结合语义分割图的注意力机制文本生成图像

梁成名，李云红，李丽敏，苏雪平，朱绵云，朱耀麟
（西安工程大学电子信息学院，西安，710048）

摘要 针对生成对抗网络生成图像存在结构不完整、内容不真实、质量差的问题，提出一种结合语义分割图的注意力机制文本到图像生成模型(SSA-GAN)。首先采用一种简单有效的深度融合模块，以全局句子向量作为输入条件，在生成图像的同时，充分融合文本信息。其次结合语义分割图像，提取其边缘轮廓特征，为模型提供额外的生成和约束条件。然后采用注意力机制为模型提供细粒度词级信息，丰富所生成图像的细节。最后使用多模态相似度计算模型计算细粒度的图像-文本匹配损失，更好地训练生成器。通过CUB-200和Oxford-102 Flowers数据集测试并验证模型，结果表明：所提模型(SSA-GAN)与StackGAN、AttnGAN、DF-GAN以及RAT-GAN等模型最终生成的图像质量相比，IS指标值最高分别提升了13.7%和43.2%，FID指标值最高分别降低了34.7%和74.9%，且具有更好的可视化效果，证明了所提方法的有效性。

关键词 文本生成图像；语义分割图像；生成对抗网络；注意力机制；仿射变换
DOI 10.3969/j.issn.2097-1915.2024.04.016
中图分类号 TP391.41 **文献标志码** A **文章编号** 2097-1915(2024)04-0118-10

A Semantic Segmentation Graph in Combination with Attention Mechanism Text Generation Images

LIANG Chengming, LI Yunhong, LI Limin, SU Xueping, ZHU Mianyun, ZHU Yaolin
(School of Electronics and Information, Xi'an Polytechnic University, Xi'an 710048, China)

Abstract Aimed at the problems that generative adversarial network is incomplete in structure, unreal in content and poor in quality of images generated, an attention mechanism text-to-image generation model combined with semantic segmentation graph (SSA-GAN) is proposed. First, taking global sentence vectors as input conditions, a simple and effective deep fusion module is utilized for fully fusing text information while generating images are generating simultaneously. Second, the semantically segmented images are combined to extract their edge profile features to provide additional generative and constraint conditions for the model, and the attention mechanism is used to provide fine-grained word-level information for the model to enrich the details of the generated images. Finally, a multimodal similarity computation model is used to compute fine-grained image-text matching loss to further train the generator. The model is tested and validated by CUB-200 and Oxford-102 Flowers datasets, and the results show that the proposed model (SSA-GAN) improves the quality of the final generated images. Compared to the models such as Stack-

收稿日期：2023-10-07
基金项目：国家自然科学基金(62203344)；陕西省自然科学基础研究重点项目(2022JZ-35)；陕西高校青年创新团队项目
作者简介：梁成名(1998—)，男，河南信阳人，硕士生，研究方向为电子信息。E-mail：1850716201@qq.com
通信作者：李云红(1974—)，女，辽宁锦州人，教授，博士，研究方向为红外热像测温技术、图像处理、信号与信息处理技术等。E-mail：hitliyunhong@163.com

引用格式：*梁成名，李云红，李丽敏，等. 结合语义分割图的注意力机制文本生成图像[J]. 空军工程大学学报，2024，25(4)：118-127. LIANG Chengming, LI Yunhong, LI LiMin, et al. A Semantic Segmentation Graph in Combination with Attention Mechanism Text Generation Images [J]. Journal of Air Force Engineering University, 2024, 25(4): 118-127.*

GAN, AttnGAN, DF-GAN, and RAT-GAN, the IS increases in metrics values by 13.7% and 43.2%, respectively. And the FID in metric values is reduced to 34.7% and 74.9%, respectively.

Key words text generates images; semantic segmentation image; attention mechanism; generate adversarial network; affine transformation

随着深度学习技术[1-2]日益成熟，图像的获取方式越来越多样化，传统的图像获取方式效率低下，已无法满足当今人们的需求。如何使用深度学习技术自动生成符合语义要求的图像以满足人们日益增长的需求成为当今的研究热点，一系列图像生成任务应运而生[3-4]。其中，基于描述语句的图像合成是视觉与语言处理交叉领域的热门课题。它可应用于照片编辑、计算机辅助设计及虚拟场景的生成等。同时，该课题的研究也促进了跨视觉和语言领域的多模态学习与推理能力的研究进展，这也是近年来最为活跃的研究领域之一[5-8]。

近年来，大多数文本到图像生成的方法都是基于生成式对抗网络[9]（generative adversarial network，GAN）所提出的。2016 年，GAN-CLS 和 GAN-INT-CLS 算法[10]首次将深度卷积生成式对抗网络模型（deep convolutional generative adversarial network，DCGAN）用于文本描述到图像生成任务。GAN-CLS 算法通过加入分类匹配判别器，确定了图片的真伪性及其文字描述与图片的对应情况。而 GAN-INT-CLS 算法则通过插值的方法，在已有的文本特征中进行插值运算，有效增加了生成图像的多样性。通过实验证实，2 种算法均可生成 64×64 分辨率的图像。但是，由于模型直接将高维度的文本特征向量送到生成器中，造成了文本表示的稀疏性，从而在生成的图像中出现了结构缺失和图像失真的现象。继 GAN-INT-CLS 之后，Reed 等[11]提出了 GAWWN 模型，该模型使用关键点（keypoint）和边界（bounding box）对对象的位置进行信息标记，提升了生成图像的质量，生成了 128×128 分辨率的图像。为进一步提高生成图像的质量，Zhang 等[12]提出了堆叠式生成对抗网络（StackGAN），通过将 2 个条件生成式对抗网络进行叠加，由低到高分别生成不同分辨率的图像，进一步改善了生成图像的质量。但其因为缺乏词级别的细粒度信息，最终很难生成令人满意的高质量图像。因此，Xu 等[13]提出了注意力生成对抗网络（AttnGAN），首次证明了分层条件 GAN 能够自动关注相关单词，形成图像生成的条件。Qiao 等[14]为了使全局和局部的细节都能得到有效的关注，在 AttnGAN 的基础上再次引入了全局注意力，提出了 Mirror-GAN 模型，使文字描述和图像之间拥有更好的语义一致性。但其生成的图像存在结构不完整、内容不真实的情况。最近，很受欢迎的 DF-GAN[15]、RAT-GAN[16]和 GigaGAN[17]模型虽然提升了生成图像的质量，但结构不完整、内容不真实的问题仍未得到解决。主要原因有：①没有同时有效地兼顾细粒度的词级信息和全局文本信息，这限制了细粒度视觉特征合成的能力；②没有引入额外的生成和约束条件，导致后续处理很难将图像细化到更令人满意的结果。

