基于语义增强的单阶段文本生成图像方法

兰才俊[1] 姚剑敏[1,2] 胡海龙[1,2] 陈恩果[1] 严 群[1]
LAN Caijun YAO Jianmin HU Hailong CHEN Enguo YAN Qun

摘 要 文本到图像生成是一项极具挑战性的跨模态任务，目标是根据给定文本描述生成对应的图像。尽管现阶段相关研究在视觉呈现方面效果优异，但仍存在细节表达不够精细、语义一致性欠佳等问题。基于此，文章提出了一种基于语义增强的生成对抗模型，将文本进行编码后送入条件增强模块进行处理，丰富文本语义特征。在生成网络中，添加一个自适应块，在仿射变换前将上一层的输出和文本语义信息输入自适应块进行进一步的信息增强。并通过引入对比损失，提高文本与生成图像之间的语义一致性。将这一方法在 MSCOCO 和 CUB birds 200 两个数据集上进行训练测试，实验结果表明，与其他模型相比，性能得到了较高提升。

关键词 语义增强；文本生成图像；生成对抗网络；对比损失

doi：10.3969/j.issn.1672-9528.2025.01.001

0 引言

文本到图像生成任务是基于给定的文本描述生成语义一致的、真实的图像。这是一项具有广泛应用前景的任务，如艺术生成[1]、计算机辅助设计[2]、图像编辑[3]。由于其广泛的应用和挑战，文本到图像的生成已经成为计算机视觉和自然语言处理中的活跃领域。

基于扩散和自回归模型的方法，如 DALL·E 2[4]、Parti[5]等，已经显示出很强的生成能力，并且明显优于此前的条件生成对抗网络[6]。虽然这些模型已经取得了很大的进展，但均依赖于迭代推理，模型规模巨大，需要极长的计算时间和极高的硬件要求。与 CGAN 相比，后者只需一次正演即可生成图像，且模型尺寸相对较小，具有计算速度快、计算量小的优点，基于 CGAN 的方法[7-8]展现出了优秀的图像效果。

现有的基于 CGAN 的文本到图像生成模型大致可以分为两类：单阶段模型和多阶段模型。多阶段模型[9]通过堆叠一系列生成器和鉴别器，逐渐增加图像的细节和分辨率，最终获得逼真的图像，从粗到细的方式生成图像；单级模型[10]仅引入一对生成器和鉴别器来生成与文本描述语义一致的图像。与多阶段模型相比，单阶段模型更有利于模型的收敛和稳定性，本文使用的是单阶段架构。

虽然最新的单阶段生成模型已经取得了优异的成果，但是在语义一致性和细节描述方面还存在细节不足，导致生成图片与文本的描述不够逼真和自然。针对语义不够丰富的问题，本文引入条件增强模块，丰富文本向量的语义信息。针对语义描述一致性不足的问题，在生成器阶段进行仿射变换前设计了一个自适应块进行了再一次的信息增强，最后，引入对比学习，加入了对比损失，减小相同语义图像对之间的距离，增大不同语义图像对之间的距离，使生成的图像更加真实。本文的贡献如下：

（1）在文本编码器后进行条件增强，在给定少量文本图像数据对的情况下提供更多的增强数据。

（2）在生成器阶段，引入自适应模块，在仿射变换前将文本信息和图像特征通过自适应块进行进一步的信息增强。

（3）优化损失函数，加入对比损失，提高生成图像与文本的语义一致性。

1 相关工作

1.1 文本到图像生成

早期的 CGAN 方法通过一次前向传播即能生成图像，提高生成效率。此外，研究人员根据生成器和鉴别器数量，将基于 CGAN 的方法划分为两类：多阶段模型和单阶段模型。随后，自回归模型和扩散模型的发展进一步推动了文本到图像生成任务的突破。自回归模型通过逐步解码输入的文本特征来生成相应的图像，如 Parti 将图像生成视为翻译任务，Cogview2[11]采用分层变换和并行自回归的方式生成图像，DALL·E2 结合预训练的 CLIP 和扩散模型生成图像等。虽然生成的图像质量明显提高，但这些方法需要耗时和资源密集

1. 福州大学物理与信息工程学院 福建福州 350108
2. 晋江市博感电子科技有限公司 福建泉州 362216
[基金项目] 国家自然科学基金（62175032）；福建省杰出青年基金项目（2024J010046）

