研究与开发

基于多文本描述的图像生成方法

聂开琴，倪郑威
（浙江工商大学信息与电子工程学院，浙江 杭州 310018）

摘　要：针对单条文本描述生成的图像质量不高且存在结构错误的问题进行研究，采用多阶段生成对抗网络模型，并提出对不同文本序列进行插值操作，从多条文本描述中提取特征，以丰富给定的文本描述，使生成图像具有更多细节。为了生成与文本更为相关的图像，引入了多文本深度注意多模态相似度模型以得到注意力特征，并与上一层视觉特征联合作为下一层的输入，从而提升生成图像的真实程度和文本描述之间的语义一致性。为了能够让模型学会协调每个位置的细节，引入了自注意力机制，让生成器生成更加符合真实场景的图像。优化后的模型在CUB和MS-COCO数据集上进行验证，生成的图像不仅结构完整，语义一致性更强，视觉上的效果更加丰富多样。

关键词：文本生成图像；生成对抗网络；计算机视觉；语义一致性；自注意力
中图分类号：TP183
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2024142

Image synthesis method based on multiple text description

NIE Kaiqin, NI Zhengwei
College of Information and Electronic Engineering, Zhejiang Gongshang University, Hangzhou 310018, China

Abstract: Aiming at the challenges associates with the low quality and structural errors existed in the images generated by a single text description, a multi-stage generative adversarial network model was used to study, and it was proposed to interpolate different text sequences to enrich the given text descriptions by extracting features from multiple text descriptions and imparting greater detail to the generated images. In order to enhance the correlation between the generated images and the corresponding text, a multi-captions deep attentional multi-modal similarity model that captured attention features was introduced. These features were subsequently integrated with visual features from the preceding layer, serving as input for the subsequent layer. This integration improved the realism of the generated images and enhanced their semantic consistency with the text descriptions. In addition, a self-attention mechanism to enable the model to effectively coordinate the details at each position was incorporated, resulting in images that were more aligned with real-world scenarios. The optimized model was verified on the CUB and MS-COCO datasets, demonstrating the generation of images with intact structures, stronger semantic consistency, and richer visual diversity.

Key words: text-to-image, generative adversarial network, computer vision, semantic consistency, self-attention

收稿日期：2024−01−09；**修回日期：**2024−05−06
基金项目：浙江省自然科学基金资助项目（No.LQ22F010008）
Foundation Item: The Natural Science Foundation of Zhejiang Province (No.LQ22F010008)

0 引言

基于文本描述的图像生成是图像生成领域的一个重要研究方向，涵盖了计算机视觉和自然语言处理两方面内容，是一个多模态方向任务[1-2]。文本生成图像任务是使用描述性文本作为后续生成图像的控制参量，利用模型反复训练，最终生成与描述内容相符合、高分辨率和高真实性的图像。该任务具有相当广泛的应用前景，可在工业设计、计算机辅助设计、幼儿教育以及医疗领域发挥极大的作用。

近些年来，随着深度学习技术的快速发展，文本描述生成图像的研究取得了可观的进展。常用方法为深度循环注意力书写器（deep recurrent attention writer，DRAW）[3]、变分自动编码器（variational auto-encoder，VAE）[4]和生成对抗网络（generative adversarial network，GAN）[5]等。目前文本生成图像的方法[6-10]大多建立在GAN及其变体的基础上，主要分为两大类。一类是利用卷积神经网络模型提取出文本描述特征，再结合上采样模块直接生成目标分辨率图像，这种方法所利用的网络结构大多较简单，生成图像细节模糊、质量不佳；另一类是通过多阶段生成器生成不同分辨率的图像，再通过不同阶段判别器对各阶段生成图像进行优化，逐步提高生成图像的分辨率，最终生成高质量图像。然而，该类方法仍然存在生成图像与输入文本语义不对应的问题。

文本生成图像是一个较为复杂的过程，自然语言与视觉图像是两种不同的模态信息，在这两者之间建立联系是一个关键步骤。由于自然语言具有多种含义，表达方式也较为抽象，单条文本描述中缺乏对象的细节信息。当文本描述中的某些特征信息发生改变时，生成图像就只包含每条文本中的部分特征，利用同一幅图像的若干条文本描述进行训练可以补充更多的细节信息，使得生成图像更接近于真实图像。另一方面，多阶段GAN可用于解决生成图像分辨率低的问题，在每个生成阶段计算出单词特征与图像局部特征之间的相似度，提高生成图像与文本描述之间的语义一致性。

综上所述，为了同时获取多条文本中的丰富信息，本文对不同文本描述采用流形插值操作，丰富生成图像的细节信息；与此同时，在生成器的初始阶段引入自注意力机制，让模型自主协调生成图像中各个位置的细节，改善原模型生成不符合常态的图像；并在图像的各个生成阶段计算文本局部特征和图像全局特征之间的相似度，采用多文本深度注意多模态相似度模型提升生成图像与文本描述之间的内容一致性，从而提高生成图像的质量。

1 相关工作

最初文本生成图像任务是通过计算文本描述关键词与图像之间的相关性来确定对应文本图像，也就是跨模态监督学习任务，该方法局限性很大。直到2016年，Reed等[11]提出了集成分类器的生成对抗网络（generative adversarial network with integrated classifiers，GAN-INT-CLS）模型，生成了分辨率为64×64（单位为像素，下文不再标出）的图像，自此GAN成为文本生成图像的主流发展方向，GAN及其后续改进模型都在图像生成领域发展迅速。GAN-INT-CLS模型生成的图像缺乏对物体细节的描述，而且生成图像的分辨率很低。后来，Reed等[12]又提出了生成对抗何处何物网络（generative adversarial what and where network，GAWWN）模型，添加辅助条件可以控制图像特定目标的位置，从而进一步提高图像的精确度。

Zhang等[13]提出堆叠生成对抗网络（stacked generative adversarial network，StackGAN）模型，解决了GAN模型无法产生高清图像的问题。该

