融合 Transformer 与 DF-GAN 的文本生成图像方法

马静[1]，车进[1,2]，孙末贤[1]

（1.宁夏大学电子与电气工程学院，宁夏 银川 750000；2.宁夏大学研究生院，宁夏 银川 750000）

摘　要：文本生成图像任务中的文本编码器不能深度挖掘文本信息，导致后续生成的图像语义不一致。针对该问题，提出一种 DXC-GAN 文本生成图像方法。引入 Transformer 系列中的 XLNet（Xtra Long Network）预训练模型替换原始文本编码器，捕获大量文本的先验知识，实现对上下文信息的深度挖掘。添加 CBAM（Convolutional Block Attention Module）注意力模块，使生成器更加关注图像中的重要信息，从而解决生成图像细节不完整和空间结构错误问题。在判别器中引入对比损失，与模型中匹配感知梯度惩罚和单向输出结合，使得相同语义图像之间更加接近，不同语义图像之间更加疏远，从而增强文本与生成图像之间的语义一致性。实验结果表明：与 DF-GAN 相对比，DXC-GAN 在 CUB 数据集上的 IS（Inception Score）与 FID（Fréchet Inception Distance）分别提升了 4.42%和 17.96%；在 Oxford-102 数据集上，IS 为 3.97，FID 为 37.82；相较于 DF-GAN，DXC-GAN 在鸟类图像生成方面有效避免了多头少脚等畸形问题，同时在花卉图像生成上也显著减少了花瓣残缺等图像质量问题；此外，DXC-GAN 还增强了文本与图像的对齐性，显著提升了图像的完整度和生成效果。

关键词：生成对抗网络；文本生成图像；XLNet；CBAM；对比损失

中图分类号：TP18　　**文献标志码：**A　　**DOI：**10.19678/j.issn.1000-3428.0069611

Text-to-Image Generation Method Combining Transformer and DF-GAN

MA Jing[1], CHE Jin[1,2], SUN Moxian[1]

(1.School of Electronic and Electrical Engineering, Ningxia University, Yinchuan 750000, Ningxia, China;
2.Graduate School, Ningxia University, Yinchuan 750000, Ningxia, China)

【Abstract】To address the failure of the text encoder to deeply mine text information in text-to-image generation tasks, which leads to semantic inconsistency in the subsequently generated images, a DXC-GAN method for text-to-image generation is proposed. This method introduces the Xtra Long Network (XLNet) pretraining model from the Transformer series to replace the original text encoder, enabling the capture of prior knowledge from a vast amount of text for deep mining of contextual information. A Convolutional Block Attention Module (CBAM) is added to increase the generator's focus on important information in images, thus solving the issues of incomplete image details and incorrect spatial structure. In the discriminator, contrastive loss is introduced and combined with match-aware gradient penalty and unidirectional output in the model, making images with the same semantics closer and those with different semantics further apart, thereby enhancing the semantic consistency between text and generated images. The experimental results show that compared to the DF-GAN model, the Inception Score (IS) and Fréchet Inception Distance (FID) on the CUB dataset for the proposed model improved by 4.42% and 17.96%, respectively. On the Oxford-102 dataset, the IS is 3.97 and the FID is 37.82. Evidently, compared to DF-GAN, DXC-GAN effectively avoids deformities such as multi-headedness and foot deficiency in bird image generation and significantly reduces image quality issues such as missing petals in flower image generation. Furthermore, it enhances the alignment between text and images, significantly improving the completeness and generation effect of images.

【Key words】Generative Adversarial Network (GAN); text-to-image generation; XLNet; CBAM; contrastive loss

0　引言

将文本转化为图像是一项涉及自然语言处理（NLP）和计算机视觉的跨模态交叉训练任务[1-2]，其目的是从特定文本描述中生成高质量的图像。目前，文本生成图像技术已经取得了显著的进展，但传统的方法通常依赖于基于规则的图像合成技术，这些方法在生成内容的真实感和多样性方面存在一定的局限性。近年来，生成对抗网络（GAN）[3]作为一种强大的深度学习框架，被广泛应用于图像转换、视觉推理和图像生成等领域，并取得了令人瞩目的成果。

基金项目：国家自然科学基金（62366042）。

作者简介：马静，女，硕士研究生，主研方向为文本生成图像；车进（通信作者），教授、博士；孙末贤，硕士研究生。

收稿日期：2024-03-18　　**修回日期：**2024-08-29　**E-mail：**1581005897@qq.com

文本生成图像利用 GAN 模型，通过文本编码器将文本编码成图像，以生成具有语义一致性的逼真图像。尽管 GAN 在文本生成图像方面取得了一些成功，但由于任务复杂性和对语义一致性与多样性的要求，当前的方法仍存在一些问题，如生成图像模糊、低分辨率、缺乏真实感、细节不够丰富、文本描述与图像之间的语义一致性较差等。因此，如何利用 GAN 来克服这些问题，生成更真实、高分辨率、语义一致和多样化的图像，是文本生成图像任务面临的挑战。

针对以上问题，本文提出 DXC-GAN 文本生成图像方法。在 DF-GAN[4] 中采用 XLNet（Xtra Long Network）[5] 文本编码器替换长短时记忆网络（LSTM）[6] 文本编码器对文本的上下文信息进行深度挖掘，更好地理解文本之间的关联和语义，提高文本编码的效率和准确性，改善生成结果的语义一致性。在图像生成阶段引入 CBAM（Convolutional Block Attention Module）[7] 注意力机制，动态地调整图像中不同位置的特征权重，从而增强模型对重要信息的关注，抑制对无关信息的响应。另外，在判别器中加入对比学习[8] 方法，增强对抗训练中判别器对真实样本和生成样本的区分能力，从而改善 GAN 的性能，提高生成图像的质量。