的迭代过程。相比之下，CGAN 只需一次前向传播就能生成图像，训练时间大幅缩短。

1.2 单阶段模型

Reed 等人 [12] 率先基于单级模型开展文本到图像生成任务的研究。然而受限于该模型，其仅能依据文字描述合成低分辨率图像。后来，Tao 等人提出 DF-GAN。该方法通过在生成器上叠加多个 UpBlock，对图像特征进行上采样，从而能够直接生成高分辨率且逼真的图像。DF-GAN 解决了当时多阶段模型训练复杂、文本信息利用不足的问题，具有结构简单、易于收敛的优点。由于单级模型的优越性，诸多后续工作都围绕着这一结构展开。Zhang 等人 [13-14] 提出了 DT-GAN 和 DiverGan 两种方法，并引入了条件归一化，设计了两个新的注意模块，使得文本能够更有效地融入生成过程。Ye 等人 [15] 通过开发 RAT-GAN，将递归神经网络与所有融合块进行连接，模拟独立融合块之间的长期依赖关系，生成逼真可信的图像。

1.3 多阶段模型

多阶段方法 StackGAN[16] 和 StackGAN++[17]，使用多个条件生成对抗网络（CGAN）从粗到精的生成图像实例。为进一步完善图像细节，Xu 等人 [18] 提出 AttnGAN 算法，利用注意力机制细化与文字相关的图像区域，增强图像的视觉真实感。Zhu 等人 [19] 提出的 DM-GAN 利用记忆网络存储丰富的语义信息，使网络在处理复杂文本描述时能够更好地捕捉图像的细节和语义。Ruan 等人 [20] 提出的 DAE-GAN，可以准确捕捉文本的多个方面，生成多样化和高质量的图像。Tan 等人 [21] 提出的 DR-GAN 算法，引入分布正则化机制，通过最小化真实图像和生成图像之间的分布差异，使生成图像的分布更接近真实图像。上述工作均基于多阶段的方法，通过引入注意力机制、记忆网络以及分布正则化等技术，进一步提升了文本到图像生成的性能和效果。

2 网络结构

本文所提出的方法模型结构如图 1 所示，整个模型由 3 个部分组成，分别为文本编码器、生成器和鉴别器。模型整体遵循 DF-GAN 基本架构。

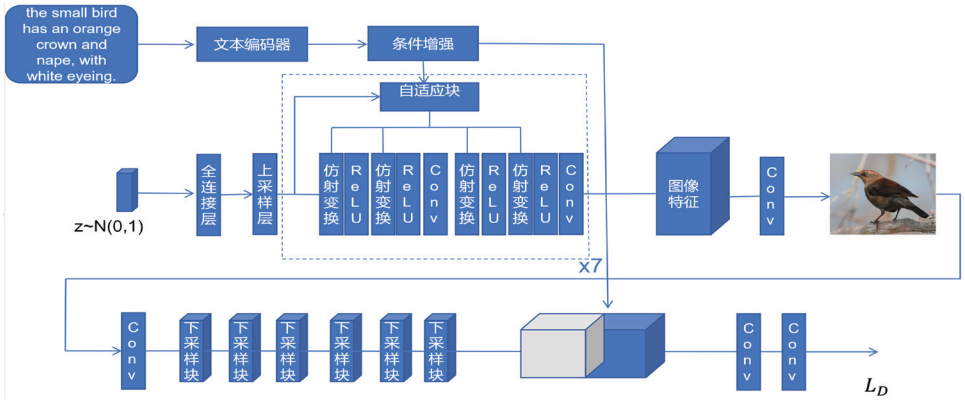

图 1 整体模型图

2.1 条件增强模块

当文本特征的数量较少时，可以使用条件增强模块来增强语义信息，生成更多样的图像特征。文本特征首先通过编码器得到初始的文本特征 φ_t，然后条件增强模块引入一个新的变量 $\hat{c}$ 作为约束。此前的文本生成方法均是直接固定编码后的特征，未充分利用语义信息。本文将文本编码后的特征 φ_t 进行随机采样得到 $\hat{c}$。具体来说，就是从一个独立的高斯分布 $N(\mu(\varphi_t), \Sigma(\varphi_t))$ 进行随机采样，获取增强后的文本特征，$\mu(\varphi_t)$ 表示平均值，$\Sigma(\varphi_t)$ 表示对角协方差矩阵，条件增强模块如图 2 所示，公式表示为：