为改善该问题，本文提出了结合语义分割图的注意力机制文本生成图像模型（combining semantic segmentation graphs with attention mechanism text generation images，SSA-GAN）。提出了深度融合模块（deep fusion module，DFMBlock），在融合全局文本信息和图像特征的同时，使文本信息在生成过程中得到充分保留，加强了文本与图像的融合，防止文本信息在图像生成过程中的逐渐丢失，避免对生成图像的语义一致性造成影响。引入语义分割图作为初始阶段生成器的额外生成条件，并将其加入条件损失函数的约束条件之中，约束模型生成结构完整、内容真实的低分辨率图像。使用注意力机制为模型提供细粒度词级信息，丰富图像细节，提高生成图像的质量。多模态相似度计算模型（multimodal similarity calculation model，MSCM）为模型提供图像-文本匹配损失，更好地训练生成器。

1　模型结构设计

模型主要由注意生成网络和多模态相似度计算模型（MSCM）构成。其中注意力生成网络模型包含多个生成器，用于生成不同分辨率的图像。图 1 中的全局句子向量和词级向量都来自文本编码器对图像的文本描述编码。首先，生成网络在初始阶段以全局句子向量、随机噪声向量和语义分割图的特征向量经过深度融合模块后，生成边缘轮廓完整的低分辨率（64×64）图像。其次，利用各子区域的图像向量、词

向量和注意力层，形成词上下文向量，然后将区域图像向量与相应的词上下文向量结合，形成多模态上下文向量。最后，在此基础上生成周围子区域的新图像特征，使得各阶段生成的图形都具有更多丰富的细节。多模态相似度计算模型（multimodal similarity calculation model，MSCM）是模型的另一个组成部分，它将图像的子区域和句子的单词映射到一个公共语义空间。利用全局句子级信息和细粒度词级信息计算生成的图像与句子之间的余弦相似度，为训练生成器提供图像-文本匹配损失。

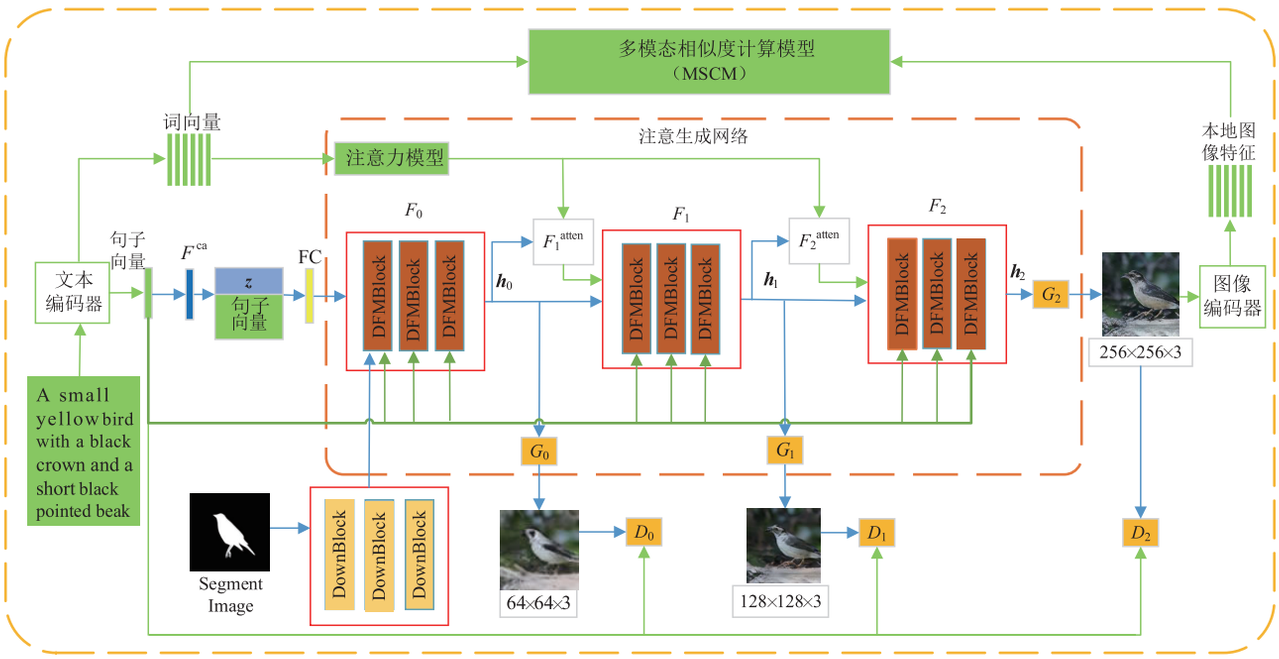

图 1 SSA-GAN 的宏观结构

1.1 注意生成网络

注意生成网络模型使生成网络能够根据与图像子区域最相关的单词绘制图像的不同子区域。所提出的网络模型具有 m 个生成器（$G_0,G_1,\cdots,G_{m-1}$），取隐状态（$\boldsymbol{h}_0,\boldsymbol{h}_1,\cdots,\boldsymbol{h}_{m-1}$）作为输入，生成不同尺度的图像（$\dot{\boldsymbol{x}}_0,\dot{\boldsymbol{x}}_1,\cdots,\dot{\boldsymbol{x}}_{m-1}$）。具体如式（1）所示：

$$\begin{cases}\boldsymbol{h}_0=F_0(\boldsymbol{z},F^{ca}(\bar{\boldsymbol{e}}),\boldsymbol{I}_s)\\ \boldsymbol{h}_i=F_i(\boldsymbol{h}_{i-1},F_i^{attn}(\boldsymbol{e},\boldsymbol{h}_{i-1}))\\ \dot{\boldsymbol{x}}=G_i(\boldsymbol{h}_i),i=0,1,\cdots,m-1\end{cases}\tag{1}$$

式中：$\boldsymbol{z}$ 为从标准正态分布中采样而来的噪声向量；$\bar{\boldsymbol{e}}$ 和 $\boldsymbol{e}$ 分别为全局句子向量和词向量；$\boldsymbol{I}_s$ 是语义分割图像经过下采样模块（DownBlock）提取的语义分割图像特征向量；F^{ca}、F_i^{attn}、F_i、G_i 被建模为神经网络，F^{ca} 为条件增强[12]，F_i^{attn} 为在模型的第 i 个阶段所采用的注意力模型。$F_i^{attn}(\boldsymbol{e},\boldsymbol{h})$ 的输入分别为单词特征向量 $\boldsymbol{e}\in\boldsymbol{R}^{D\times T}$ 和上一个隐藏层的图像特征 $\boldsymbol{h}\in\boldsymbol{R}^{\hat{D}\times N}$。