1 相关工作

1.1 基于 GAN 的文本生成图像方法

文本生成图像方法具有广泛的应用基础，在自然图像领域取得了重大进展。近些年，基于 GAN 的文本生成图像方法在提高任务性能方面取得了重要进展。REED 等[9] 提出的 GAN-INT-CLS 方法首次引入对抗过程来生成图像，但其限制在生成 64×64 像素的图像，且图像质量较差。为了生成高质量图像，ZHANG 等[10] 提出多阶段模型 StackGAN。该模型的第一阶段生成低分辨率图像，第二阶段修复这些图像以生成高分辨率图像。随后，他们又在 StackGAN 的基础上提出了一种新的端到端模型 StackGAN++[11]，增加了更多的生成阶段来提高模型的收敛速度和生成图像的质量。然而，该模型生成的图像仍然存在纹理模糊、局部细节丢失等问题，文本与图像之间的关联度仍有待改善。为了解决这些问题，XU 等[12] 提出了 AttnGAN，在 StackGAN 基础上采用注意力机制和深度注意多模态相似性模型（DAMSM），以生成更符合语句描述的图像。MirrorGAN[13] 则是通过重述生成图像、加入重建损失增强语义一致性。ZHU 等[14] 提出 DM-GAN，主要是在 AttnGAN 基础上引入动态记忆存储模块，并专注于捕捉字词特征，通过这项技术，模型能够生成更加生动、精彩的图像内容。尽管堆叠式 GAN 优化了训练不稳定的问题，但产生了生成器间的交织问题，使得生成图像看起来既模糊又显现出局部细节。

为解决上述问题，DF-GAN 采用了简化结构，仅使用一个生成器和一个判别器进行训练。它能够生成高分辨率图像，并通过匹配感知梯度惩罚以及单向输出来提升文本到图像的语义匹配，无需额外的网络结构。然而，该模型仅考虑了句子级别的信息，缺乏细致合成视觉特征的能力。为了解决这个问题，LIAO 等[15] 提出了 SSA-GAN，使用语义-空间批量归一化模块、残差块和掩码预测器，深度融合文本和视觉特征以优化生成器的图像合成。然而，该模型在细节方面仍存在一些问题，如生成图像缺少主体必要特征、边缘模糊等。

1.2 Transformer

Transformer 是一个以自注意力机制为核心的神经网络模型[16]，最早被应用于机器翻译领域，取得显著成功。相比传统的循环神经网络[17]，Transformer 通过并行处理和处理长距离依赖，提高了训练速度，它已成为深度学习中的经典模型，在 NLP 和计算机视觉等领域广泛应用。流行的基于 Transformer 的 NLP 模型有基于 Transformers 的双向编码器表征（BERT）[18-19]、生成式预训练 Transformer（GPT）[20] 和 XLNet[5]。BERT 采用了双向上下文理解，并通过大规模无监督预训练提升了在多种任务上的性能，但预训练和微调需要大量计算资源，且对长文本处理有局限。GPT 以自回归方式表现出色于文本生成，尤其在连贯长文本方面，但作为单向模型可能对某些任务表现不佳，并需要大量训练数据。XLNet 结合了前两种方法，它既支持双向编码又能自回归生成，在 20 个任务上超越了 BERT，并在 18 个任务中刷新了纪录，包括了情感分析、机器问答和自然语言推断等多个应用领域。

1.3 注意力机制

注意力机制源自人类对外部信息的处理能力。由于人脑无法同时处理大量复杂的信息，人们会选择性地关注需要关注的信息，过滤无关信息。常用的注意力机制有空间注意力机制（SAM）、通道注意力机制（CAM）和混合注意力机制（CBAM）[7]。CAM 关注图像中通道之间的信息，SAM 关注图像中位置信息，CBAM 是由 CAM 和 SAM 组成的模

型，通过顺序注意力结构将注意力从通道扩展到空间。空间注意力使网络更关注对分类起决定作用的像素区域，而通道注意力则处理特征图通道的分配关系。综合考虑这两种注意力分配的方式可以提升模型的性能。

1.4 对比学习

对比学习是一种自监督学习[21]方法，早期在计算机视觉和图形学习中应用[22]。基于记忆库机制的对比学习架构被采用，其核心思想是通过最大化相似样本之间的距离、最小化不相似样本之间的距离来学习特征表示。将每个样本表示为向量，并使用神经网络模型将其映射到特征空间。对于每对样本，计算它们在特征空间的距离，并根据它们的标签（相似或不相似）调整模型参数。相似样本对可以使用欧氏距离或余弦相似度等进行衡量，目标是最小化它们的距离。对于不相似样本对，目标则是最大化它们的距离。

2 网络模型

2.1 模型架构

本文以 DF-GAN 为基础架构进行优化改进，改进后的网络结构如图 1 所示（彩色效果见《计算机工程》官网 HTML 版，下同）。基于 Transformer 架构构建文本编码器，将其原有的 LSTM 编码器替换为 XLNet 文本编码器，从而更好地捕捉文本信息；在卷积层前添加 CBAM 模块，同时对两个维度进行注意力分配，提高图像分辨率；通过在判别器中引入对比损失模块，使得相同图像对之间的距离越来越近，不同图像对之间的距离越来越远，从而增强模型语义一致性。