模型生成器由两个阶段组成，分别在两个阶段生成分辨率不同的图像，并且在两个生成阶段均利用判别器对图像进行优化处理。其中，第一阶段是将噪声和文本信息作为输入，捕获形状、背景和颜色等大致信息，生成分辨率为64×64的图像；第二阶段再一次输入文本描述获取更加丰富的细节，对第一阶段的输出进行细化，生成细致精确的高分辨率图像。在后来的工作中又提出了堆叠生成对抗网络加强版（stacked generative adversarial network plus plus， StackGAN++ ）模型[14]，3个阶段生成器依次得到了分辨率为64×64、128×128、256×256的图像，进一步提升了生成图像质量。同时训练多阶段生成器解决了生成图像分辨率低的问题，但依旧存在生成图像与文本描述之间关联性较低的缺陷。因此Xu等[15]提出了注意力生成对抗网络（attentional generative adversarial network，AttnGAN）模型，该模型在StackGAN++模型的基础上引入了注意力机制，利用单词级别与句子级别的文本特征向量作为输入，引入全局注意力机制和深度注意多模态相似机制，使得生成图像相应部位能够精确对应部分文本。AttnGAN模型提高了生成图像与文本描述间的语义一致性，然而在复杂文本和多文本描述中生成的图像效果一般。

2　网络模型

本文提出的模型以AttnGAN为基础框架，加上了流形插值操作，并整合了自注意力模块和多文本深度注意多模态相似度模型。网络模型结构如图1所示。

公共数据集中的每幅图像都对应多条文本描述，在给定了一条文本描述之后，随机检索出其他相容项，并采用插值操作来丰富其语义特征。多文本注意力生成对抗网络由3组生成器与判别器共同构成，分别处理分辨率为64×64、128×128、256×256的图像。多文本深度注意多模态相似度模型包含文本编码器与图像编码器2个部分，分别用于提取文本特征和视觉图像特征，此模块生成的损失也被纳入生成器损失函数中。在生成器的初始阶段引入自注意力模块，让模型学习协调生成图像中各个位置的细节，改善原模型生成异常图像。

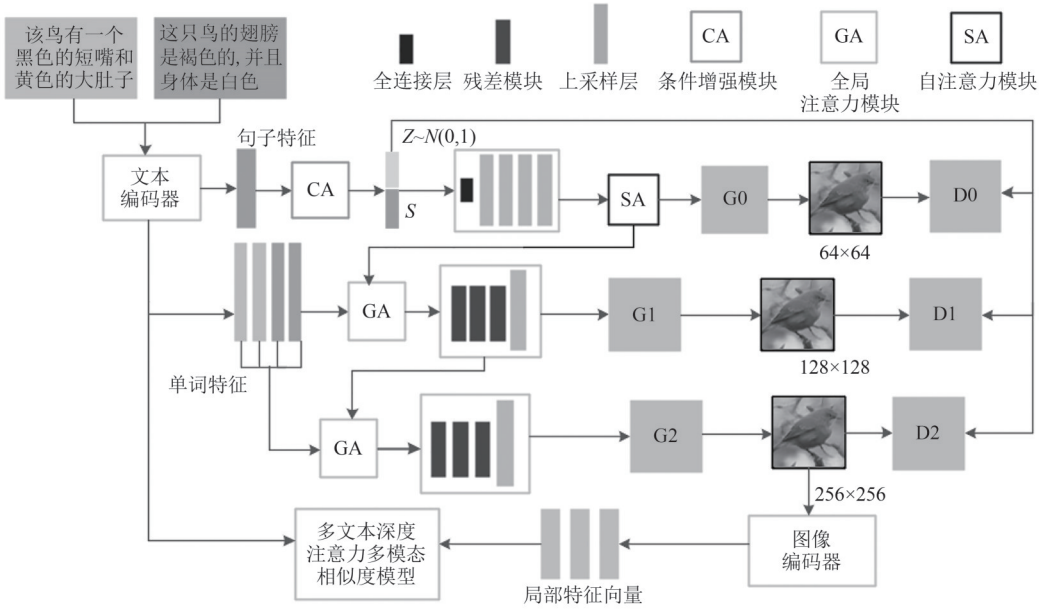


图1　网络模型结构

2.1 流形插值

根据文本描述生成图像效果不佳的其中一个重要因素就是文本描述缺乏细节信息合成[16-25]。为了提高生成图像的质量，可以对不同文本向量进行流形插值操作[26-27]，融合文本的丰富特征，使模型能够生成更加细致精确的图像。由于插值后的文本向量是合成的，判别器没有对应的图像来进行训练，但判别器可以学习预测图像和文本向量是否匹配。因此，在插值后的文本向量上满足判别器，使得生成器也可以学习在训练样本之间填补数据流形上的空白。

深度学习网络在学习文本向量之间的线性插值趋近于数据流形，即代表经过插值的文本向量含义比较接近两者分别编码得到的文本向量的中间含义。该过程可以看作在生成器目标函数上增加一项来进行最小化。

$$E_{\boldsymbol{t}_1,\boldsymbol{t}_2\sim P_{\text{data}}}\left[\log\left(1-D(G(z,(1-\beta)\boldsymbol{t}_1+\beta\boldsymbol{t}_2))\right)\right] \quad (1)$$

其中，z是从正态分布中提取的噪声，P_{data}表示真实样本数据分布，β是文本$\boldsymbol{t}_1$与$\boldsymbol{t}_2$融合的比例，D、G分别是GAN模型的判别器和生成器。$\boldsymbol{t}_1$和$\boldsymbol{t}_2$在向量空间中的位置与关系不存在任何约束，可能对应于不同图像，甚至不同类别对象的描述。

2.2 多文本注意力生成对抗网络

当前基于GAN的文本生成图像模型通常对单条文本描述进行编码生成文本向量，直接将其作为生成图像的条件信息放入模型中，该方法生成的图像缺乏单词级别的细粒度信息。由于Attn-GAN框架可利用部分单词绘制图像不同的子区域，达到图文一致的效果，因此本文在其基础上结合了自注意力机制[28-29]多文本深度注意多模态相似度模型，构建了多文本注意力生成对抗网络模型。

