领域和场合的需求，包括艺术文化、娱乐、商业、社区和个人活动。每种类型的海报都有自己独特的设计风格和内容要求，从而使海报生成模型能够学习更多的海报设计风格和知识，增强模型的泛化能力。

2.评价标准

对于文本生成图像模型，其不仅要生成清晰的图像，还要保持图像元素的多样性。因此，为了评估图像的质量，采用国际通用评价指标Inception Score（简称IS）和Frechet Inception Distance（简称FID）。其中IS通常作为评价多样性的指标，FID作为评价生成图像质量的指标。

IS对应的计算公式为：

$$IS(G)=exp(E_{x\sim P_g}D_{KL}(p(y|x)\|p(y))) \quad (9)$$

其中，$x\sim P_g$表示生成模型生成样本x，y表示初始v3分类模型中生成图像的输出类别。当边缘分布$p(y)$和条件分布$p(y|x)$之间的KL散度较大时，模型生成的样本具有多样性。

FID的计算公式如式（4）所示：

$$FID(P,G)=||\mu_p-\mu_g||^2+Tr(\Sigma_p+\Sigma_g-2(\Sigma_p\Sigma_g)^{\frac{1}{2}}) \quad (10)$$

其中，P是真实图像，G是生成图像。

3.实验结果分析

通过将GAN_SR3与StackGan[12]和DDPM在相同的数据集进行充分训练后，进行测试。

以同一文本“生成带有粽子元素并带有端午节字样的图像”作为输入，分别输入GAN_SR3模型，StackGan模型和DDPM模型，得到如图6所示的图像生成结果。

图6　DDPM、StackGan、GAN_SR3生成的图像

Fig.6 The images generated by DDPM, StackGan, GAN_SR3

从图6可以看出，GAN_SR3模型生成的图像相较于StackGan[11]来说内容更加丰富，物品更加贴近文本描述，并且拥有更高的分辨率。这是由于第二阶段的SR3模型通过三个扩散模型使得生成的内容与文本信息匹配得更加紧密，置信度更高，细节信息也更多，分辨率也更高。

同时，GAN_SR3生成的图案比DDPM模型生成的图案更加符合文本的要求，也能更加准确的生成所需的文字。原因在于GAN_SR3生成的图像对文本的理解来源于训练集，在训练中，文本提示中经常包含相关的文字标题与位置信息，而DDPM对于文字标题和位置的生成更加随机，不可控，导致其往往不能高效率的生成我们所需的海报图像。

此外，通过输入类似的文本提示，引导GAN_SR3生成相关图像。例如，输入文本“生成带有少年向往远方意境并带有五四青年节字样的图像”生成图7所示图片。

图7　GAN_SR3生成的关于青年节的图像

Fig.7 Image of Youth Day generated by GAN_SR3

最后，在测试数据集上使用IS和FID指标来测量模型生成图像的质量和多样性，通过推理时间来比较模型生成图片需要耗费的时间。通过在不同的数据集上生成1000张图片，并将其全部输入Inception v3分类网络，得

到如表1所示结果：

表1　不同模型实验结果

Table 1 Experimental results of different models

数据集	电影海报			网络海报		
	IS	FID	Inf.time	IS	FID	Inf.time
StackGan	3.27	51.07	12.1s	3.74	49.58	11.4s
DDPM	3.45	52.67	8.7s	4.10	44.31	8.1s
GAN_SR3	3.70	42.52	8.6s	4.05	39.97	7.9s

四、结论

图像合成技术是计算机视觉领域的一个重要研究领域，它在人们的日常生活中有巨大的发展潜力。其中，基于文本内容生成图像是一个具有广泛实用价值的热门问题，可应用于海报设计、电影制作、动画制作、书籍封面设计等多个领域。在本文中，我们提出了一种新的方法，将CLIP用于文本嵌入模型，以获得更高质量的文本特征。同时，为了提高生成图像的质量，采用了分阶段的图像生成策略。

在第一阶段，利用文本信息生成低分辨率图像。然后，在第二阶段，使用第一阶段生成的图像作为输入，通过三个DDPM逐步提高图像的分辨率。为了防止第二阶段输入图像在编码过程中丢失底层特征，每个DDPM之间引入了跳跃链接。这同时加强了多个尺度上特征图像的融合，进一步提升了生成图像的质量。

参考文献：

[1] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[2] LIU Q, CHEN M, ZHOU D. Fast haze removal from a single image[C]//2013 25th Chinese Control and Decision Conference (CCDC). IEEE, 2013: 3780-3785.

[3] Kingma D P, Welling M. Auto-Encoding Variational Bayes[J]. stat, 2014, 1050: 1

[4] KARRAS T, LAINE S, AITTALA M, et al. Analyzing and improving the image quality of stylegan[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 8110-8119.

[5] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models form natural language supervision[C]//International conference on machine learning. PMLR,2021:8748-8763

[6] RAMESH A, DHARIWAL P, NICHOL A, et al. Hierarchical Text-Conditional Image Generation with CLIP Latents[J]. arXiv e-prints, 2022: arXiv: 2204.06125.

[7] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[J]. Advances in neural information processing systems, 2020, 33: 6840-6851.

[8] REED S, AKATA Z, YAN X, et al. Generative adversarial text to image synthesis[C]//International conference on machine learning. PMLR, 2016: 1060-1069.

[9] REED S, AKATA Z, MOHAN S, et al. Learning what and where to draw[J]. Advances in neural information processing systems, 2016, 29.

[10] ZHANG H, XU T, LI H, et al. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks[C]//Proceedings of the IEEE international conference on computer vision. 2017: 5907-5915.

[11] ZHANG H, XU T, LI H, et al. Stackgan++: Realistic image synthesis with stacked generative adversarial networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(8): 1947-1962.

[12] XU T, ZHANG P, HUANG Q, et al. Attngan: Fine-grained text to image generation with attentional generative adversarial networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 1316-1324.

[13] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein generative adversarial networks[C]//International conference on machine learning. PMLR, 2017: 214-223.

[14] KANG M, ZHU J Y, ZHANG R, et al. Scaling up gans for text-to-image synthesis[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 10124-10134.

[15] YANG A, PAN J, LIN J, et al. Chinese CLIP: Contrastive Vision-Language Pretraining in Chinese[J]. arXiv e-prints, 2022: arXiv: 2211.01335.

[16] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein generative adversarial networks[C]//International conference on machine learning. PMLR, 2017: 214-223.

[17] SAHARIA C, HO J, CHAN W, et al. Image super-resolution via iterative refinement[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(4): 4713-4726.

[18] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015: 234-241.