$$\hat{c} = \mu + \sigma \otimes \varepsilon \tag{1}$$

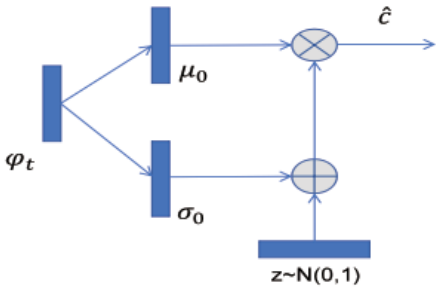

图 2 条件增强模块

为防止生成器模型过拟合，本文在目标函数中引入了一项额外约束项。使用 KL 散度来衡量生成器的采样分布与潜在语义空间的相似程度。

$$D_{\mathrm{KL}}\left(N\left(\mu(\varphi_t), \Sigma(\varphi_t)\right) \| N(0, I)\right) \tag{2}$$

式中：$N(0, I)$ 表示采样分布；$N(\mu(\varphi_t), \Sigma(\varphi_t))$ 表示语义空间分布。

2.2 自适应块

本文改进了基于单阶段的文本图像中的生成器结构，引入自适应块，文本特征在进行仿射变换前将由自适应块更新，再进行信息增强。图 3 显示了自适应块的体系结构。自适应块有两个输入：当前层特征图 h_{t-1} 和句子向量 $\boldsymbol{s}$。卷积核大小与特征图的维度相同。h'_{t-1} 为深度卷积变换输出。语义条件 s_c 是 h'_{t-1} 和 s 连接的结果，将传递到多层感知器中。使用 Tanh 作为最终的激活函数来计算值为 (-1,1) 的注意图。后采用残差结构得到 s 的自适应信息增强型句子向量 $\boldsymbol{s}'$，并利用其进行仿射变换。

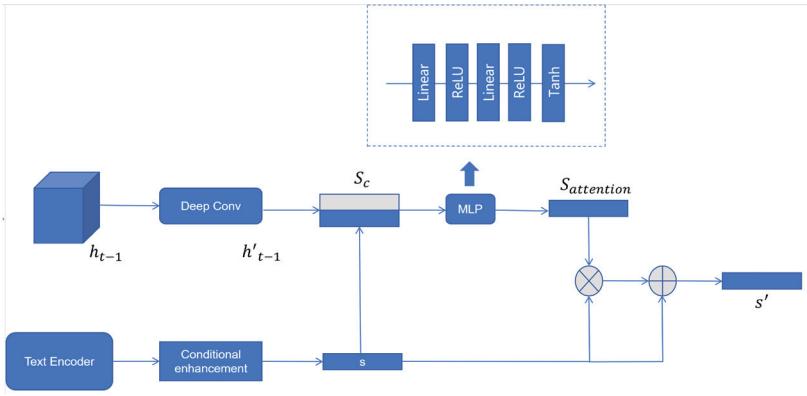

图 3 自适应块

2.3 对比学习

本文采用 NT-Xent 作为对比学习损失，对比学习可以使模型学习到同类实例之间的共同特征，以及不同实例之间的差异，通过最小化相关实例间的距离，最大化不相关实例间的距离，通过对比学习可以有效提高生成图像与输入文本描述的语义一致性。为衡量两个向量的相似程度，本文使用了余弦相似度作为度量指标：

$$\cos(\boldsymbol{\mu},\boldsymbol{v})=\frac{\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{v}}{\|\boldsymbol{\mu}\|\cdot\|\boldsymbol{v}\|} \tag{3}$$

（1）文本与图像对比损失

对于一个图像 x_i 和非其文本描述句子特征 s，L_{text} 的公式为：

$$L_{\text{text}}=-\log\frac{\exp\left(\frac{\cos\left(f_{\text{img}}(x_i),f_{\text{text}}(s_i)\right)}{\eta}\right)}{\Sigma_{j=1}^{N}\exp\left(\frac{\cos\left(f_{\text{img}}(x_i),f_{\text{text}}(s_j)\right)}{\eta}\right)} \tag{4}$$