1.1.1 注意力模型

注意力模型首先通过感知层将单词和图像特征转换到公共语义空间，再根据图像的隐藏特征 $\boldsymbol{h}$ 计算单词上下文向量 $\boldsymbol{c}_j$。$\boldsymbol{h}$ 的每一列都是图像子区域的特征向量。对于第 j 个子区域，其词上下文的向量是与 $\boldsymbol{h}_j$ 相关的词向量的动态表示，即：

$$\boldsymbol{c}_j=\sum_{i=0}^{T-1}\beta_{j,i}\boldsymbol{e}'_i\tag{2}$$

式中：$\beta_{j,i}=\dfrac{\exp(s'_j,i)}{\sum_{k=0}^{T-1}\exp(s'_j,k)}$；$s'_{j,i}=\boldsymbol{h}_j^{\mathrm{T}}\boldsymbol{e}'_i$ 和 $\beta_{j,i}$ 分别为模型在生成图像第 j 个子区域时对第 i 个词的关注权值。

然后，通过式（3）生成图像特征集的词上下文矩阵，最后，将图像特征与相应的词-上下文特征相结合，生成下一阶段的图像。

$$F^{attn}(\boldsymbol{e},\boldsymbol{h})=(\boldsymbol{c}_0,\boldsymbol{c}_1,\cdots,\boldsymbol{c}_{N-1})\in\boldsymbol{R}^{\hat{D}\times N}\tag{3}$$

与一般方法不同的是：SSA-GAN 所使用的注意力模型以词级特征向量取代整个文本描述的特征向量作为注意力模型的输入，为条件 GAN 提供单词级别的细粒度信息，这有助于高质量图像的生成。

1.1.2 深度融合模块

如图 2 所示，深度融合模块（DFMBlock）由 3 个上采样模块（UPBlock）所构成，其中 UPBlock 又由 Affine 仿射块、ReLU 层和 Conve 层构成。在 UPBlock 中的 2 个 Affine 仿射块之间添加一个 ReLU 层，将非线性引入融合过程，扩大条件表示空间，促进视觉特征的多样性。

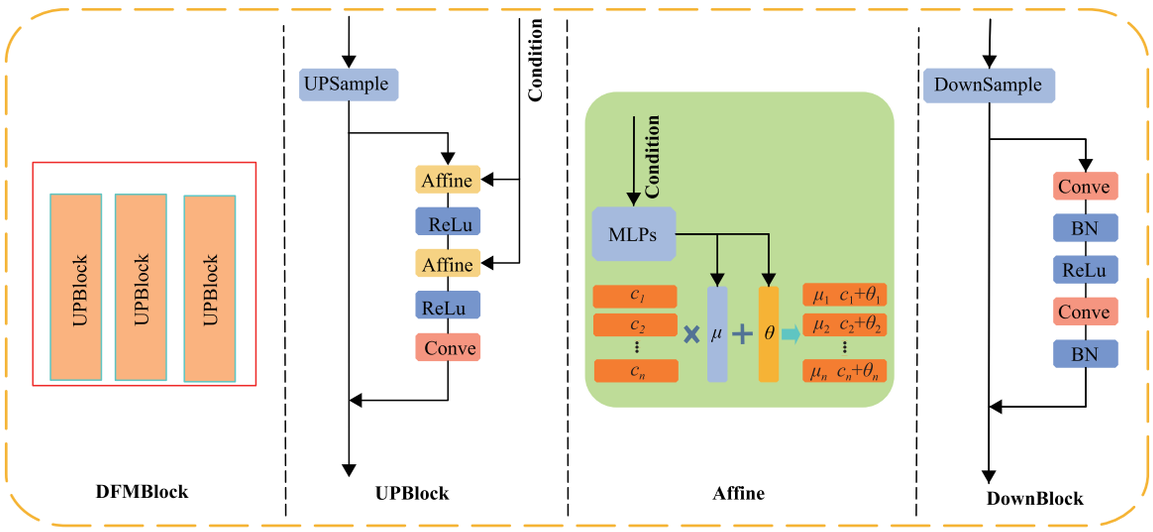

图 2　深度融合模块(DFMBlock)、UPBlock 、Affine、DownBlock 结构

Affine 仿射块由 2 个 MLP(多层感知器)组成，分别为 MLP 预测语言条件下的通道尺度参数 μ 和预测其移位参数 θ，即：

$$\Delta\mu=\mathrm{MLP}(\bar{e}),\Delta\theta=\mathrm{MLP}(\bar{e}),\quad \dot{\mu}=\mu+\Delta\mu,\dot{\theta}=\theta+\Delta\theta \tag{4}$$

Affine 仿射变换模块先使用参数 μ 对视觉特征图进行通道方向的标度运算，然后使用移位参数 θ 进行通道方向的移位运算，即：

$$\mathrm{AFF}(\hat{x}_i\mid\bar{e})=\mu_i,\hat{x}_i+\theta_i \tag{5}$$

式中：$\hat{x}_i$ 为视觉特征图的第 i 个通道信息；μ_i 和 θ_i 为视觉特征图第 i 通道的缩放参数和移位参数。

1.1.3　语义分割图像

语义分割图像是对应于文本描述的原图像数据集的分割图像，作为低分辨率图像生成条件之一。语义分割图像经过下采样模块(DownBlock)提取图像的轮廓特征，即对图像进行编码处理得到图像的特征向量。DownBlock 包括一系列的 Conve 层、BN 层、Relu 层。最终提取的语义分割图像特征向量可表示为：

$$\boldsymbol{I}_s=\frac{1}{k}\sum\nolimits_i\boldsymbol{I}_i,\boldsymbol{I}_s\in\boldsymbol{R}^D \tag{6}$$

式中：k 为局部区域的个数；$\boldsymbol{I}_i$ 为图像的第 i 个区域的特征向量，即局部特征向量，将局部特征向量的平均值作为图像的全局特征向量 $\boldsymbol{I}_s$。因此，初始阶段生成器的输入，除了细粒度的词级向量、全局文本特征向量、随机噪声向量外，还增加了语义分割图像的特征向量。