初始阶段，利用条件增强模块F^{ca}将句向量$\boldsymbol{e}$转化为低维度文本条件向量$\boldsymbol{c}$，具体做法是从句向量的高斯分布中得到均方差矩阵和协方差矩阵，经过计算可得到条件向量$\boldsymbol{c}$。随后将其与符合标准正态分布的噪声向量z拼接得到联合向量，经过上采样模块F_0得到隐藏向量$\boldsymbol{h}_0'$，此时$\boldsymbol{h}_0'$中已经包含目标合成对象的大致信息。再利用自注意力模块F^{SA}，使得到的最终向量$\boldsymbol{h}_0$中包含更多的空间与位置信息。第一阶段生成器G_0合成图像特征$\hat{x}_0$的生成过程为：

$$\boldsymbol{h}_0'=F_0\left(z,F^{\text{ca}}(\boldsymbol{e})\right) \quad (2)$$

$$\boldsymbol{h}_0=F^{\text{SA}}(\boldsymbol{h}_0') \quad (3)$$

$$\hat{x}_0=G_0(\boldsymbol{h}_0) \quad (4)$$

其中，$\boldsymbol{h}_0$是自注意力模块的输出结果。

首先，将$\boldsymbol{h}_0'$转换到特征空间f和g，通过1×1的卷积计算得到注意力系数矩阵$\boldsymbol{\beta}$，如式（5）、式（6）所示。

$$f(\boldsymbol{h}_0')=W_{\text{f}}(\boldsymbol{h}_0') \quad (5)$$

$$g(\boldsymbol{h}_0')=W_{\text{g}}(\boldsymbol{h}_0') \quad (6)$$

其中，W_{f}、W_{g}是感知层。

$\boldsymbol{\beta}_{j,i}$用来表示合成图像第j个区域时的第i个位置的关注程度，N为隐藏向量的特征位置数，计算过程为：

$$\boldsymbol{\beta}_{j,i}=\frac{\exp\left(\boldsymbol{\alpha}_{ij}\right)}{\sum_{i=1}^{N}\left(\boldsymbol{\alpha}_{ij}\right)} \quad (7)$$

$$\boldsymbol{\alpha}_{ij}=f(\boldsymbol{h}_0')^{\text{T}}g(\boldsymbol{h}_0') \quad (8)$$

通过自监督机制学习特征图中的空间与位置信息，为图像中重要的细节信息赋予了更大的权重值，有利于初始阶段生成更有意义的图像，接着将$\boldsymbol{h}_0'$转换到第三个特征空间u，如式（9）所示。

$$u(\boldsymbol{h}_0')=W_u(\boldsymbol{h}_0') \quad (9)$$

其中，W_u是特征空间u的感知层，用于改变特征的维度大小。随后将权重信息$\boldsymbol{\beta}_{j,i}$与$u(\boldsymbol{h}_0')$相乘得到带有注意力机制的图像特征矩阵$\boldsymbol{m}_j$。

$$\boldsymbol{m}_j=\sum_{i=1}^{N}\boldsymbol{\beta}_{j,i}u(\boldsymbol{h}_0') \tag{10}$$

最后，通过1×1的卷积计算将得到的图像特征矩阵$\boldsymbol{m}_j$转换到特征空间v，从而使得到的图像特征尺寸与输入的图像特征尺寸大小相同。

$$v\left(\boldsymbol{m}_j\right)=W_v\boldsymbol{m}_j \tag{11}$$

后续生成图像过程如式（12）、式（13）所示，其中F_i^{attn}是第i个阶段的注意力模型，用于生成图像子区域的上下文矢量。

F_i^{attn}的输入分为两部分，分别是单词特征向量$\boldsymbol{w}$和前一阶段得到的隐藏特征向量$\boldsymbol{h}_{i-1}$，经过第i阶段的上采样模块F_i后得到隐藏向量$\boldsymbol{h}_i$。

$$\boldsymbol{h}_i=F_i\left(\boldsymbol{h}_{i-1},F_i^{\text{attn}}\left(\boldsymbol{w},\boldsymbol{h}_{i-1}\right)\right) \tag{12}$$

$$\hat{x}_i=G_i\left(\boldsymbol{h}_i\right) \tag{13}$$

$\boldsymbol{h}_i$中的每一列分别对应图像的所有子区域，且每一个子区域都可以利用单词向量加权和的形式来动态表示，得到的结果称为上下文矢量，计算过程为：

$$\boldsymbol{c}_j=\sum_{i=0}^{T-1}\boldsymbol{\beta}_{j,i}\boldsymbol{w}_i' \tag{14}$$

其中，$\boldsymbol{\beta}_{j,i}$表示图像子区域中某个单词所获得的关注程度，$\boldsymbol{c}_j$用来表示图像子区域的上下文矢量。

为了使生成图像包含句子级别和单词级别特征，最终将多文本注意力生成网络的目标函数表达为：

$$\begin{cases}L=L_G+\lambda L_{\text{MDAMSM}}\\ L_G=\sum_{i=0}^{m-1}L_{G_i}\end{cases} \tag{15}$$

式（15）中的超参数λ决定了多文本深度注意力多模态相似度模块对生成器损失函数L的影响程度。L_G包含各阶段的生成器损失，L_{MDAMSM}为第2.2.1节介绍的语义一致性网络损失。

L_{G_i}包含各个生成阶段的联合近似无条件分布损失和条件分布损失。无条件损失用于判定是真实图像还是生成图像，条件损失用于判定生成图像与句子向量是否匹配。G_i的损失函数为：

$$L_{G_i}=-\frac{1}{2}E_{\hat{x}_i\sim P_{G_i}}\left[\log\left(D_i(\hat{x}_i)\right)\right]-\frac{1}{2}E_{\hat{x}_i\sim P_{G_i}}\left[\log\left(D_i(\hat{x}_i,\boldsymbol{e})\right)\right] \tag{16}$$

各阶段判别器的损失函数L_{D_i}与生成器损失函数L_{G_i}类似，分为条件损失L_1^{con}和非条件损失$L_2^{\text{un-con}}$这两部分。不同于生成器，判别器D_i通过最小化定义的交叉熵损失函数进行分类判定，表示为：

$$\begin{cases}L_D=\sum_{i=0}^{m-1}L_{D_i}\\ L_{D_i}=L_1^{\text{con}}+L_2^{\text{un-con}}\end{cases} \tag{17}$$

L_1^{con}用于判断生成图像与文本描述内容是否一致，$L_2^{\text{un-con}}$用于区分真实图像与生成图像，计算式为：

$$\begin{cases}L_1^{\text{con}}=-\frac{1}{2}E_{x_i\sim P_{\text{data}_i}}\left[\log\left(D_i(x_i,\boldsymbol{e})\right)\right]-\frac{1}{2}E_{\hat{x}_i\sim P_{G_i}}\left[\log\left(1-D_i(\hat{x}_i,\boldsymbol{e})\right)\right]\\ L_2^{\text{un-con}}=-\frac{1}{2}E_{x_i\sim P_{\text{data}_i}}\left[\log\left(D_i(x_i)\right)\right]-\frac{1}{2}E_{\hat{x}_i\sim P_{G_i}}\left[\log\left(1-D_i(\hat{x}_i)\right)\right]\end{cases} \tag{18}$$