（2）生成图像与真实图像对比损失

对于一个合成图像 $\hat{x}_i$ 和具有相同文本描述的真实图像 x_i，L_{img} 公式为：

$$L_{\text{img}}=-\log\frac{\exp\left(\frac{\cos\left(f_{\text{img}}(x_i),f_{\text{img}}(\hat{x}_i)\right)}{\eta}\right)}{\Sigma_{j=1}^{N}\exp\left(\frac{\cos\left(f_{\text{img}}(\hat{x}_i),f_{\text{img}}(\hat{x}_j)\right)}{\eta}\right)} \tag{5}$$

式中：f_{text} 和 f_{img} 为训练好的文本编码器和图像编码器；η 为温度超参数，N 为批次大小。

2.4 目标函数

生成器的目标函数为：

$$L_G=L_{\text{adv\_G}}+\lambda_1 L_{\text{text}}+\lambda_2 L_{\text{img}}+D_{\text{KL}} \tag{6}$$

$$L_{\text{adv\_G}}=-E_{G(z)\sim P_g}[D(G(z),e)] \tag{7}$$

式中：λ_1 为文本与图像之间的对比损失的权重；λ_2 为真实图像与生成图像对比损失的权重。

判别器的目标函数为：

$$\begin{aligned}L_D=&-E_{x\sim P_r}[\min(0,-1+D(x,e))]\\&-E_{G(z)\sim P_g}[\min(0,-1-D(G(z),e))]/2\\&-E_{x\sim P_{\text{mis}}}[\min(0,-1-D(x,e))]/2\\&+kE_{x\sim P_r}\left[\left(\|\nabla_x D(x,e)\|+\|\nabla_e D(x,e)\|\right)^p\right]\end{aligned} \tag{8}$$

3 实验

3.1 数据集

本文在常用的 T2I 数据集上进行了实验：CUB-Bird[22] 和 MS-COCO[23]。前者有试验鸟 2933 只（50 类）和训练鸟 8855 只（150 类）。每只鸟都有 10 条英语文本来描述细微的视觉场景。后者由 82 783 个训练图像和 40 504 个测试图像组成，每个图像都有 5 条文本注释。

3.2 实验环境

本文所提方法在 PyTorch 上实现，实验的硬件条件是 NVIDIA RTX 3090 GPU，训练时使用 Adam 优化器进行优化，其中 β_1=0.0、β_2=0.9、batch=16。生成器的学习率被设置为 0.000 1，而鉴别器的学习率被设置为 0.000 4。

3.3 评价指标

为验证所提出方法的可行性，本文选择使用常见的评估指标费雷歇距离 FID[24] 来衡量模型的性能，FID 是用于衡量生成的图像分布和真实图像分布之间的一致性。较低的 FID 意味着生成的图像更接近真实图像。

$$\text{FID}=\|\mu_r-\mu_g\|_2^2+\text{tr}\left(\Sigma_r+\Sigma_g-2(\Sigma_r\Sigma_g)^{\frac{1}{2}}\right) \tag{9}$$

式中：r、g 为真实图像表示和生成图像表示；μ 表示均值；$\boldsymbol{\Sigma}$ 表示协方差；tr 为矩阵的迹。

3.4 定量评价

本文将测试结果与一些当前主流的 GAN 模型在 CUB 和 COCO 两个数据集上进行对比，对比结果如表 1 所示，由于本模型是在 DF-GAN 的基础上进行的改进，与 DF-GAN 相比，在 CUB 数据集上本模型将 FID 从 14.81 降到 13.56，在 COCO 数据集上将 FID 从 21.42 降到了 20.54，从数据分析的角度来看，本文提出的方法具有一定的优势，生成的图像质量得到提升，生成的图像与文本描述的语义性更加一致。