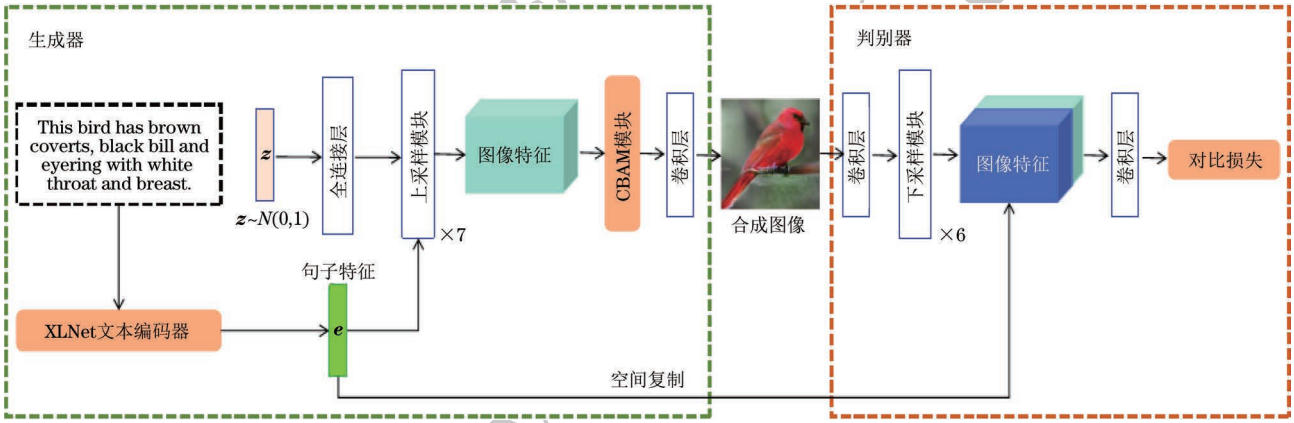

图 1 网络框架图

Fig. 1 Network framework diagram

2.2 XLNet 文本编码器

原有模型中使用 LSTM 文本编码器，该文本编码器在处理序列数据时表现出色，但不能很好地理解文本语境和全局依赖关系。因此，本文提出一种基于 XLNet 模型的文本编码器，利用自回归预训练的方法，更好地捕捉文本中的全局依赖关系，深度挖掘文本信息。加入 XLNet 模型后的文本编码器流程图如图 2 所示，整体架构由输入、文本预处理、XLNet 预训练编码器和输出 4 个部分组成。首先，在 PyTorch 中加载 XLNet 模型，使用 Hugging Face 的 Transformers 库，引入 XLNetTokenizer 和 XLNetModel 类。XLNetTokenizer 负责对 XLNet 模型进行分词，XLNetModel 则是 PyTorch 提供的 XLNet 模型网络结构。然后，使用 from_pretrained 方法加载 XLNetTokenizer 和 XLNetModel，以存储模型词汇表，提供 token embedding。

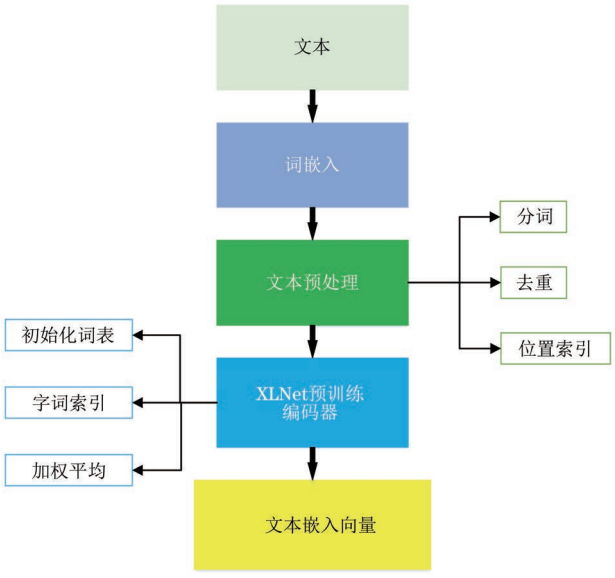

图 2 XLNet 文本编码器流程图

Fig. 2 XLNet text encoder flowchart

XLNet 是一种基于 Transformer 架构的自回归预训练模型，它结合了自回归和自编码两种方法，能够提高语言建模的性能。在 XLNet 中，文本编码器的独特设计主要体现在两个方面：自回归性和双流注

意力机制。类似于 GPT 模型，XLNet 也是一个自回归的模型。这意味着在训练过程中，模型会根据输入序列中前面的标记来预测下一个标记。不同之处在于，XLNet 采用了一种称为“双流自回归机制”的方法。这种机制允许模型同时利用自回归和自编码两种方法，以更好地捕捉上下文信息，它还使用了一种称为“双流注意力机制”的设计，在传统的 Transformer 中，只有单个注意力流，即每个位置的输出都是由整个输入序列的所有位置计算得到的。而在 XLNet 中，每个位置都有两个注意力流：自回归流和自编码流。这意味着每个位置的输出都同时考虑了前面的上下文和后面的上下文，从而提高了建模效果。

在复现 XLNet 模型时，需要注意的参数设置主要包括自回归参数、双流注意力机制参数、超参数调整。在训练时，需要指定模型应该采用多长的上下文来预测下一个标记。这通常通过调整模型的最大位置编码来实现。同时，需要注意训练过程中的截断问题，即如何处理过长的序列以适应模型的内存限制。在模型的初始化和训练过程中，需要确保双流注意力机制被正确实现。这涉及到在模型的注意力层中正确配置两个注意力流。超参数调整和其他深度学习模型一样。XLNet 的性能也受到许多超参数的影响，如学习速率、批量大小、层数、隐藏单元数等。在复现时，需要进行适当的超参数调整以获得最佳性能。XLNet 的预训练任务通常是掩码语言建模和下一个句子预测。在复现时，需要确保正确设置这些任务的参数，并使用适当的数据集进行训练。

2.3 CBAM 注意力机制

针对传统卷积神经网络在处理不同尺度、形状和方向信息时具有局限性，导致生成图像存在语义不一致的问题，如图 3 所示，本文在生成器中添加 CBAM，将通道注意力模块和空间注意力模块串联，分别嵌入到卷积神经网络中的不同层，以增强特征表示。

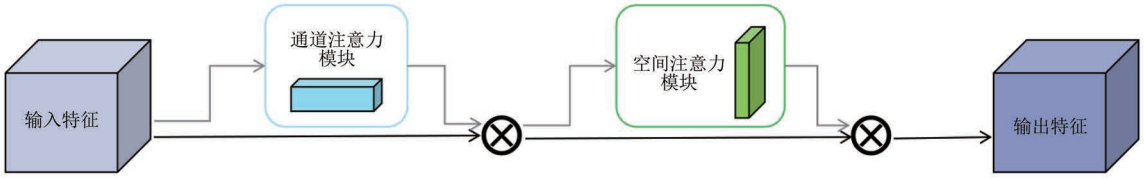

图 3 CBAM 整体结构

Fig. 3 Overall structure of CBAM

由图 4 可知，通道注意力模块首先会对输入特征 F 进行平均池化和最大池化操作，然后将这两种池化结果传递给共享的全连接层处理。接着，将全连接层的输出与原始特征相加，经过激活函数得到权重向量。最后采用逐通道的乘法加权方法，将权重向量应用到输入特征层上。具体的计算过程可以参考式(1)：

$$M_C(F)=\sigma(\mathrm{MLP}(\mathrm{AvgPool}(F))+\mathrm{MLP}(\mathrm{MaxPool}(F)))=\sigma(W_1(W_0(F^{c}_{avg}))+W_1(W_0(F^{c}_{max}))) \tag{1}$$