其中，x_i是来自第i个阶段的真实图像分布P_{data_i}，$\hat{x}_i$是来自第i个阶段生成器模型分布P_{G_i}的生成图像分布。

2.2.1　多文本深度注意力多模态相似度模型

多文本深度注意力多模态相似模型是一个半监督学习模型，用于学习多文本注意力模型，并对整个图像与整个文本序列之间的匹配概率进行监督。多文本深度注意多模态相似度模型通过学习2个神经网络（文本编码器和图像编码器），将文本词向量与图像子区域映射到相同语义空间，从而实现单词级别上图像—文本语义一致性，用于优化降低图像生成过程中产生的细粒度损失。

本文采用基于ImageNet预训练的Inception-V3模型作为图像编码器，用于将图像映射到语

义向量。Inception-V3的中间层网络可获取到输入图像子区域的局部特征，再经过后面的层可得到图像全局特征。首先利用上采样层up_sample，将输入图像分辨率重新调整为299×299，经过Mixed_6e层后得到图像局部向量$\boldsymbol{f}$，最后由平均池化层得到图像全局向量$\bar{\boldsymbol{f}}$。为了保证生成图像与文本描述的语义一致性，需要通过感知器层W，将图像特征与文本特征映射到相同语义空间，该过程如式（19）、式（20）所示。

$$\boldsymbol{v}=W\boldsymbol{f} \tag{19}$$

$$\bar{\boldsymbol{v}}=W\bar{\boldsymbol{f}} \tag{20}$$

采用双向长短期记忆（Bi-directional long short-term memory，Bi-LSTM）网络作为文本编码器，用于获取文本描述对应的语义向量。在Bi-LSTM中，每个单词都具有2个隐藏状态，每个方向分别对应一个，通过串联2个隐藏状态可用于表示一个单词的语义。此外，全局句子向量是将Bi-LSTM的最后一个隐藏状态连接形成的。

为了提升训练效果，计算出单词向量$\boldsymbol{w}$与图像子区域向量$\boldsymbol{v}$之间的相似度矩阵$\boldsymbol{s}$之后，再对其进行归一化处理，得到句子的每个单词与图像各个子区域之间的点积相似度$\bar{\boldsymbol{s}}_{j,i}$。

$$\boldsymbol{s}=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{v} \tag{21}$$

$$\bar{\boldsymbol{s}}_{j,i}=\frac{\exp\left(\boldsymbol{s}_{i,j}\right)}{\sum_{k=0}^{T-1}\left(\boldsymbol{s}_{k,j}\right)} \tag{22}$$

然后利用建立好的注意力模型来计算每个单词所对应的区域上下文向量$\boldsymbol{c}_i$，使用与句子的第i个单词相关的图像子区域来进行动态表示。

$$\boldsymbol{c}_i=\sum_{j=0}^{288}\boldsymbol{\alpha}_j\boldsymbol{v}_j \tag{23}$$

$$\boldsymbol{\alpha}_j=\frac{\exp\left(\gamma_1\bar{\boldsymbol{s}}_{i,j}\right)}{\sum_{i=0}^{288}\left(\gamma_1\bar{\boldsymbol{s}}_{i,k}\right)} \tag{24}$$

其中，$\boldsymbol{v}_j$表示第j个图像子区域特征，γ_1是决定关注度大小的一个影响因素，它决定了计算一个单词的区域上下文向量时对其相关图像子区域特征所需要的注意力，该注意力构成了注意力矩阵$\boldsymbol{\alpha}$，$\boldsymbol{\alpha}_j$为第j个图像子区域的注意力权重。

最后，第i个单词和图像之间的相关性由上下文矢量$\boldsymbol{c}_i$和单词特征向量$\boldsymbol{w}$之间的余弦相似度来定义，整个图像Q和整个文本描述D的相似度为：

$$R\left(\boldsymbol{c}_i,\boldsymbol{w}_i\right)=\frac{\boldsymbol{c}_i^{\mathrm{T}}\boldsymbol{w}_i}{\left\|\boldsymbol{c}_i\right\|\left\|\boldsymbol{w}_i\right\|} \tag{25}$$

$$R\left(Q,D\right)=\log\left(\sum_{i=1}^{T-1}\exp\left(\gamma_2R\left(\boldsymbol{c}_i,\boldsymbol{w}_i\right)\right)\right)^{\frac{1}{\gamma_2}} \tag{26}$$

其中，γ_2是一个影响因子，它决定了在多大程度上放大最相关的单词和区域上下文对的重要性。

图2为多文本深度注意力多模态相似度模型结构，文中使用该模型来引导生成器合成与标题相关的图像，因此损失函数表达为：

$$\hat{L}_{\mathrm{MDAMSM}}\left(\mathrm{Img},\mathrm{txt}\right)=\sum_{k=0}^{N_{\mathrm{txt}}}L_{\mathrm{MDAMSM}}\left(\mathrm{Img},\mathrm{txt}_k\right) \tag{27}$$

$$\begin{aligned}L_{\mathrm{MDAMSM}}\left(\mathrm{Img},\mathrm{txt}\right)=&L_1^w\left(\boldsymbol{f}_{\mathrm{img}}^{\mathrm{part}}\left(\mathrm{Img}\right),\boldsymbol{f}_{\mathrm{word}}^{\mathrm{txt}}\left(\mathrm{txt}_k\right)\right)+\\&L_2^w\left(\boldsymbol{f}_{\mathrm{img}}^{\mathrm{part}}\left(\mathrm{Img}\right),\boldsymbol{f}_{\mathrm{word}}^{\mathrm{txt}}\left(\mathrm{txt}_k\right)\right)+\\&L_1^s\left(\boldsymbol{f}_{\mathrm{img}}^{\mathrm{full}}\left(\mathrm{Img}\right),\boldsymbol{f}_{\mathrm{sent}}^{\mathrm{txt}}\left(\mathrm{txt}_k\right)\right)+\\&L_2^s\left(\boldsymbol{f}_{\mathrm{img}}^{\mathrm{full}}\left(\mathrm{Img}\right),\boldsymbol{f}_{\mathrm{sent}}^{\mathrm{txt}}\left(\mathrm{txt}_k\right)\right)\end{aligned} \tag{28}$$

其中，$\boldsymbol{f}_{\mathrm{img}}^{\mathrm{part}}$和$\boldsymbol{f}_{\mathrm{img}}^{\mathrm{full}}$是利用基于Inception-V3模型的图像编码器提取的局部特征和全局特征，$\boldsymbol{f}_{\mathrm{word}}^{\mathrm{txt}}$和$\boldsymbol{f}_{\mathrm{sent}}^{\mathrm{txt}}$是文本编码器提取的单词向量和句子向量，图像Img和文本描述txt之间的相似度是从图像Img中提取到的特征向量与其对应的注意力表示之间计算所得的余弦相似度；L_1^w和L_2^w是图像局部向量与其对应描述词向量相似性的交叉熵损失函数，L_1^s和L_2^s是描述图像全局向量与其对应描述句向量相似性的交叉熵损失函数。