1.2　多模态相似度计算模型

MSCM 通过学习 2 个神经网络，将部分图像和句子中的词映射到一个公共语义空间，以便于计算图像-文字和句子级别之间的相似度，进而计算图像生成所产生的细粒度损失。文本编码器是一种双向长短期记忆网络(long short-term memory，LSTM)。它可以把一个词所对应的 2 个隐藏状态连接起来，用以描述词的意思。图像编码器是一种卷积神经网络(convolutional neural network，CNN)，它可以把图像信息映射到语义向量空间。模型的图像编码器建立在 ImageNet[18] 上预训练的 Inception-v3 模型[19]之上。首先将输入图像缩放为 299×299 像素，然后，从 Inception-v3 的“mixed_6e”层中提取局部特征矩阵 $\boldsymbol{f}\in\boldsymbol{R}^{768\times289}$，其中 $\boldsymbol{f}$ 的每一列都是图像一个子区域的特征向量，768 为局部特征向量的维数，289 为图像中的子区域数。同时，从 Inception-v3 的最后一个平均池化层中提取全局特征向量 $\bar{\boldsymbol{f}}\in\boldsymbol{R}^{2\,048}$。最后通过添加感知层，将图像特征转换为文本特征的公共语义空间。感知层函数为：

$$\boldsymbol{v}=\boldsymbol{W}\boldsymbol{f},\bar{\boldsymbol{v}}=\bar{\boldsymbol{W}}\bar{\boldsymbol{f}} \tag{7}$$

式中：$\boldsymbol{v}\in\boldsymbol{R}^{D\times289}$；$\boldsymbol{v}_i$ 为图像第 i 个子区域的视觉特征向量；$\bar{\boldsymbol{v}}\in\boldsymbol{R}^D$ 为整个图像的全局向量；D 为多模态(即图像和文本模态)特征空间的维度。注意力驱动的图形-文本匹配评分是基于图像和文本之间的注意力模型来计算图像-句子对的匹配程度。首先计算句子中所有可能的单词对和图像中的子区域的相似度矩阵，计算式为：

$$\boldsymbol{s}=\boldsymbol{e}^{\mathrm{T}}\boldsymbol{v} \tag{8}$$

式中：$\boldsymbol{s}\in\boldsymbol{R}^{T\times289}$；$\boldsymbol{s}_{i,j}$ 为第 i 个单词与图形的第 i 个子区域间的相似度点积。模型还对相似度矩阵进行归一化：

$$\bar{\boldsymbol{s}}_{i,j}=\frac{\exp(\boldsymbol{s}_{i,j})}{\sum_{K=0}^{T-1}\exp(\boldsymbol{s}_{k,j})} \tag{9}$$

然后，注意力模型为每个词（查询）计算一个区域上下文向量。区域的上下文向量 $\boldsymbol{c}_i$ 是与句子的第 i 个单词相关的图像子区域的动态表示。它的计算是对多个区域视觉向量的加权和，即：

$$\boldsymbol{c}_i=\sum_{j=0}^{288}\alpha_j\boldsymbol{v}_j \tag{10}$$

式中：$\alpha_j=\dfrac{\exp(\gamma_1\bar{\boldsymbol{s}}_i,j)}{\sum_{k=0}^{288}\exp(\gamma_1\bar{\boldsymbol{s}}_i,k)}$；$\gamma_1$ 为权重参数，它决定了在计算单词的区域上下文向量时，对其相关子区域特征的关注程度。最后计算 $\boldsymbol{c}_i$ 和 $\boldsymbol{e}_i$ 之间的余弦相似度来定义第 i 单词与图像之间的相关性。即 $R(\boldsymbol{c}_i,\boldsymbol{e}_i)=(\boldsymbol{c}_i^{\mathrm{T}}\boldsymbol{e}_i)/(\|\boldsymbol{c}_i\|\|\boldsymbol{e}_i\|)$。整个图像（$Q$）与整个文本描述（$D$）之间的注意力驱动图像-文本匹配得分定义为：

$$R(Q,D)=\log\Big(\sum_{i=1}^{T-1}\exp(\gamma_2R(c_i,e_i))\Big)^{\frac{1}{\gamma_2}} \tag{11}$$

式中：γ_2 为权重参数，用来决定最相关单词与区域上下文之间的重要性。当 $\gamma_2\to\infty$，$R(Q,D)$ 近似于 $\max_{i=1}^{T-1}R(\boldsymbol{c}_i,\boldsymbol{e}_i)$。

2 损失函数

为了生成具有多级条件（即句子级、单词级、语义分割图特征级）的高质量图像，模型最终的损失函数被定义为：

$$\mathcal{L}=\mathcal{L}_{\mathrm{G}}+\lambda\mathcal{L}_{\mathrm{MSCM}} \tag{12}$$

式中：$\mathcal{L}_{\mathrm{G}}=\sum_{i=0}^{m-1}\mathcal{L}_{G_i}$，$\mathcal{L}_{\mathrm{MSCM}}=\mathcal{L}_1^{\mathrm{w}}+\mathcal{L}_2^{\mathrm{w}}+\mathcal{L}_1^{\mathrm{s}}+\mathcal{L}_2^{\mathrm{s}}$；$\lambda$ 为一个用于平衡式的 2 项的超参数；$\mathcal{L}_{\mathrm{G}}$ 为 GAN 损失，它来联合逼近条件分布和无条件分布；$\mathcal{L}_{\mathrm{MSCM}}$ 为多模态相似性模型损失，它通过计算细粒度图像-文本匹配损失以训练生成器。在改进模型的第 i 个阶段，有一个鉴别器 D_i 与生成器 G_i 相对应。G_i 的对抗损失定为：

$$\underbrace{\mathcal{L}_{G_i}=-\frac{1}{2}E_{\boldsymbol{x}_i\sim\boldsymbol{p}_{G_i}}[\log(D_i(\dot{\boldsymbol{x}}_i))]}_{\text{无条件损失}}-\underbrace{\frac{1}{2}E_{\boldsymbol{x}_i\sim\boldsymbol{p}_{G_i}}[\log(D_i(\dot{\boldsymbol{x}}_i,\bar{\boldsymbol{e}},\boldsymbol{I}_s))]}_{\text{条件损失}} \tag{13}$$

式中：$\boldsymbol{I}_s$ 为语义分割图特征向量；$\bar{\boldsymbol{e}}\in\boldsymbol{R}^D$ 为双向 LSTM 的最后一个隐状态连接而成的全局句子向量。交替训练 G_i 和鉴别器 D_i，并通过最小化定义的交叉熵损失函数将输入分类为真或假类别，定义的交叉熵函数为：

$$\begin{aligned}\mathcal{L}_{D_i}=&\underbrace{-\frac{1}{2}E_{\boldsymbol{x}_i\sim\boldsymbol{p}_{\mathrm{data}_i}}[\log D_i(\boldsymbol{x}_i)]-\frac{1}{2}E_{\dot{\boldsymbol{x}}_i\sim\boldsymbol{p}_{G_i}}[\log(1-D_i(\dot{\boldsymbol{x}}_i)]}_{\text{无条件损失}}+\\&\underbrace{-\frac{1}{2}E_{\boldsymbol{x}_i\sim\boldsymbol{p}_{\mathrm{data}_i}}[\log D_i(\boldsymbol{x}_i,\bar{\boldsymbol{e}},\boldsymbol{I}_s)]-\frac{1}{2}E_{\dot{\boldsymbol{x}}_i\sim\boldsymbol{p}_{G_i}}[\log(1-D_i(\dot{\boldsymbol{x}}_i,\bar{\boldsymbol{e}},\boldsymbol{I}_s)]}_{\text{条件损失}}\end{aligned} \tag{14}$$