表 1 不同模型 FID 结果对比表

模型	CUB-FID	COCO-FID
DM-GAN	16.09	32.64
DF-GAN	14.81	21.42
AttnGAN	23.98	35.49
DAE-GAN	15.19	28.12
Ours	13.56	20.54

3.5 定性评价

将真实图像、本文方法生成的图像和 DF-GAN 模型生成的图像作可视化对比，给定同一句文本，对比 3 张图片效果。

在 CUB 数据集上的可视化效果如图 4 所示。在第 1 幅图像中，本模型生成的图像生动描述了黄色的腹部和浅褐色的头部，DF-GAN 生成的图像虽然和文本描述的大致符合，但生成的图像不够自然。从其他多幅图像对比中，也可看出本模型相对于 DF-GAN 生成的图像质量得到一定的提升，对于文本描述的一些细节，在生成图像中都得到了较好描述，总而言之，本文提出的方法能够更好地使生成图像符合文本描述。

图 4 CUB-Bird 数据集实验

在 COCO 数据集上的生成图像对比如图 5 所示，COCO 数据集图像中的种类丰富，场景较复杂，所以模型的生成图像只能尽可能保证真实性。在第 3 幅图像中，生成的食物较 DF-GAN 更加真实，图像更加清晰。第 4 幅生成图像中较好的描述了比萨和丰富的食物。通过在两个数据集上的可视化对比，证明改进的模型相较于 DF-GAN 具有一定提升。

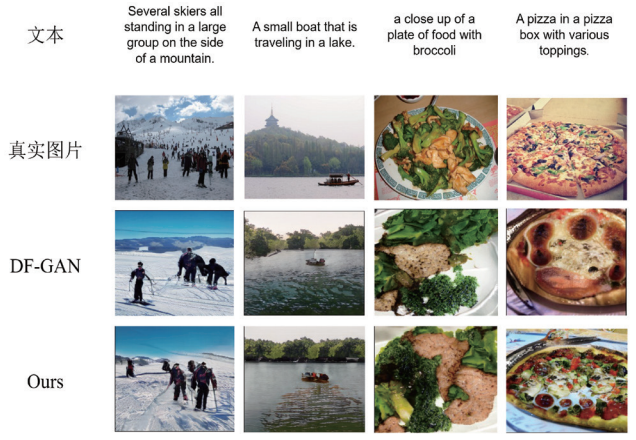

图 5 COCO 数据集实验

4 结论

本文设计了一种基于条件语义增强的文本生成图像方法，在 DF-GAN 的基础上，引入条件增强模块，丰富文本特征，通过在生成器中添加自适应块，进一步对语义信息进行了增强，为了提高生成图像与文本的语义一致性，改进了损失函数，引入了对比损失。通过实验证明了改进后模型的有效性。下一步在生成器中引入更多的语义信息，如单词信息融入生成器中，如何指导生成器生成更加细腻的图像。

参考文献：

[1] ZHI J L. PixelBrush: art generation from text with GANs[J/OL]. Computer science,2017[2024-03-19].https://www.semanticscholar.org/paper/PixelBrush-%3A-Art-Generation-from-text-with-GANs-Zhi/9cfdf0c3248c3703481ee93ff8abd4c08fc0a75f.

[2] SANGHI A, CHU H, LAMBOURNE J G, et al. CLIP-Forge: towards zero-shot text-to-shape generation[C/OL]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).Piscataway:IEEE,2022[2024-04-09].https://ieeexplore.ieee.org/document/9878592.

[3] KAWAR B, ZADA S, LANG O, et al. Imagic: text-based real image editing with diffusion models[C/OL]// 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).Piscataway:IEEE,2023[2024-05-14].https://ieeexplore.ieee.org/document/10203581.

[4] RAMESH A, DHARIWAL P, NICHOL A, et al. Hierarchical text-conditional image generation with clip latents[DB/OL]. (2022-04-13)[2024-05-19].https://doi.org/10.48550/arXiv.2204.06125.

[5] YU J H, XU Y Z, KOH J Y, et al. Scaling autoregressive models for content-rich text-to-image generation[DB/OL]. (2022-06-22)[2024-04-11].https://doi.org/10.48550/arXiv.2206.10789.

[6] MIRZA M, OSINDERO S. Conditional generative adversarial nets[DB/OL]. (2014-11-06)[2024-06-19].https://doi.org/10.48550/arXiv.1411.1784.

[7] KANG M G, ZHU J Y, ZHANG R, et al. Scaling up GANs for text-to-image synthesis[C/OL]// 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway:IEEE,2023[2024-05-10].https://ieeexplore.ieee.org/document/10205294.