2.2.2 联合训练目标函数

多文本合成的图像也应该保证图文语义一致性。在多条文本约束的条件下训练，总目标函数应该改为：

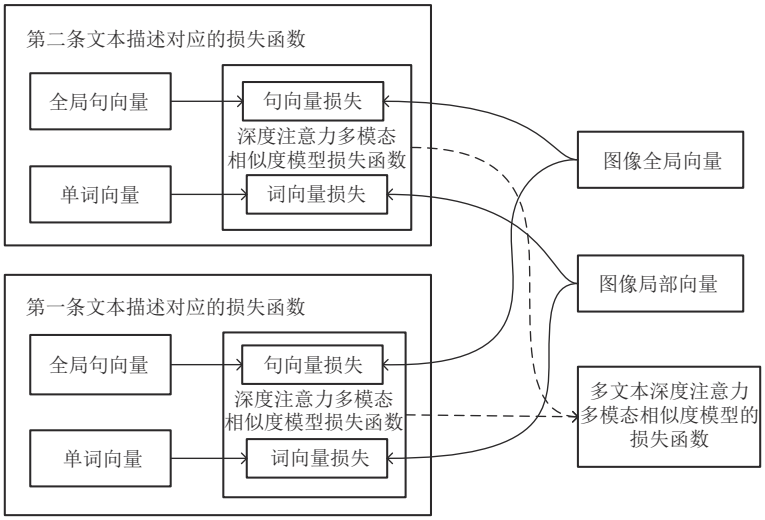


图2　多文本深度注意力多模态相似度模型结构

$$F\left(G_1,\cdots,G_k|\mathrm{TXT},D_1,\cdots,D_k\right)=\lambda E_{\hat{x}_i\sim P_{G_k}(\mathrm{TXT})}\left[\hat{L}_{\mathrm{MDAMSM}}\left(\hat{x},\mathrm{TXT}\right)\right]+$$
$$\sum_{i=1}^{K}\left\{E_{x\sim P_{\mathrm{data}}(\mathrm{TXT})}\left[\log D_i\left(x|\mathrm{TXT}\right)\right]+E_{\hat{x}\sim P_{\mathrm{G}_i}(\mathrm{TXT})}\left[\log\left(1-D_i\left(\hat{x}|\mathrm{TXT}\right)\right)\right]\right\}\tag{29}$$

其中，$\hat{x}\sim P_{\mathrm{G}_i}(\mathrm{TXT})$是给定文本TXT，由生成器$G_i$合成的图像；$\lambda$为决定多文本深度注意力多模态相似度模型对生成器损失函数影响程度的超参数；k为生成器的级数。

3　实验与分析

3.1　实验环境与数据集

本文实验环境为Ubuntu18.04，Cuda11.4，Python版本为3.8，实验代码采用Pytorch学习框架，在GPU上运行。

本文使用的数据集为CUB-200-2011[30]和MS-COCO[31]，其中CUB数据集是目前用于细粒度分类研究的基准图像数据集。该数据集共包含200个鸟类子类，11 788张图像，其中训练集和测试集分别有8 522张图像和2 933张图像。数据集中的每张图像都有10条对应的文本描述，描述内容包含鸟类的颜色、头部、翅膀等属性信息。MS-COCO数据集包含200个种类，约有12万张图像，每张图像还有5句人工生成的高质量英文描述，每个图像的注释包括对象边界框、对象分割掩码和图像描述等可用于图像描述生成和理解任务。训练次数设置为600，batch_size大小设置为20，生成器与判别器的学习率均为0.000 2，优化器采用Adam。

3.2　评价指标

本文采用IS（inception score）和FID（frechet inception distance）分数来量化本文实验结果。

IS通常用来评价GAN生成图像的清晰度和多样性。本文实验用到的IS指标是Xu等[32]提出的一套完整的评价算法。算法原理为：

$$\mathrm{IS}(G)=\exp\left(E_{x\sim P_g}\left[D_{\mathrm{KL}}\left(p\left(y|x\right)\|\left(p\left(y\right)\right)\right]\right)\tag{30}$$

其中，$p(y)$表示模型预测标签y的条件概率，$p(y|x)$表示生成图像样本x经过图像编码器得到的边缘概率，通过计算这两者之间的KL散度来衡量2个分布间的相似度。IS值越大，说明生成图像质量越高，细节越丰富。

FID分数是另一种常用指标，用于计算真实图像和生成图像在特征空间上的距离。FID分数越低，则表明生成图像与真实图像越接近，也就

意味着生成图像越真实。算法原理为：

$$\mathrm{FID}=\left\|\boldsymbol{\mu}_{\mathrm{r}}-\boldsymbol{\mu}_{\mathrm{g}}\right\|^{2}+\mathrm{tr}(\Sigma x+\Sigma g2\left(\Sigma x\Sigma g\right)^{\frac{1}{2}}) \quad (31)$$

其中，$\boldsymbol{\mu}_{\mathrm{r}}$、$\boldsymbol{\mu}_{\mathrm{g}}$为真实图像和生成图像均值，$\Sigma x$和$\Sigma g$分别代表真实图像和生成图像特征协方差。

3.3 结果分析

本文将从评价指标与视觉效果2个角度来分析实验效果。首先是利用评价指标IS和FID进行量化分析，将本文方法与原始AttnGAN生成图像在CUB数据集和MS-COCO数据集上进行对比；紧接着将本文方法和原始AttnGAN生成图像形成视觉对比，从而验证本文方法的有效性。

3.3.1 评价指标对比

本文提出的多文本注意力生成对抗网络模型在24 GB内存的显卡上进行训练，并在CUB、MS-COCO数据集上进行指标对比，不同方法在CUB数据集上的IS和FID见表1。最终在CUB数据集上的IS指标分数为4.19，FID指标分数为32.78，在MS-COCO数据集上的IS指标和FID指标分数分别为19.78和39.26。