式中：σ 为 Sigmoid 函数；$W_0\in\mathbb{R}^{C\times C/r}$；$W_1\in\mathbb{R}^{C\times C/r}$；MLP 的权重由 W_0 和 W_1 共享，且 $F^{c}_{avg}\in\mathbb{R}^{1\times H\times W}$ 前有 ReLU 激活函数。

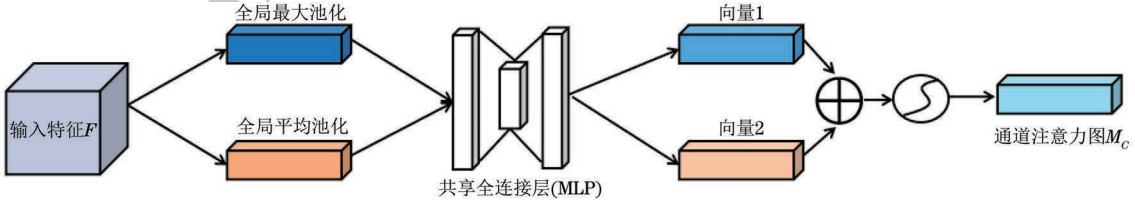

图 4 通道注意力模块结构

Fig. 4 Channel attention module structure

如图 5 所示，空间注意力模块与通道注意力模块的区别在于它关注输入图像特征的哪些部分信息更重要，可以看作是对通道注意力模块的补充。在空间注意力模块中，输入的图像特征 F' 会经历通道维度的最大池化和平均池化操作，从而生成包含不同尺度上下文信息的特征。接着，将这两种池化后的特征在通道维度上进行拼接，形成一个具有不同尺度上下文信息的特征向量，通过卷积层的处理，生成空间注意力权重。最后，将这些权重通过逐通道的乘法加权到输入特征层上。具体的计算过程可以参考式(2)：

$$M_S(F)=\sigma(f^{7\times7}([\mathrm{AvgPool}(F);\mathrm{MaxPool}(F)]))=\sigma(f^{7\times7}([F^{s}_{avg};F^{s}_{max}])) \tag{2}$$

式中：σ 为 Sigmoid 函数；$F^{s}_{avg}\in\mathbb{R}^{1\times H\times W}$；$F^{s}_{max}\in\mathbb{R}^{1\times H\times W}$；$f^{7\times7}$ 表示 7×7 大小的卷积核。

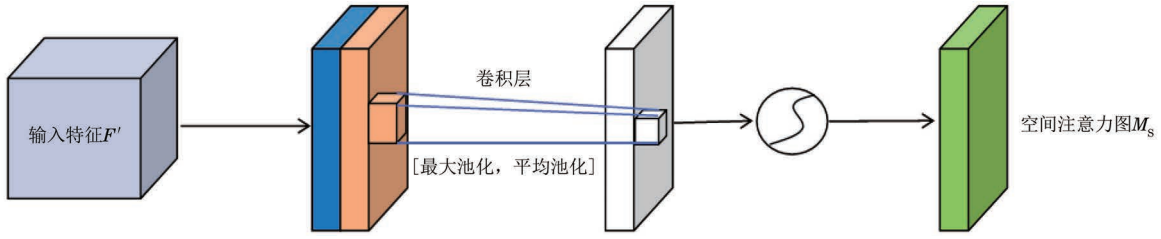

图 5 空间注意力模块结构

Fig. 5 Spatial attention module structure

综上所述，CBAM 的总体流程如下：首先，输入图像特征 F，$F\in\mathbb{R}^{C\times H\times W}$。然后，通过通道注意力模块生成一维通道注意力图 M_C，$M_C\in\mathbb{R}^{C\times 1\times 1}$。接着，通过空间注意力模块生成二维空间注意力图 M_S，$M_S\in\mathbb{R}^{1\times H\times W}$。最后，将通道注意力模块和空间注意力模块的输出特征进行逐元素相乘，得到最终的注意力增强特征。具体的计算过程可以参考式(3)和式(4)：

$$F'=M_C(F)\otimes F \tag{3}$$

$$F''=M_S(F')\otimes F' \tag{4}$$

2.4 对比损失

本文将 NT-Xent 引入判别器作为对比学习损失[23]，目的是缩小相关文本图像间的距离、增大不相关文本图像间的距离，使得相同文本描述生成的图像更近、不同文本描述生成的图像更远，以此确保模型语义一致性。

对于最小化相关文本图像之间的距离，如图 6 所示，根据文本描述 s 和对应图像 x，定义得分函数为：

$$S_{\text{sent}}(x,s)=\cos(f_{\text{img}}(x),f_{\text{sent}}(s))/\eta \tag{5}$$

式中：$\cos(\boldsymbol{u},\boldsymbol{v})=\dfrac{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{v}}{\|\boldsymbol{u}\|\cdot\|\boldsymbol{v}\|}$ 表示余弦相似度，也作为两个向量是否相似的评价标准；η 表示温度超参数；f_{img} 为图像编码器，用于提取图像的整体特征向量；f_{sent} 为文本编码器，用于提取文本的全局特征向量。则图像 x_i 和它所对应的文本描述 s_i 之间的对比损失的计算式为：

$$L_{\text{sent}}(x_i,s_i)=-\log_a\frac{\exp(\cos(f_{\text{img}}(x_i),f_{\text{sent}}(s_i))/\eta)}{\sum_{n=1}^{M}\exp(\cos(f_{\text{img}}(x_i),f_{\text{sent}}(s_n))/\eta)} \tag{6}$$

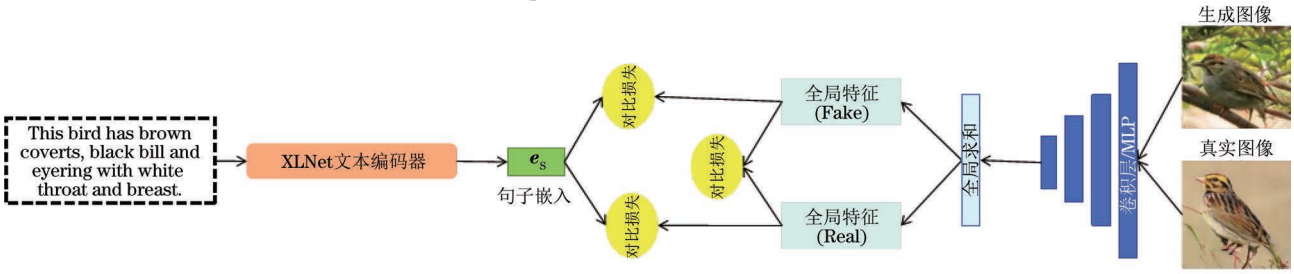

图 6 相关文本图像对比损失示意图

Fig. 6 Schematic diagram of the loss for comparing related text images

根据同一文本描述对应的真实图像和生成图像，定义它们之间的得分分数为：

$$S_{\text{img}}(x,\tilde{x})=\cos(f'_{\text{img}}(x),f'_{\text{img}}(\tilde{x}))/\eta \tag{7}$$