式中：$\boldsymbol{x}_i$ 为第 i 个尺度的真实图像分布 $\boldsymbol{p}_{\mathrm{data}_i}$；$\dot{\boldsymbol{x}}_i$ 为相同尺度模型分布 $\boldsymbol{p}_{G_i}$。用无条件损失和条件损失判断图像的真假和图像与句子是否匹配。模型的鉴别器可进行并行训练，因为它们在结构上是不相交的，并且都集中在单个图像尺度上。多模态相似度计算模型的损失函数设计以半监督的方式学习注意力模型，其中唯一的监督是整个图像和整个句子（单词顺序）之间的匹配。对于一批图像-句子对 $\{(Q_i,D_i)\}_{i=1}^{M}$，计算句子 D_i 与图像 Q_i 匹配的后验概率为：

$$P(D_i\mid Q_i)=\frac{\exp(\gamma_3R(Q_i,D_i))}{\sum_{j=1}^{M}\exp(\gamma_3R(Q_i,D_j))} \tag{15}$$

式中：γ_3 为实验的平滑因子。在这批句子中，只有 D_i 匹配了图像 Q_i，其余的都是不匹配的描述。模型将损失函数定义为图像与其对应的文本描述匹配的负对数后验概率，即：

$$\mathcal{L}_1^{\mathrm{w}}=-\sum_{i=1}^{M}\log P(D_i\mid Q_i) \tag{16}$$

式中：w 表示单词，同理可得 $\mathcal{L}_2^{\mathrm{w}}$ 和 $\mathcal{L}_3^{\mathrm{w}}$。

3 实验结果

3.1 数据集及评价指标

本文使用 Oxford-102 Flowers[20] 和 Caltech UCSD Birds-200（CUB-200）[21] 数据集以及它们对应的语义分割图像集进行实验和验证分析，故 2 个数据集的训练集和测试集的图像数量都为原来的 2 倍，实验所使用的数据集具体如表 1 所示。

表 1 实验数据集

参数	Oxford-102 Flowers		CUB-200	
	训练集	测试集	训练集	测试集
图像数量	7 034×2	1 155×2	8 855×2	2 933×2
文本描述数量/图像	10	10	10	10

定量实验采用 IS(inception score)[20] 和 FID (frechet inception distance)[21]2 种指标对生成的结果图进行客观评价。IS 指标用于衡量生成图像的质量和多样性，通常 IS 值越高越好。FID 指标通过计算生成分布与真实分布之间的特征距离，来衡量生成的图像与真实的图像之间的接近程度，FID 值越低，则表示 2 个分布间的距离越接近，生成的图像就越接近真实样本。

3.2　实验环境及参数设置

实验环境如表 2 所示。实验的超参数设定为：$\gamma_1=5$，$\gamma_2=5$，$\gamma_3=10$，$M=50$，batch size 为 64，epoch 为 600，生成器和鉴别器的初始学习速率均为 0.000 2，平衡参数 $\lambda=5$。优化器为 Adam，参数 $\beta_1=0.5$，$\beta_2=0.99$。使用 Oxford-102 Flowers 和 CUB-200 以及它们对应的语义分割数据集，在表 2 的实验环境下，对生成器和鉴别器进行训练，训练时长为 7 d。

表 2　实验环境

名称	具体信息
操作系统	Ubuntu20.04.5 LTS 64-bit
内存	24 G
GPU	NVIDIA RTX3090 (1 块)
CPU	i9-10900x
开发工具	Pytorch1.9.1；python3.6；pycharm

3.3　注意力机制作用

图 3 为注意力作用效果图，图中第 1 行字体是图像的具体文本描述，其中红色字体是图像在生成过程中关注度最高的前 5 个单词。第 1 行图片分别是模型生成的 64×64、128×128、256×256 分辨率的图像，第 2 行图像是注意力机制重点关注的图像细节。从图中可看出 64×64 分辨率的图像虽然形状结构完整，但细节不够清晰，比如鸟身体的花纹和眼部细节都比较模糊。在注意力模型的作用下，模型抓住文本描述的重点语义信息从单词级别的水平逐渐细化生成的图像，最终生成了具有丰富细节的高质量图像。

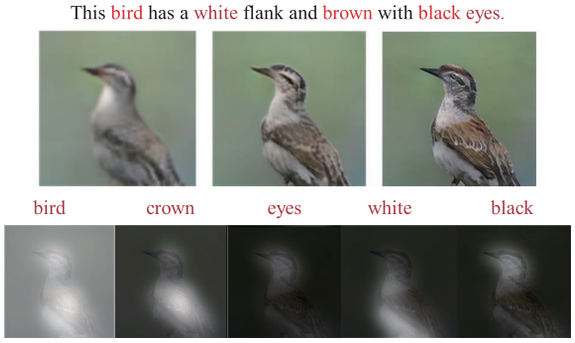

图 3　SSA-GAN 的注意力机制作用效果

3.4　结合语义分割图的作用

为了验证所提方法的有效性，对 SSA-GAN 生成的低分辨率(64×64)图像进行可视化，并与其对应的语义分割图像进行对比分析，对比结果如图 4 所示。图中第 1 行对应于原图的语义分割图像，第 2 行是在初阶段结合了语义分割图像生成的低分辨率(64×64)图像，由可视化结果可以看出，SSA-GAN 很好地勾勒出了鸟和花的形状，基本生成了与约束形状相一致的图像，保证了生成图像边缘结构的完整。

文中的对比实验选取经典的 AttnGAN 和当前受欢迎的 DF-GAN、RAT-GAN 网络模型，在 CUB-200、Oxford-102 Flowers 数据集上对生成的低分辨率图像(64×64)和高分辨率(256×256)图像分别进行对比分析。因为 DF-GAN、RAT-GAN 网络模型为单阶段式生成模型，故其原码中并没有生成 64×64 的低分辨率图像，而是直接生成了 256×256 的高分辨率图像。为了方便对比，在 DF-GAN、RAT-GAN 网络模型源码中添加了 64×64 低分辨率图像的可视化代码，需要说明的是这并不会影响原始模型生成的高分辨率图像的质量。高分辨率图像和低分辨率图像对比结果分别如图 5 和图 6 所示。由图 5 生成的结果可以看出，所有模型都在一定程度上生成了与文本描述和真实图像基本符合的鸟和花，但不同模型之间生成结果具有一定的差异。AttnGAN、DF-GAN 和 RAT-GAN 模型生成的部分结果存在结构缺失、内容不真实的现象。例如图 5 中第 2 列生成的鸟，AttnGAN 生成的结果发生了严重畸变，已经不像一只鸟。DF-GAN 生成的结果虽然整体结构较为完整，但是头部细节过于模糊，甚至无法辨认鸟的眼睛所在位置。RAT-GAN 生成的结果细节相对丰富，但存在严重的结构缺失，生成的鸟没有了尾巴，且脚部生成不够真实。RAT-GAN 在第 3 列生成的结果更是不符合实际，鸟直接停留在悬在空中的不明物体之上。AttnGAN 在最后一列生成的花出现了高度模糊，而 DF-GAN、RAT-GAN 则能生成结构相对完整、轮廓清晰的花图像。但相比之下，SSA-GAN 生成的结果结构更加完整，细节也更为丰富。