表1 不同方法在CUB数据集上的IS和FID

方法	CUB		MS-COCO	
	IS	FID	IS	FID
GAN-INT-CLS[11]	2.8±0.04	68.79	7.88±0.07	60.62
GAWWN[12]	3.62±0.07	—	—	—
StackGAN[13]	3.70±0.04	51.89	8.41±0.03	74.05
StackGAN++[14]	4.04±0.05	15.30	8.30±0.08	81.59
AttnGAN[15]	4.36±0.03	27.86	25.89±0.47	35.49
多文本AttnGAN	3.54±0.09	35.14	8.74±0.06	69.54
本文方法	4.19±0.09	32.78	19.78±0.05	39.26

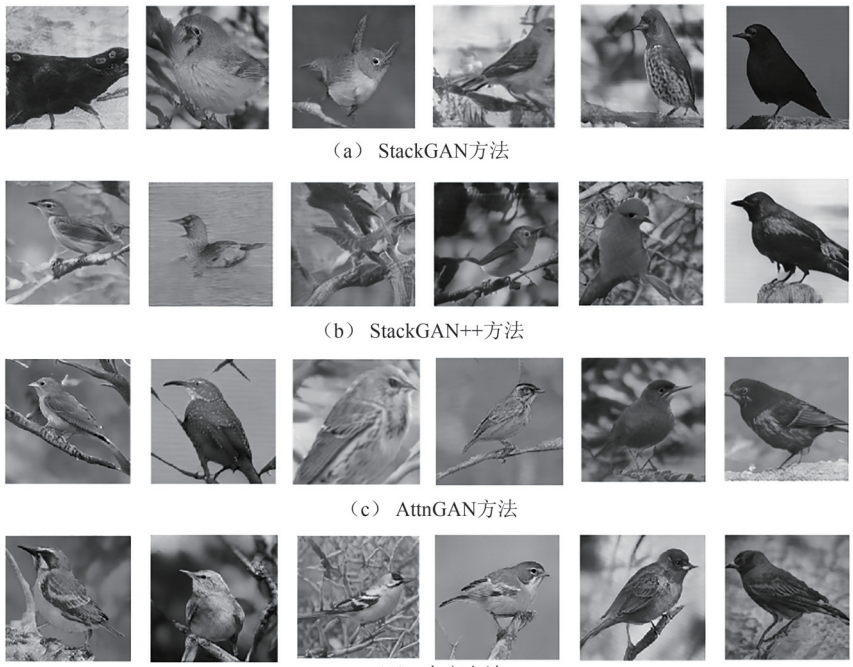
对比表1中的实验数据发现，本文方法在CUB和MS-COCO数据集上的实验结果与多种单文本生成图像的方法相比，IS指标有一定的提升，FID指标降低明显。本文方法生成图像所得IS、FID指标与单文本AttnGAN方法相比有所差距，但同样是处理多文本的情况下，AttnGAN方法的效果反而不如本文方法理想。实验结果表明，本文方法生成的图像质量良好，且图像的细节信息以及图像清晰度均有提高，图像内容更加接近真实图像。

3.3.2 合成图像对比

在视觉效果方面，CUB数据集上的生成图像对比如图3所示，图3的每一列都是在相同文本条件下合成的图像。其中，StackGAN、StackGAN++和AttnGAN方法均采用官方实现模型，在同一实验环境下进行实验。

可以看出，图3的第1、2、3列中，本文方法生成图像比较完整，且各个部位细节丰富合理，不同部位间的过渡十分自然。而其他模型生成的图像细节部分缺失明显，且结构上存在明显错误，整体看上去十分突兀。在图3的第4、5、6列中，可以明显看出本文方法生成的图像具有完整结构，目标对象结构与背景衔接自然。而StackGAN提出了堆叠生成对抗网络模型，侧重于提高生成图像的分辨率，但是忽略部分鸟类属性，以至于生成的鸟类缺少细粒度信息，例如生成鸟类的结构不完整，缺少爪子、喙等部位；StackGAN++进一步提升了生成图像的分辨率，其他问题却没能解决；AttnGAN增加了注意力机制，能够增强文本到图像的语义一致性，但是没有很好地学习到鸟类的重要属性特征，导致生成图像缺乏部分细节信息，容易出现结构性错误问题，背景混乱，甚至会出现2个头等情况。

除CUB数据集外，本文方法也在更有挑战性的MS-COCO数据集上进行了实验，并与其他几个具有代表性的方法进行了视觉性比较，3种方法均在同一文本下生成图像，MS-COCO数据集上的生成图像对比如图4所示。可以看出，StackGAN和StackGAN++模型无法生成现实图像，只具有部分目标对象的大致形状。AttnGAN模型引入了局部注意力机制来获取生成器在图像生成过程中不同子区域的权重信息，最终生成图像效果



（a） StackGAN方法

（b） StackGAN++方法

（c） AttnGAN方法

（d） 本文方法

图3　CUB数据集上的生成图像对比

（a）StackGAN方法

（b）StackGAN++方法

（c）AttnGAN方法

（d）本文方法

图4　MS-COCO数据集上的生成图像对比

虽然也有提升，但是与本文方法相比，仍缺少许多局部细节信息。尽管MS-COCO数据集文本中包含的场景内容更加复杂，但本文方法在该数据集上生成图像的语义一致性依旧很好，目标对象的轮廓和布局也更加合理，如第2列文本提到的“堆满了盘子”，第5列文本中的“海滩”。

3.3.3　消融实验

为了验证本文提出的流行插值操作、多文本

深度注意力多模态相似度模型与自注意力模块的有效性，在CUB数据集上设置了4组对比试验。

本文的基础网络为AttnGAN，MI表示流形插值操作，MC-DAMSM表示多文本深度注意多模态相似度模型，SA表示自注意力模块，3个模块结合多文本注意力生成对抗网络模型可得到CUB数据集上的消融实验结果，见表2。结果表明，流形插值操作、多文本深度注意多模态相似度模型和自注意力模块都在一定程度上提高了生成图像质量，证明了本文方法的有效性。

本文方法采用流形插值来结合多文本语义特征，一定程度上丰富了文本特征，使得生成图像具有更多的细节。接着引入多文本深度注意多模态相似度模块，增强图像与文本的语义一致性，不仅生成了高分辨率图像，生成的鸟类具备了丰富的细节，鸟类的形态也更加接近真实图像。自注意力模块的加入不仅让模型学习到图像中重要的位置和空间信息，改善了生成图像中目标对象结构错误的情况，还提升了文本生成图像模型的稳定性。CUB数据集上消融实验结果对比如图5所示。

表2　CUB数据集上的消融实验结果

	MI	MC-DAMSM	SA	IS	FID
多文本AttnGAN（BS）				3.54±0.09	35.14
	√			3.87±0.11	37.43
	√	√		4.05±0.11	35.75
	√	√	√	4.19±0.09	32.78