式中：f'_{img} 表示真实图像和生成图像的一个共同的图像编码器。则真实图像 x_i 和生成图像 $\tilde{x}_i$ 之间的对比损失的计算式为：

$$L_{\text{img}}(x_i,\tilde{x}_i)=-\log_a\frac{\exp(S_{\text{img}}(x_i,\tilde{x}_i))}{\sum_{n=1}^{M}\exp(S_{\text{img}}(x_i,\tilde{x}_n))} \tag{8}$$

对于最大化不相关文本图像之间的距离，如图 7 所示，对于两个不同文本描述所生成的图像，定义它们之间的得分分数为：

$$S'_{\text{img}}(\tilde{x}_i,\tilde{x}_j)=\cos(f'_{\text{img}}(\tilde{x}_i),f'_{\text{img}}(\tilde{x}_j))/\eta \tag{9}$$

则它们之间的对比损失的计算式为：

$$L'_{\text{img}}(\tilde{x}_i,\tilde{x}_j)=-\log_a\frac{\exp(S'_{\text{img}}(\tilde{x}_i,\tilde{x}_j))}{\sum_{k=1}^{2M}\exp(S'_{\text{img}}(\tilde{x}_i,\tilde{x}_k))} \tag{10}$$

式中：$\tilde{x}_i$ 和 $\tilde{x}_j$ 为两个不同语义的生成图像；$\tilde{x}_i$ 和 $\tilde{x}_k$ 为同一批次的不相关语义的生成图像。在计算生成图像之间的对比损失时，批大小(Batch Size)大小不变，生成图像的数量须是原来的 2 倍，所以式中的 M 变为 $2M$。

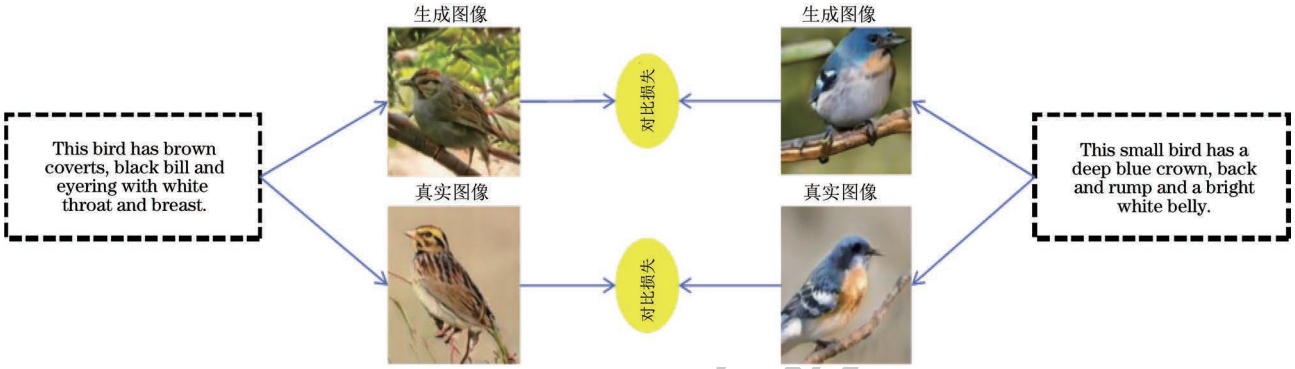

图 7 不相关文本图像对比损失示意图

Fig.7 Schematic diagram of the loss for comparing irrelevant text image

2.5 目标函数

生成器网络的目标函数定义为：

$$L_G=L_{adv\_G}+\lambda_1 L_{DAMSM}+\lambda_2 L_{sent}+\lambda_3 L_{img}+\lambda_4 L'_{img} \tag{11}$$

$$L_{adv\_G}=-E_{G(z)\sim Pg}[D(G(z),\boldsymbol{e})] \tag{12}$$

式中：λ_1 是 DAMSM 损失的权重；λ_2 是文本与图像之间的对比损失的权重；λ_3 是生成图像与真实图像之间的对比损失的权重；λ_4 是两个不同文本所对应的生成图像之间的对比损失的权重；L_{DAMSM} 表示 DAMSM 计算生成图像区域和单词之间的损失，用来衡量生成图像和语义间的匹配度。

判别器网络的目标函数定义为：

$$L_D=-E_{x\sim Pr}[\min(0,-1+D(x,\boldsymbol{e}))]-\frac{1}{2}E_{G(z)\sim Pg}[\min(0,-1-D(G(z),\boldsymbol{e}))]-\frac{1}{2}E_{x\sim Pmis}[\min(0,-1-D(x,\boldsymbol{e}))]+kE_{x\sim Pr}[(\|\nabla_x D(x,\boldsymbol{e})\|+\|\nabla_e D(x,\boldsymbol{e})\|)^P] \tag{13}$$

与 DF-GAN 模型相同，本文利用铰链损失、匹配感知梯度惩罚(MA-GP)和单向输出，使得对抗训练更稳定，生成器生成的图像更符合文本描述。

3 实验结果与分析

3.1 实验设置

本文实验基于 Ubuntu 20.04 系统，搭载 GPU 为 GeForce RTX 3090(24 GB)，使用 Python 3.9 编程语言、PyTorch 深度学习框架，CUDA 版本为 11.3。优化器选择的是 Adam 优化器，其参数设置为 $\beta_1=0$，$\beta_2=0.999$；生成器和判别器的学习率为 $\alpha=0.0001$ 和 $\alpha=0.0004$；轮数(Epoch)为 800；模型使用的 Batch Size 为 32，损失函数中超参数 $\lambda=0.2$。