高分辨率图像是在低分辨率图像的基础上弥补低分辨率图像的不足，生成的具有更加丰富细节的高质量图像。分析图 5 和图 6 中的生成结果可知，低分辨率图像出现的模糊问题，在生成的高分辨率图像中有所改善。但是，低分辨图像中存在的结构缺失问题，在生成的高分辨率图像中仍然无法得到有效解决。例如图 6 中第 2 列的生成结果依然存在结构缺失的问题。由此可知，生成结构完整、质量较好的低分辨率图像尤为重要。相较于 AttnGAN、DF-GAN 和 RAT-GAN 模型，SSA-GAN 生成的高分辨率图像的结构更加完整、图像细节更加丰富。图 7 为 SSA-GAN 生成的部分图像。

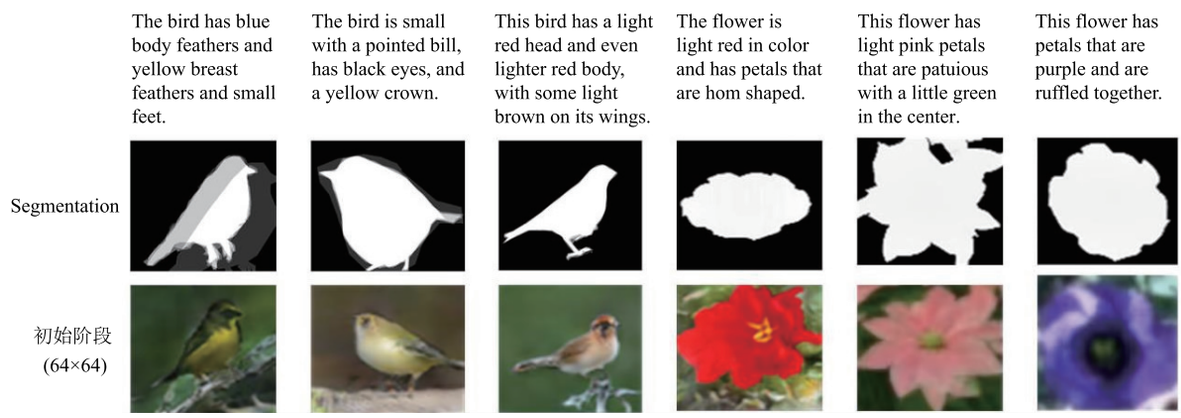

图 4　SSA-GAN 在不同数据集上结合语义分割图生成的结果

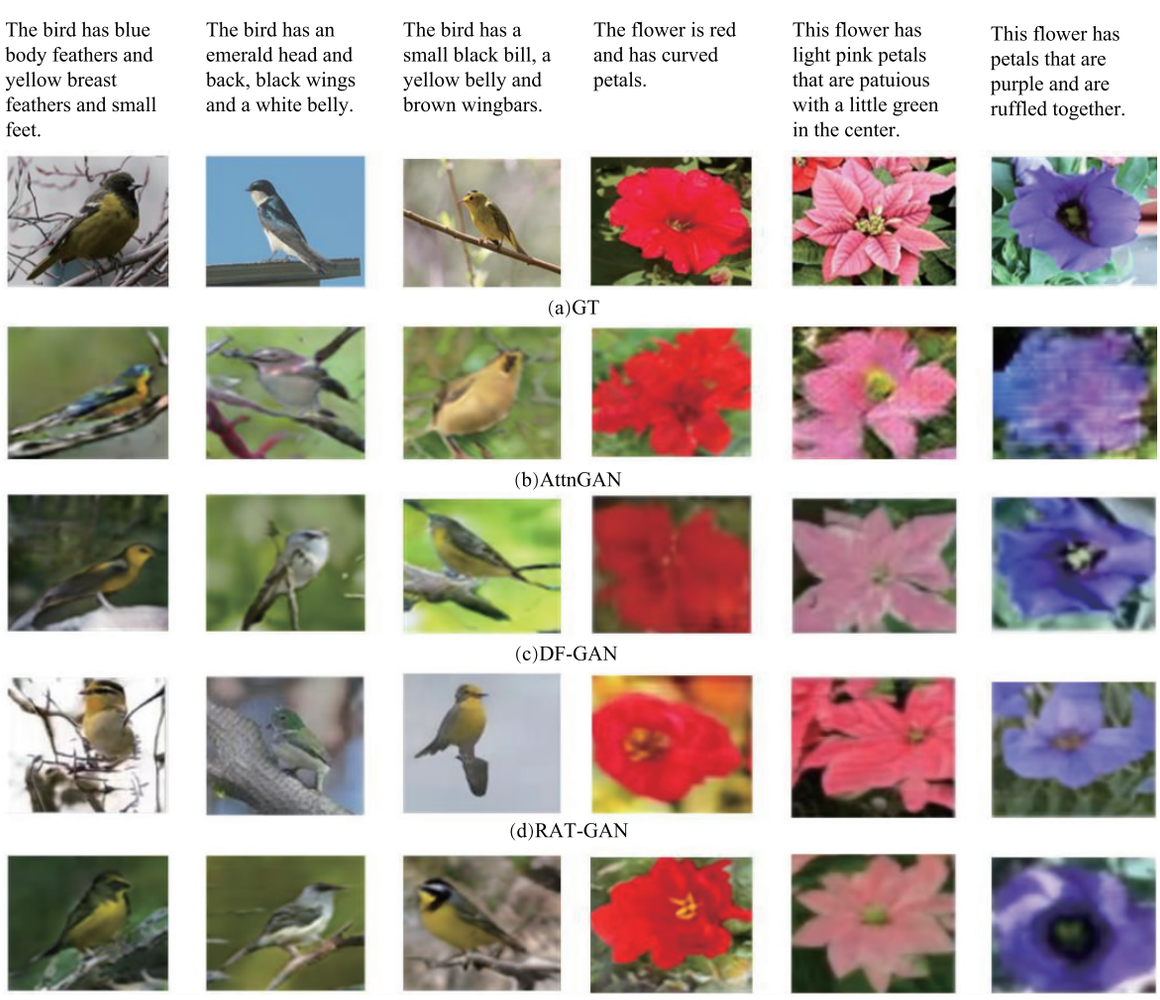

图 5　不同模型生成的低分辨率(64×64)图像结果对比

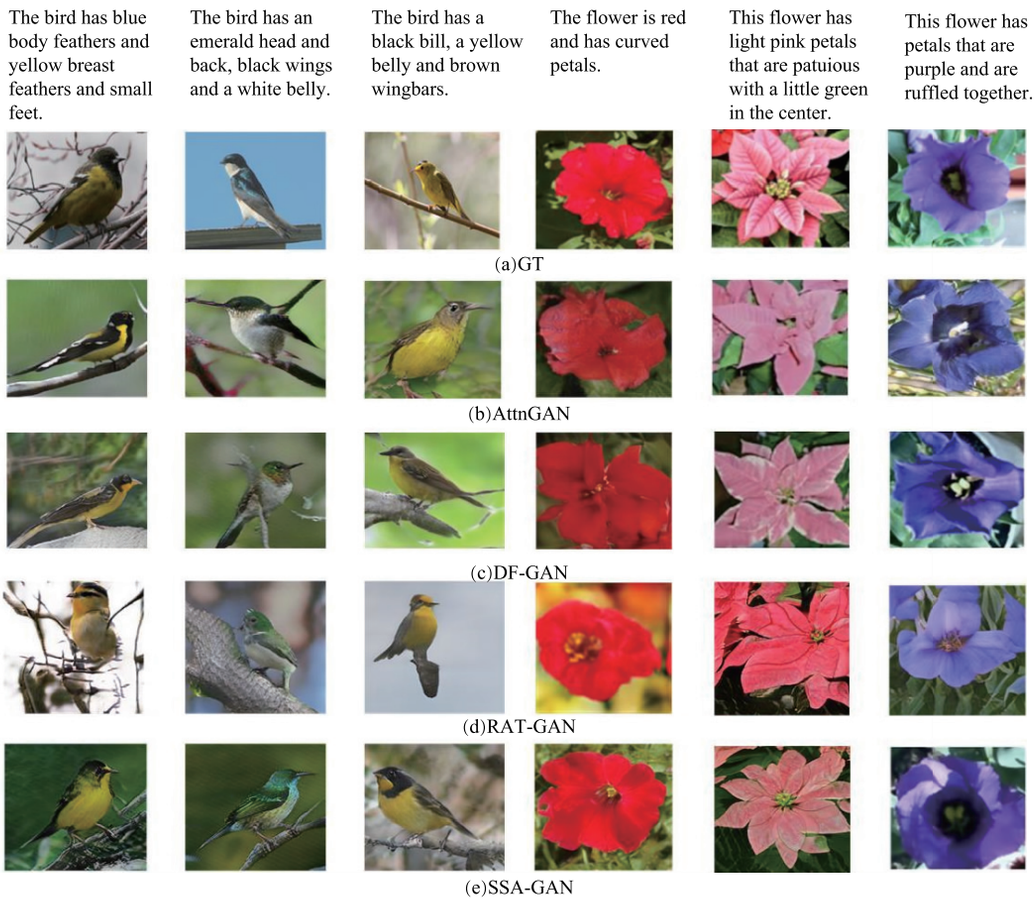

图 6　不同模型生成的高分辨率(256×256)图像结果对比

图 7　SSA-GAN 生成的部分高分辨率(256×256)图像

3.5　定量结果比较

表 3 和表 4 分别为不同模型生成的低分辨率和高分辨率图像的定量结果。从表 3 和表 4 可知，相较于所对比的模型，SSA-GAN 在 Oxford-102 Flowers、CUB-200 数据集上均取得了最优的 IS 和 FID 指标值。对于生成的低分辨率图像，由表 3 中的数据计算可知，SSA-GAN 在 CUB-200 数据集上相较于 StackGAN 模型，IS 指标值提升了 44.2%，FID 指标值降低了 75.3%，在 Oxford-102 Flowers 数据集上相较于 AttnGAN 模型，IS 指标值提升了 53.3%，FID 指标值降低了 35%。

对于生成的高分辨率图像，由表 4 中的数据计

算可知，SSA-GAN 在 CUB-200 数据集上相较于 StackGAN 模型，IS 指标值提升了 43.2%，FID 指标值降低了 74.9%，在 Oxford-102 Flowers 数据集上相较于 AttnGAN 模型，IS 指标值提升了 13.7%，FID 指标值降低了 34.7%。

表 3 低分辨率(64×64)图像的 IS 和 FID 指标值变化