在实验过程中，多文本AttnGAN会合成一些失败的图像，鸟类生成失败情况对比如图6所示，每一行都是在相同文本条件下合成的图像。如图6（a）中多文本AttnGAN方法生成了2个鸟嘴，图6（b）列举了生成2个鸟头的情况，

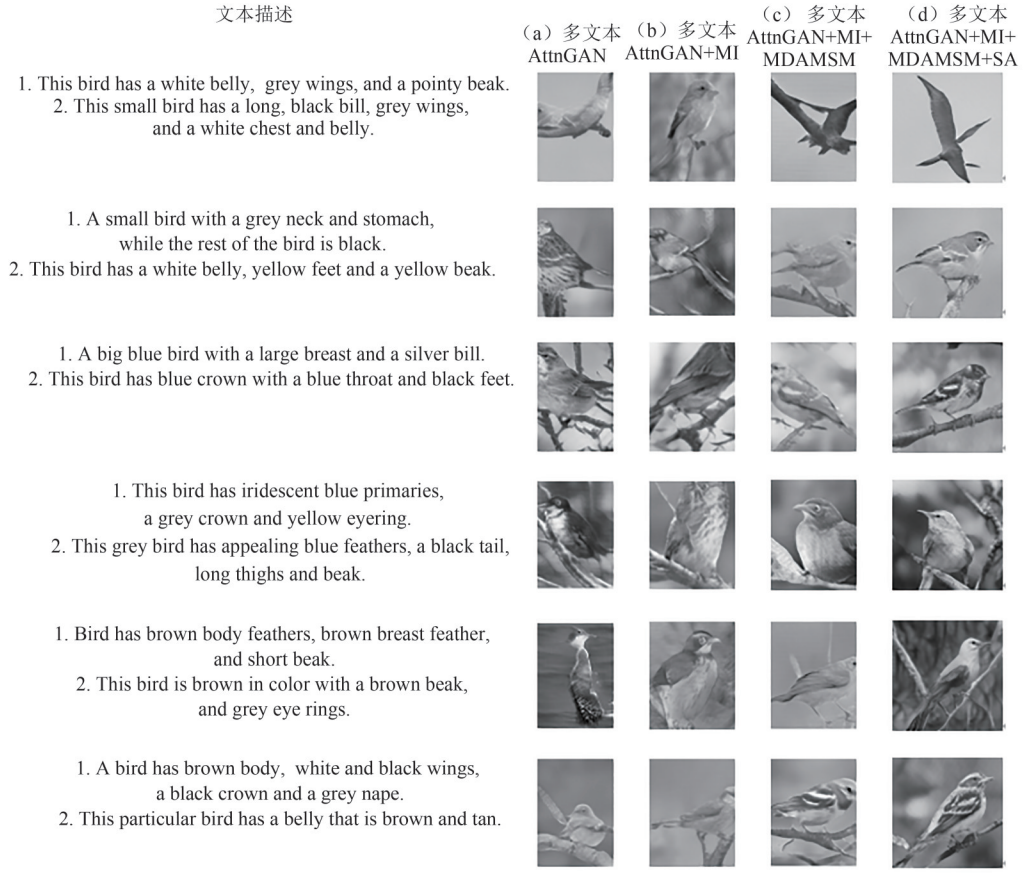
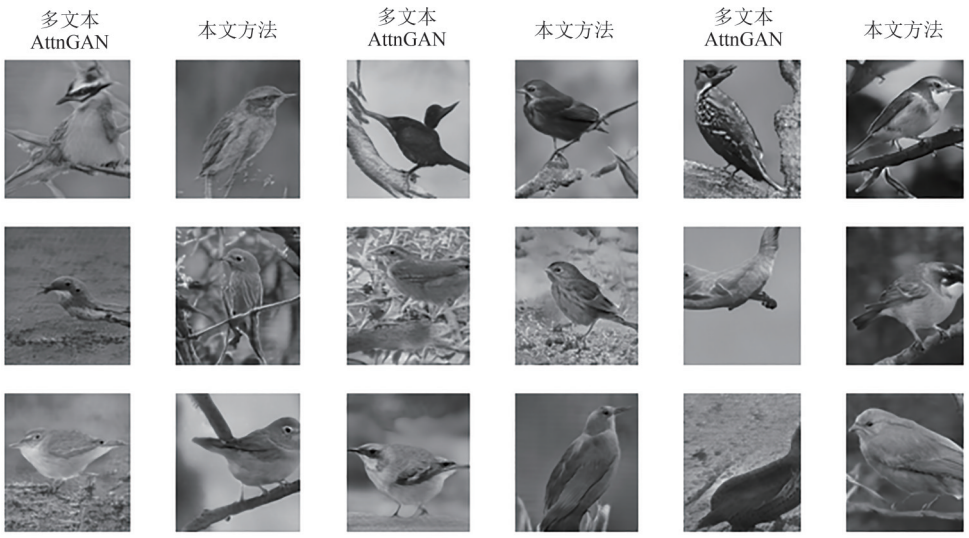


图5　CUB数据集上消融实验结果对比



（a）“多嘴”情况对比　（b）“多头”情况对比　（c）整体结构错误情况对比

图6　鸟类生成失败情况对比

图6（c）中多文本AttnGAN方法生成的图像根本看不出是鸟类，不符合现实生活中正常鸟类形态。与上述情况不同的是，本文方法生成的鸟类图像具有正确的鸟类结构，能够正确形成鸟头、鸟嘴的个数信息，明显改善了多文本AttnGAN方法生成图像与文本条件特征不符的问题。通过对比可以看出，本文方法生成图像不仅能够准确捕捉到鸟类信息，还能对生成对象与背景进行区分，而多文本AttnGAN方法生成图像则会产生一定程度上的失真。

4　结束语

本文针对生成图像细节不足以及存在结构性错误的问题，提出了一种以AttnGAN为基础框架的文本生成图像模型。通过在文本描述间进行流形插值操作，丰富给定文本描述的属性特征，解决生成图像细节不足的问题。接着引入多文本深度注意多模态相似度模块，优化生成图像与文本描述之间的相似度，生成贴合语义特征的图像。最后在生成器的初始阶段引进自注意力模块，在优化生成图像质量的同时，也提升了模型的稳定性。在CUB和MS-COCO数据集上进行实验，IS指标与FID指标均有明显的提升。实验结果表明，本文提出的方法能够充分提取利用文本描述中的属性特征，生成细节丰富、清晰多样的图像，并且模型也具有一定的泛化性。但是还存在一些问题，未来可以从以下几个方向进行探索。

（1）扩充文本生成图像数据集。对于本文提出的模型，目前仅利用CUB数据集和MS-COCO数据集对模型进行训练，还未对中文数据集进行应用。未来将继续探索与制作相关数据集，将文本生成图像的研究成果应用于实际，进一步优化模型的稳定性与应用前景。

（2）加强对复杂场景图像生成的研究。现有模型在一般简单的花鸟数据集上生成图像质量较好，然而在VisualGenome和ImageNet这种包含复杂场景和多目标对象的数据集上，生成图像效果不尽如人意。因此，下一阶段的主要任务就是加强在复杂场景图像生成方面的研究。

参考文献：

[1] WU X, XU K, HALL P. A survey of image synthesis and editing with generative adversarial networks[J]. Tsinghua Science and Technology, 2017, 22(6): 660-674.