3.2 实验数据集及评价指标

3.2.1 数据集

在文本图像生成领域，常见的数据集有 CUB 数据集、Oxford-102 数据集和 COCO 数据集等。本文选择了 CUB 鸟类数据集和 Oxford-102 花卉数据集作为实验数据集。CUB 鸟类数据集是一个用于鸟类识别和定位的常见数据集，该数据集包含约 11 788 幅来自 200 种鸟类的图像，每种鸟类约有 50 幅图像，并且每幅图像都有鸟类边界框和标签的注释。Oxford-102 花卉数据集与 CUB 数据集类似，该数据集涵盖了 102 个不同的花卉类别，共计 8 189 幅图像。每幅图像都配备了 10 个不同的文本描述，为文本图像生成提供了丰富的素材和标注数据。

3.2.2 评价指标

本文使用 IS(Inception Score)[24] 和 FID(Fréchet Inception Distance)[25-26] 两个评价指标来量化实验结果。

IS 是评估文本生成图像模型性能的重要指标，能够客观地评估生成的图像质量和图像多样性。IS 主要是利用预训练的 Inception v3 模型来评估边缘分布与条件分布之间的 KL 散度，其计算公式为：

$$I=\exp(E_x D_{KL}(p(y|x)\,\|\,p(y))) \tag{14}$$

式中：E 表示期望；D_{KL} 表示边缘分布和条件分布之间的 KL 散度。IS 得分高低与 KL 散度的值呈现正相关。一个较高的 IS 得分意味着 KL 散度值也相应较高。

FID 是评估文本生成图像领域模型性能的另一个重要指标。它旨在衡量两个数据集之间的相似度，特别是原始图像数据分布与生成图像数据分布间的相似性。FID 也是利用预训练的 Inception 模型对图像进行特征编码，随后计算两者之间的 Fréchet 距离，计算公式如下：

$$D_{FID}=\|\mu_{x_1}-\mu_{x_2}\|^2+\mathrm{tr}\left(\boldsymbol{\Sigma}_{x_1}+\boldsymbol{\Sigma}_{x_2}-2\left(\boldsymbol{\Sigma}_{x_1}\boldsymbol{\Sigma}_{x_2}\right)^{\frac{1}{2}}\right) \tag{15}$$

式中：x_1、x_2 分别为真实图像、生成图像；μ_{x_1}、μ_{x_2} 是各自特征向量均值；$\boldsymbol{\Sigma}_{x_1}$、$\boldsymbol{\Sigma}_{x_2}$ 为特征向量的协方差矩阵。FID 值越低，说明真实图像和生成图像的分布就越接近，两个图像就越相似，生成的图像质量就越高。

3.3 实验结果与分析

3.3.1 定量评价

使用与模型 DF-GAN 相同的测试方法对生成图像进行测试。在 CUB 数据集上，每一个文本描述可生成 10 幅图像，测试集总共包含 29 330 幅图像，对其计算 FID 和 IS 分数。在 Oxford-102 数据集上，每一个文本描述生成 1 幅图像，总共生成 2 060 幅图像，对其计算 FID 和 IS 分数。对比 MirrorGAN、DM-GAN 和 DF-GAN 等模型，详细的定量结果如表 1 和表 2 所示，其中，加粗数据表示最优值。在 CUB 数据集上，相较于 DF-GAN，本文模型将 FID 从 14.81 降至 12.15，将 IS 从 4.76 提升至 4.98；在 Oxford-102数据集上，相较于DF-GAN，本文模型将 FID 从 42.61 降至 37.82，将 IS 从 3.80 提升至 3.97。从而可知，本文提出的模型生成的图像更接近真实图像，语义一致性得到一定增强，图像质量也有一定提升。

表 1 不同数据集不同方法的 FID 对比

Table 1 FID comparison of different methods on different datasets

方法	FID	
	CUB	Oxford-102
StackGAN	51.89	55.28
AttnGAN	23.98	—
DM-GAN	16.09	43.92
SE-GAN	18.17	42.28
DAE-GAN	15.19	—
MirrorGAN	18.34	43.29
SSA-GAN	15.61	—
DF-GAN(预训练)	14.81	42.61
Ours	**12.15**	**37.82**

表 2 不同数据集不同方法的 IS 对比

Table 2 IS comparison of different methods on different datasets

方法	IS	
	CUB	Oxford-102
StackGAN	3.70±0.04	3.20±0.01
AttnGAN	4.36±0.03	—
DM-GAN	4.75±0.07	3.91±0.06
DAE-GAN	4.42±0.04	—
CFA-HAGAN	4.54±0.04	3.98±0.03
MirrorGAN	4.56±0.05	3.84±0.01
DF-GAN(预训练)	4.76±0.04	3.80±0.02
Ours	**4.98±0.03**	**3.97±0.01**

3.3.2 定性分析

图 8 展示了本文模型 DXC-GAN 与其他模型(如 MirrorGAN、DM-GAN 和 DF-GAN)在 CUB 数据集上生成图像效果对比。通过对比可以看出，本文模型的生成图像与文本描述之间的契合度更高。在有的示例中，本文模型生成的鸟的形状更完整，且鸟的属性与文本描述更为接近。举例来说，对于给定的文本描述“This bird is white with grey and has a long, pointy beak.”，MirrorGAN 生成的鸟的轮廓模糊，存在失真问题；DM-GAN 生成的鸟虽然展现了脚部特征，但仍然不够完整，且缺乏真实感；DF-GAN 生成的鸟轮廓清晰，但只有一只爪子；本文模型生成的鸟类图像具有灰色的背部和尾部、明亮的白色腹部，同时脚部特征清晰且完整，与给定的文本描述相符。

图 9 展示了在 Oxford-102 数据集上对 MirrorGAN、DM-GAN、DF-GAN 以及本文模型的定性比较结果。观察结果显示，本文模型生成的花与真实图像更为相符，且花的轮廓细节更为清晰。例如，对于给定的文本描述“This flower has large smooth white petals that turn yellow toward the center.”，尽管 MirrorGAN 生成的花在颜色上与描述相符，但在花瓣的完整性上存在缺失，显得不够完美；DM-GAN 生成的花在颜色方面符合描述，但两朵花连在一起，使得花的辨识度较低；DF-GAN 生成的花虽然在某些方面表现不错，但花瓣的完整性略显欠缺；本文模型生成的花具有大而光滑的白色花瓣，随着朝向中心的变化逐渐转变为黄色，同时在细节上也满足了给定的文本描述语义，呈现出更加逼真的效果。