模型	Oxford-102 Flowers		CUB-200	
	IS	FID	IS	FID
StackGAN[12]			2.65	57.20
AttnGAN[13]	3.19	24.89	3.21	24.22
DF-GAN[15]	3.21	17.23	3.46	18.60
RAT-GAN[16]	3.32	16.21	3.67	14.18
SSA-GAN	3.36	16.18	3.82	14.12

表 4 高分辨率(256×256)图像的 IS 和 FID 指标值变化

模型	Oxford-102 Flowers		CUB-200	
	IS	FID	IS	FID
StackGAN[12]			3.70	55.58
AttnGAN[13]	3.57	24.65	4.36	23.98
MirrorGAN[14]			4.56	18.34
AttnGANCL[24]	3.75	24.35	4.42	16.32
DM-GAN[25]			4.75	16.09
DF-GAN[15]	3.80	17.15	4.86	14.81
RAT-GAN[16]	4.02	16.14	5.28	13.98
SSA-GAN	4.06	16.08	5.30	13.94

3.6 消融实验

利用消融实验进一步验证所提方法的有效性。实验将不使用结合语义分割图方法(Se)、多模态相似度计算模型(MSCM)、注意力模型(At)和每个阶段只使用一个深度融合模块(即 $N=1$)作为基准模型(Baseline)。实验对 Baseline 和 Baseline＋Se 以及 Baseline＋Se 且 N 分别为 3、5、8 的配置下，在 Oxford-102 和 CUB-200 数据集上计算所生成高分辨率(256×256)图像的 IS 和 FID 指标值，实验结果如表 5 所示。

从表 5 所示的实验结果可知：①Baseline＋Se、Baseline＋At、Baseline＋ MSCM 生成结果的各项指标值都优于 Baseline，且 Baseline＋Se＋MSCM＋At 生成结果的指标值要明显优于 Baseline 和 Baseline＋Se、Baseline＋At、Baseline＋ MSCM，证明了本文所提方法的有效性。②随着 N 值的增大，IS 和 FID 指标值先逐渐变好后逐渐变差，最终取得最优效果的配置为 Baseline＋Se，$N=3$。