[8] TAO M, BAO B K, TANG H, et al. Galip: generative adversarial clips for text-to-image synthesis[DB/OL]. (2023-01-30)[2024-03-12].https://doi.org/10.48550/arXiv.2301.12959.

[9] QUAN F N, BO L, LIU Y X. ARRPNGAN: text-to-image GAN with attention regularization and region proposal networks[J]. Signal processing: image communication, 2022,106(8): 116728.

[10] TAO M, TANG H, WU F, et al. DF-GAN: a simple and effective baseline for text-to-image synthesis[DB/OL].(2022-10-15)[2024-06-19].https://doi.org/10.48550/arXiv.2008.05865.

[11] DING M, ZHENG W D, HONG W Y, et al. Cogview2: faster and better text-to-image generation via hierarchical transformers[DB/OL].(2022-05-27)[2024-05-06].https://doi.org/10.48550/arXiv.2204.14217.

[12] REED S, AKATA Z, YAN X C, et al. Generative adversarial text to image synthesis[C]//Proceedings of the 33rd International Conference on International Conference on Machine Learning. New York: JMLR.org, 2016:1060-1069.

[13] ZHANG Z X, SCHOMAKER L. DTGAN: dual attention generative adversarial networks for text-to-image generation[DB/OL]. (2020-11-21)[2024-05-02].https://doi.org/10.48550/arXiv. 2011.02709.

[14] ZHANG Z X, SCHOMAKER L. DiverGAN: an efficient and effective single-stage framework for diverse text-to-image generation[J]. Neurocomputing, 2022, 473(7):182-198.

[15] YE S M, WANG H, TAN M K, et al. Recurrent affine transformation for text-to-image synthesis[J]. IEEE transactions on multimedia, 2023(26):462-473.

[16] ZHANG H, XU T, LI H S, et al. StackGAN: text to photo-realistic image synthesis with stacked generative adversarial networks[C/OL]//2017 IEEE International Conference on Computer Vision (ICCV).Piscataway:IEEE, 2017[2024-05-01].https://ieeexplore.ieee.org/document/8237891.

[17] ZHANG H, XU T, LI H S, 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.

[18] XU T, ZHANG P C, HUANG Q Y, et al. AttnGAN: fine-grained text to image generation with attentional generative adversarial networks[C/OL]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018[2024-05-16].https://ieeexplore.ieee.org/document/8578241.

[19] ZHU M F, PAN P B, CHEN W, et al. DM-GAN: dynamic memory generative adversarial networks for text-to-image synthesis[C/OL]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).Piscataway: IEEE, 2019[2024-06-11].https://ieeexplore.ieee.org/document/8954283.

[20] RUAN S L, ZHANG Y, ZHANG K, et al. DAE-GAN: dynamic aspect-aware GAN for text-to-image synthesis[C/OL]// 2021 IEEE/CVF International Conference on Computer Vision (ICCV).Piscataway:IEEE,2021[2024-04-12].https://ieeexplore.ieee.org/document/9710042.

[21] TAN H C, LIU X P, YIN B C, et al. DR-GAN: distribution regularization for text-to-image generation[J]. IEEE transactions on neural networks and learning systems, 2022, 34(12): 10309-10323.

[22] WAH C, BRANSON S B, WELINDER P, et al. The caltech-UCSD birds-200-2011 dataset[J/OL]. Computer science,2011[2024-06-19].https://www.semanticscholar.org/paper/The-Caltech-UCSD-Birds-200-2011-Dataset-Wah-Branson/c069629a51f6c1c301eb20ed77bc6b586c24ce32.

[23] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[DB/OL].(2015-02-21)[2024-03-19]. https://doi.org/10.48550/arXiv.1405.0312.

[24] JIA J Y, CAO X Y, WANG B H, et al. Certified robustness for top-k predictions against adversarial perturbations via randomized smoothing[DB/OL]. (2019-12-20)[2024-06-11]. https://doi.org/10.48550/arXiv.1912.09899.

【作者简介】

兰才俊（1999—），男，福建龙岩人，硕士研究生，研究方向：深度学习、图像处理等。

姚剑敏（1978—），男，福建莆田人，博士，副研究员，研究方向：人工智能、图像处理、计算机视觉等。

（收稿日期：2024-10-14）