3.4 消融实验

本文在 CUB 数据集上对提出的 3 个关键模块进行了详细的消融研究，并将实验结果记录在表 3 中。表中数据显示，在原始 DF-GAN、DF-GAN+XLNet、DF-GAN+CBAM 和 DF-GAN+XLNet+CBAM 4 个网络的基础上添加对比损失函数后，图像生成能力都得到了不同程度的提升。这表明对比学习方法对文本生成图像具有积极作用，同时也说明 XLNet 文本编码器、CBAM 注意力模块和对比学习方法对文本生成图像任务都具有积极作用。

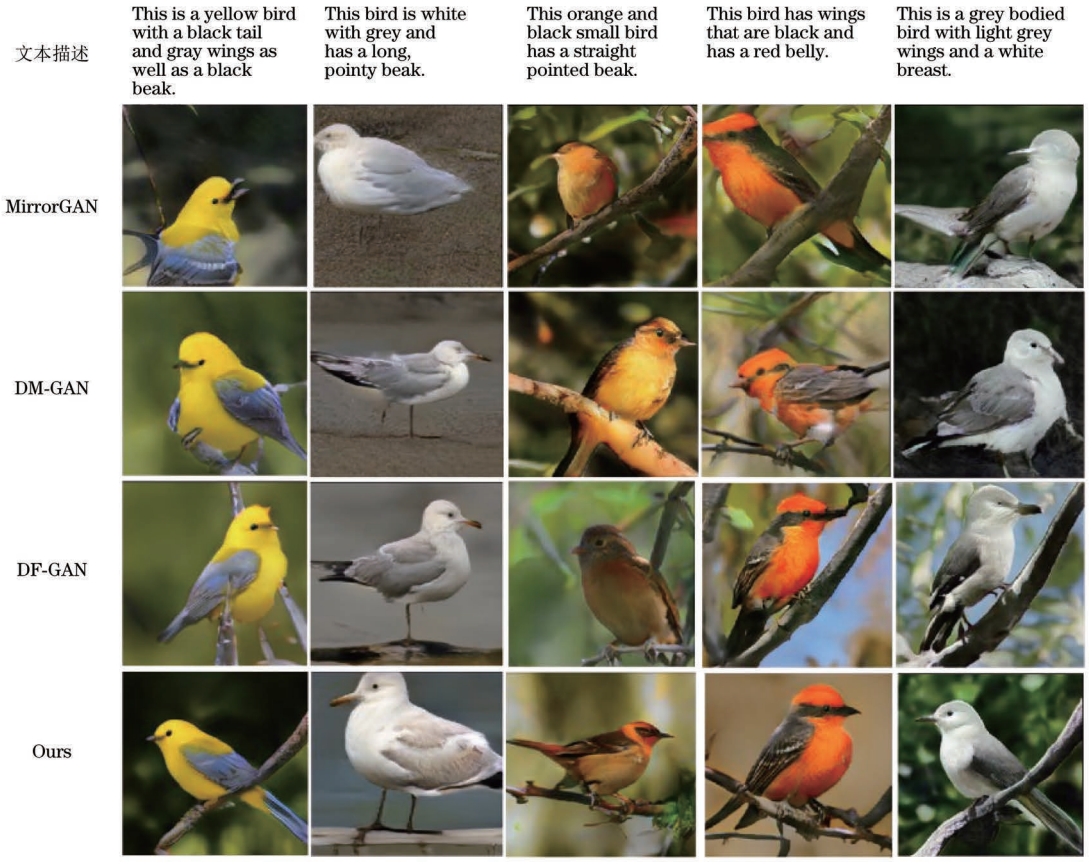

图 8 CUB 数据集生成图像对比

Fig. 8 Comparison of images generated from the CUB dataset

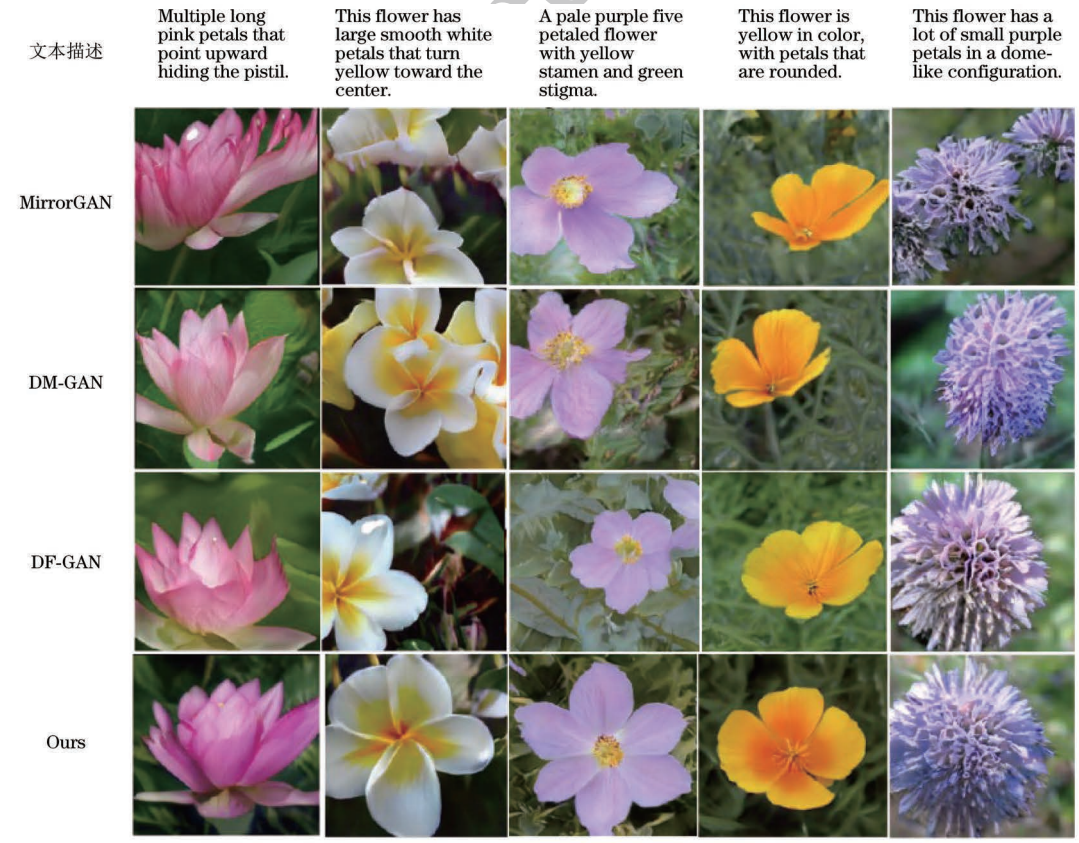

图 9 Oxford-102 数据集生成图像对比

Fig. 9 Comparison of images generated from the Oxford-102 dataset

表 3 CUB 数据集总体消融实验各项评价指标对比

Table 3 Comparison of various evaluation indicators for the overall ablation experiment on the CUB dataset