表 5 SSA-GAN 在 Oxford-102 和 CUB-200 数据集上的消融实验结果

方法	Oxford-102		CUB-200	
	IS	FID	IS	FID
Baseline	3.26	24.54	3.25	23.91
Baseline＋Se	3.45	23.62	3.41	22.38
Baseline＋At	3.52	22.57	3.65	21.87
Baseline＋MSCM	3.42	21.42	3.38	20.79
Baseline＋Se＋MSCM＋At	3.76	18.85	4.53	15.32
Baseline＋Se＋MSCM＋At，$N=3$	4.06	16.08	5.30	13.94
Baseline＋Se＋MSCM＋At，$N=5$	3.82	17.34	5.18	14.69
Baseline＋Se＋MSCM＋At，$N=8$	3.78	18.54	5.02	14.94

4 结语

本文提出结合语义分割图的方法，为模型提供了额外的生成和约束条件，有效地改善了生成图像存在的结构缺失和内容不真实的问题。提出的深度融合模块(DFMBlock)，使生成器在融合文本和图像特征的同时，充分保留了文本信息，加强了文本与图像的融合。注意力模型为生成器提供了细粒度词级信息，有效地丰富了生成图像的细节，改善了生成图像的质量不佳问题。多模态相似度计算模型(MSCM)为训练生成器提供了图像-文本匹配损失，使生成器更易于训练。通过 Oxford-102 和 CUB-200 数据集验证了所提方法的有效性。虽然本文所提出的模型(SSA-GAN)进一步改善了生成图像的结构不完整、内容不真实、质量不佳的问题，但仍存在模型占用计算资源较大等不足，在后续的工作中，需进一步改善。

参考文献

[1] 曹寅，秦俊平，高彤，等. 基于生成对抗网络的文本两阶段生成高质量图像方法[J]. 浙江大学学报(工学版)，2024，58(4)：674-683.

[2] 吴春燕，潘龙越，杨有，等. 基于特征增强生成对抗网络的文本生成图像方法[J]. 微电子学与计算机，2023，40(6)：51-61.

[3] 李云红，段姣姣，苏雪平，等. 基于改进生成对抗网络的书法字生成算法[J]. 浙江大学学报(工学版)，2023，57(7)：1326-1334，1459.

[4] TAO M，BAO B K，TANG H，et al. GALIP：Generative Adversarial CLIPs for Text-to-Image Synthesis[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver，BC，Canada：IEEE，2023：14214-14223.

[5] 刘昊，杨小汕，徐常胜. 基于动态语义记忆网络的长

尾图像描述生成 [J]. 北京航空航天大学学报，2022，48(8)：1399-1408.

[6] DING M，YANG Z，HONG W，et al. CogView：Mastering Text-to-Image Generation via Transformers [C]//Proceedings of the Advances in Neural Information Processing Systems. La Jolla：NeurIPS Foundation，2021：19822-19835.

[7] ZHANG H，KOH J Y，BALDRIDGE J，et al. Cross-Modal Contrastive Learning for Text-to-Image Generation [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville，TN：IEEE，2021：833-842.

[8] LI Y H，LIU H T，WU Q Y，et al. GLIGEN：Open-Set Grounded Text-to-Image Generation [C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver，BC：IEEE Press，2023：22511-22521.

[9] GOODFELLOW I，POUGET-ABADIE J，MIRZA M，et al. Generative Adversarial Nets [C]// Proceedings of the 27th International Conference on Neural Information Processing Systems. New York：ACM，2014：2672-2680.

[10] 赖丽娜，米瑜，周龙龙，等. 生成对抗网络与文本图像生成方法综述 [J]. 计算机工程与应用，2023，59(19)：21-39.

[11] REED S，AKATA Z，MOHAN S，et al. Learning What and Where to Draw [C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona：ACM，2016：217-225.

[12] ZHANG H，XU T，LI H S，et al. StackGAN：Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks [C]//2017 IEEE International Conference on Computer Vision (ICCV). Venice：IEEE，2017：5908-5916.

[13] XU T，ZHANG P C，HUANG Q Y，et al. AttnGAN：Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City，UT：IEEE，2018：1316-1324.

[14] QIAO T T，ZHANG J，XU D Q，et al. MirrorGAN：Learning Text-to-Image Generation by Redescription [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach，CA：IEEE，2019：1505-1514.

[15] TAO M，TANG H，WU F，et al. DF-GAN：a Simple and Effective Baseline for Text-to-Image Synthesis [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans，LA：IEEE，2022：16494-16504.

[16] YE S M，WANG H，TAN M K，et al. Recurrent Affine Transformation for Text-to-Image Synthesis [J]. IEEE Transactions on Multimedia，2024，26：462-473.

[17] KANG M，ZHU J Y，ZHANG R，et al. Scaling up GANs for Text-to-Image Synthesis [C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver，BC：IEEE，2023：10124-10134.

[18] RUSSAKOVSKY O，DENG J，SU H，et al. ImageNet Large Scale Visual Recognition Challenge [J]. International Journal of Computer Vision，2015，115(3)：211-252.

[19] SZEGEDY C，VANHOUCKE V，IOFFE S，et al. Rethinking the Inception Architecture for Computer Vision [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas，NV：IEEE，2016：2818-2826.

[20] NILSBACK M E，ZISSERMAN A. Automated Flower Classification over a Large Number of Classes [C]//2008 Sixth Indian Conference on Computer Vision，Graphics & Image Processing. Bhubaneswar：IEEE，2008：722-729.

[21] WAH C，WSH C，BRANSON S，et al. The Caltech-UCSD Birds-200-2011 Dataset [R]. Springfield：U. S. California Institute of Technology，2011.

[22] SALIMANS T，GOODFELLOW I，ZAREMBA W，et al. Improved Techniques for Training GANs [C]// Proceedings of the 30th International Conference on Neural Information Processing Systems. December 5 - 10，2016，Barcelona：ACM，2016：2234-2242.

[23] HEUSEL M，RAMSAUER H，UNTERTHINER T，et al. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium [C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach，California：ACM，2017：6629-6640.

[24] YE H，YANG X，TAKAC M，et al. Improving Text-to-Image Synthesis Using Contrastive Learning [DB/OL]. ArXiv Preprint：2107.02423，2021.

[25] ZHU M F，PAN P B，CHEN W，et al. DM-GAN：Dynamic Memory Generative Adversarial Networks for Text-to-Image Synthesis [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach，CA：IEEE，2019：5795-5803.

（编辑：徐楠楠）