[2] AGNESE J, HERRERA J, TAO H C, et al. A survey and tax-

onomy of adversarial neural networks for text-to-image synthesis[J]. arXiv preprint, 2019, arXiv: 1910.09399.

[3] GREGOR K, DANIHELKA I, GRAVES A, et al. DRAW: a recurrent neural network for image generation[J]. arXiv preprint, 2015, arXiv: 1502.04623.

[4] KINGMA D P, WELLING M. Auto-encoding variational Bayes[J]. arXiv preprint, 2013, arXiv: 1312.6114.

[5] ZENG X M, LONG L Q. Generative adversarial networks[M]. Beginning Deep Learning with TensorFlow. Berkeley, CA: Apress, 2022: 553-599.

[6] MIRZA M, OSINDERO S. Conditional generative adversarial nets[J]. arXiv preprint, 2014, arXiv: 1411.1784.

[7] RUAN S L, ZHANG Y, ZHANG K, et al. DAE-GAN: dynamic aspect-aware GAN for text-to-image synthesis[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2021: 13940-13949.

[8] YANG Z Y, WANG J F, GAN Z, et al. ReCo: region-controlled text-to-image generation[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2023: 14246-14255.

[9] PARMAR G, SINGH K, ZHANG R, et al. Zero-shot image-to-image translation[C]//ACM SIGGRAPH 2023 Conference Proceedings. 2023: 1-11.

[10] TUMANYAN N, GEYER M, BAGON S, et al. Plug-and-play diffusion features for text-driven image-to-image translation[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2023: 1921-1930.

[11] REED S, AKATA Z, YAN X C, et al. Generative adversarial text to image synthesis[C]//33rd International Conference on Machine Learning (ICML). JMLR. org, 2016.

[12] REED S, AKATA Z, MOHAN S, et al. Learning what and where to draw[J]. Advances in Neural Information Processing Systems, 2016, 29.

[13] ZHANG H, XU T, LI H S, et al. StackGAN: text to photo-realistic image synthesis with stacked generative adversarial networks[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2017: 5908-5916.

[14] ZHANG H, XU T, Ll H, et al. StackGAN++: realistic image synthesis with stacked generative adversarial networks[J]. IEEE Transactions on Pattern Analysis and Machine lntelligence, 2019: 1947-1962.

[15] XU T, ZHANG P C, HUANG Q Y, et al. AttnGAN: fine-grained text to image generation with attentional generative adversarial networks[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 1316-1324.

[16] ZHU M F, PAN P B, CHEN W, et al. DM-GAN: dynamic memory generative adversarial networks for text-to-image synthesis[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 5795-5803.

[17] ZHANG Z Z, XIE Y P, YANG L. Photographic text-to-image synthesis with a hierarchically-nested adversarial network[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 6199-6208.

[18] GAO L L, CHEN D Y, SONG J K, et al. Perceptual pyramid adversarial networks for text-to-image synthesis[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 8312-8319.

[19] TAO M, TANG H, WU S, et al. DF-GAN: a simple and effective baseline for text-to-image synthesis[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 16515-16525.

[20] ZHANG Z Q, ZHOU J J, YU W X, et al. Drawgan: text to image synthesis with drawing generative adversarial networks[C]//Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2021: 4195-4199.

[21] DASH A, GAMBOA J C B, AHMED S, et al. TAC-GAN-text conditioned auxiliary classifier generative adversarial network[J]. arXiv preprint, 2017, arXiv: 1703.06412.

[22] LI W B, ZHANG P C, ZHANG L, et al. Object-driven text-to-image synthesis via adversarial training[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 12166-12174.

[23] LIAO W T, HU K, YANG M Y, et al. Text to image generation with semantic-spatial aware GAN[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 18166-18175.

[24] CHENG J, WU F X, TIAN Y L, et al. RiFeGAN: rich feature generation for text-to-image synthesis from prior knowledge[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 10908-10917.

[25] ZHANG Z X, SCHOMAKER L. DTGAN: dual attention generative adversarial networks for text-to-image generation[C]//Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE Press, 2021: 1-8.

[26] BENGIO Y, MESNIL G, DAUPHIN Y, et al. Better mixing via

deep representations[C]//30th International Conference on Machine Learning, ICML 2013, 2013(PART 1): 552-560.

[27] REED S, SOHN K, ZHANG Y T, et al. Learning to disentangle factors of variation with manifold interaction[C]//Proceedings of the Proceedings of the 31st International Conference on Machine Learning. New York: ACM, 2014: Ⅱ-1431-Ⅱ-1439.

[28] MANSIMOV E, PARISOTTO E, LEI BA J, et al. Generating images from captions with attention[J]. arXiv e-prints, 2015, arXiv: 1511.02793.

[29] ZHANG H, GOODFELLOW I, METAXAS D, et al. Self-attention generative adversarial networks[C]//International Conference on Machine Learning. PMLR, 2019: 7354-7363.

[30] WAH C, BRANSON S, WELINDER P, et al. The caltech-UCSD birds-200-2011 dataset[J]. California Institute of Technology, 2011.

[31] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[D]. Computer Vision-ECCV 2014. Cham: Springer International Publishing, 2014: 740-755.

[32] SALIMANS T, GOODFELLOW I, ZAREMBA W, et al. Improved techniques for training GANs[J]. arXiv preprint, 2016, arXiv: 1606.03498.

[作者简介]

聂开琴（1998– ），女，浙江工商大学硕士生，主要研究方向为深度学习。

倪郑威（1989– ），男，博士，浙江工商大学副研究员，主要研究方向为机器学习、物联网、无线通信等。