方法	FID	IS
DF-GAN	14.81	4.76±0.04
DF-GAN+XLNet	13.55	4.81±0.01
DF-GAN+CBAM	14.67	4.78±0.02
DF-GAN+对比学习	14.62	4.74±0.05
DF-GAN+XLNet+CBAM	12.75	4.84±0.07
DF-GAN+XLNet+对比学习	13.22	4.89±0.01
DF-GAN+CBAM+对比学习	14.07	4.81±0.04
DF-GAN+XLNet+CBAM+对比学习	**12.15**	**4.98±0.03**

4 结束语

本文基于 DF-GAN 模型提出了 DXC-GAN 模型，以提高文本到图像的生成质量和多样性。首先，将原有的文本编码器替换为 XLNet 文本编码器，使得模型在文本理解和表达能力方面得到了显著的提升。其次，在生成器中添加 CBAM 模块，通过空间和通道注意力机制的结合，有效地提升了生成器对图像细节和结构的关注能力，从而生成更加细致丰富的图像。同时，在判别器中引入了对比损失函数，以进一步增强模型对真实图像和生成图像的区分能力，从而提高生成图像的逼真度。在 CUB 和 Oxford-102 数据集上进行大量实验，结果表明本文提出的模型生成的图像质量更高，更具多样性和真实感。本文只在 CUB 和 Oxford-102 数据集上进行实验，存在局限性，下一步将基于 COCO 数据集进行大量实验以验证 DXC-GAN 的有效性和优越性。

参考文献

[1] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from Scratch[J]. Journal of Machine Learning Research, 2011, 12: 2493-2537.

[2] ANDREW A M. Multiple view geometry in computer vision [J]. Kybernetes, 2001, 30(9/10): 1333-1341.

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

[4] TAO M, TANG H, WU S S, et al. DF-GAN: deep fusion generative adversarial networks for text-to-image synthesis [EB/OL]. [2024-01-02]. https://arxiv.org/abs/2008.05865v1.

[5] YANG Z L, DAI Z H, YANG Y M, et al. XLNet: generalized autoregressive pretraining for language understanding [J]. Advances in Neural Information Processing Systems, 2019, 8: 5753-5763.

[6] GRAVES A. Supervised sequence labelling with recurrent neural networks[M]. Berlin, Germany: Springer, 2012: 37-45.

[7] 任欢, 王旭光. 注意力机制综述[J]. 计算机应用, 2021, 41(S1): 1-6.
REN H, WANG X G. Review of attention mechanism[J]. Journal of Computer Applications, 2021, 41(S1): 1-6. (in Chinese)

[8] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations [C]// Proceedings of the 37th International Conference on Machine Learning. New York, USA: ACM Press, 2020: 1597-1607.

[9] REED S, AKATA Z, YAN X, et al. Generative adversarial text to image synthesis [C] // Proceedings of International Conference on Machine Learning. New York, USA: ACM Press, 2016: 1060-1069.

[10] 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 IEEE International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2017: 5908-5916.

[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, 2019, 41(8): 1947-1962.

[12] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2018: 1316-1324.

[13] QIAO T T, ZHANG J, XU D Q, et al. MirrorGAN: learning text-to-image generation by redescription [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2019: 1505-1514.

[14] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2019: 5795-5803.

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

[16] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 5998-6008.

[17] 刘建伟, 宋志妍. 循环神经网络研究综述[J]. 控制与决策, 2022, 37(11): 2753-2768.
LIU J W, SONG Z Y. Overview of recurrent neural networks[J]. Control and Decision, 2022, 37(11): 2753-2768. (in Chinese)

[18] DEVLIN J, CHANG M W, LEE K. BERT: pre-training of deep bidirectional Transformers for language understanding [EB/OL]. [2024-01-02]. https://arxiv.org/pdf/1810.04805.

[19] GUO M H, XU T X, LIU J J, et al. Attention mechanisms in computer vision: a survey [J]. Computational Visual Media, 2022, 8(3): 331-368.

[20] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training [EB/OL]. [2024-01-02]. https://arxiv.org/abs/1810.04805.

[21] CHEN K, WANG J, CHEN L C, et al. ABC-CNN: an attention based convolutional neural network for visual question answering[EB/OL]. [2024-01-02]. https://arxiv.org/abs/1511.05960.

[22] XU X, WANG T, YANG Y, et al. Cross-modal attention

with semantic consistence for image-text matching[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(12): 5412-5425.
[23] GUAN S Y, LOEW M. Evaluation of generative adversarial network performance based on direct analysis of generated images[C]// Proceedings of the IEEE Applied Imagery Pattern Recognition Workshop. Washington D. C., USA: IEEE Press, 2019: 1-5.
[24] OBUKHOV A, KRASNYANSKIY M. Quality assessment method for GAN based on modified metrics inception score and Fréchet inception distance[M]//RADEK S, PETR S, ZDENKA P. Software engineering perspectives in intelligent systems. Berlin, Germany: Springer, 2020: 102-114.
[25] 张佳, 张丽红. 基于条件增强和注意力机制的文本生成图像方法[J]. 测试技术学报, 2023, 37(2): 112-119.
ZHANG J, ZHANG L H. Research on text to image based on conditioning augmentation and attention mechanism[J]. Journal of Test and Measurement Technology, 2023, 37(2): 112-119. (in Chinese)
[26] ZHANG H, GOODFELLOW I, METAXAS D, et al. Self-attention generative adversarial networks[EB/OL]. [2024-01-02]. https://arxiv.org/abs/1805.08318.

文字编辑 金胡考
栏目编辑 赖玉玲