基于GAN网络的目标图像生成方法综述

王培龙,苗 壮,王家宝,李 阳,李允臣

(陆军工程大学 指挥控制工程学院,江苏 南京 210007)

摘 要: 生成对抗网络自2014年被提出以来,极大地推动了图像生成研究的进展。其通过两个神经网络的相互博弈,逐步提高鉴别真实图像与生成图像的能力,以及生成逼真图像的能力,最终使双方达到一种纳什均衡。简要介绍生成对抗网络,并围绕生成包含特定对象的图像这一问题对该网络在图像生成领域中的应用方法进行梳理,将其分为直接法、迭代法、分层法、解耦法和3D建模法5种类别。重点关注生成对抗网络在生成包含特定对象的图像方面的研究进展,并对该领域的发展方向进行展望,以期为相关人员进行图像生成研究提供参考。

关键词: 图像生成;生成对抗性网络;目标图像;解耦;人工智能

DOI: 10. 11907/rjdk. 231827 **开放科学(资源服务)标识码(OSID):**

中图分类号: TP391 **文献标识码:** A **文章编号:** 1672-7800(2024)009-0010-10

Overview of Target Image Generation Methods Based on GAN Networks

WANG Peilong, MIAO Zhuang, WANG Jiabao, LI Yang, LI Yunchen

(*Command and Control Engineering College*, *Army Engineering University of PLA*, *Nanjing* 210007, *China*)

Abstract: Since its proposal in 2014, generative adversarial networks have greatly promoted the progress of image generation research. Through the mutual game between two neural networks, it gradually improves the ability to distinguish between real images and generate images, as well as the ability to generate realistic images, ultimately achieving a Nash equilibrium between the two parties. Briefly introduce the generation of adversarial networks and sort out their application methods in the field of image generation around the issue of generating images containing targets. They are divided into five categories: direct method, iterative method, hierarchical method, decoupling method, and 3D modeling method. Focus on the research progress of generative adversarial networks in generating images containing targets, and prospect the development direction of image generation methods containing objects, in order to provide reference for relevant researchers in image generation research.

Key Words: image generation; generative adversarial network; target image; decoupling; artificial intelligence

0 引言

图像生成一直是计算机视觉领域一个重要研究方向,典型应用包括数据增广[1]、动画生成[2-3]、人脸生成等[4-5]。传统图像生成方法依赖于对数据分布进行显式建模[6],如经典的变分自动编码器[7-8](Variational Auto Encoder, VAE),其采用无监督学习的方式将高维数据映射为低维特征,然后从低维特征学习重建原始数据。该方法可以稳定地合成图像,但是质量不高[6]。2014年,Goodfellow等[9]受到零和博弈思想的启发,提出生成对抗网络(Generative Adversarial Network, GAN),其以非监督学习方式使两个子网络通过相互博弈,达到一种纳什均衡状态。GAN在数据生成方面取得了良好效果,一经提出便迅速成为该领域的研究热点,被誉为“过去20年来深度学习中最酷的想法”。

经过多年的发展,基于GAN的各种改进模型已经形成了一个庞大的家族,被广泛应用于图像编辑、图像修

收稿日期: 2023-07-27

基金项目: 江苏省自然科学基金项目(BK20200581)

作者简介: 王培龙(1993-),男,陆军工程大学指挥控制工程学院硕士研究生,研究方向为人工智能、图像处理;苗壮(1976-),男,博士,陆军工程大学指挥控制工程学院教授、博士生导师,研究方向为图像视频处理;王家宝(1985-),男,博士,陆军工程大学指挥控制工程学院副教授,研究方向为计算机视觉、图像处理;李阳(1984-),男,博士,陆军工程大学指挥控制工程学院副教授、硕士生导师,研究方向为机器视觉、机器学习;李允臣(1989-),男,陆军工程大学指挥控制工程学院硕士研究生,研究方向为人工智能、图像处理。本文通讯作者:苗壮。

复[10-12]、风格迁移[13]、目标检测等领域。现有基于GAN的图像生成方法主要聚焦于生成风景图像或纯对象图像，如海洋、草原、山川，或人脸、猫脸等，而生成一个含有人、车辆等特定对象的场景图像仍面临较大困难。这是由于生成一幅自然风景图像时，即使出现一些缺乏语义的部分也不会影响整体图像质量，人眼难以察觉其中的不合理性。而对于一张含有特定对象的图像，如车辆在公路上行驶，如果其中含有不合理信息，人们则很容易察觉。为此，本文简要介绍GAN的基本情况，重点对其应用于生成包含目标图像的研究进展进行综述，总结可行途径及存在的问题，以期帮助研究者更好地聚焦该领域的重难点问题。

1　GAN概述

1.1　GAN结构

如图1所示，一个典型的GAN包含两个结构主体，分别为生成器（Generator，G）和判别器（Discriminator，D）。G获得一组随机噪声，并输出生成数据$G(z)$；将生成数据$G(z)$和真实样本y分别输入D，D会输出$G(z)$和y的真实度概率$P(G(z))$和$P(y)$，概率接近1时输入被判别为真，即被认为是真实数据；否则被判别为假，即被认为是生成数据。一个训练理想的D的输出满足以下公式：

$$D(x)=\begin{cases}1, & x=y\\ 0, & x=G(z)\end{cases} \tag{1}$$

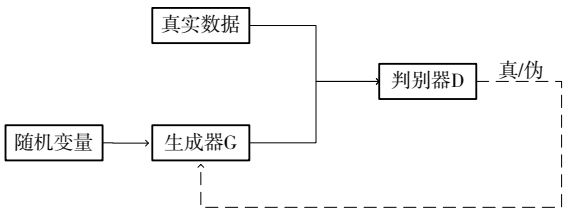

Fig. 1　GAN structure

图 1　GAN结构

同样的，一个训练理想的G可以确保$P(G(z))$恒为1。在实际训练中，应当避免出现一个过于强大的G或D，否则会导致无论生成数据是否合理，均会被判别为真的现象。理想结果为双方达到一种纳什均衡状态，此时G达到最优，而D的输出为：

$$D(x)=0.5, 其中 x=y, G(z) \tag{2}$$

1.2　网络训练

GAN需要G和D两个模型不断对抗训练来迭代优化。当G的输出接近真实数据分布，而D的输出概率为0.5时，说明两者达到了最优结果，这种迭代优化过程使模型训练成为一个复杂问题。GAN的基本训练步骤见图2。具体步骤为：①从真实样本中采样m个样本，从先验分布噪声中采集n个噪声样本并通过生成器获取n个生成样本。固定G的参数，训练D，使其尽可能准确地判别真实样本和生成样本；②循环k次更新D之后固定D的参数，令G用较小的学习率训练，目标是尽可能减小生成样本与真实样本的距离，即尽量使得D判别错误；③多次更新迭代后，最终理想情况为D对任意样本的输出概率均为0.5（纳什均衡），即D无法判断样本是来自于G的输出还是真实输出。

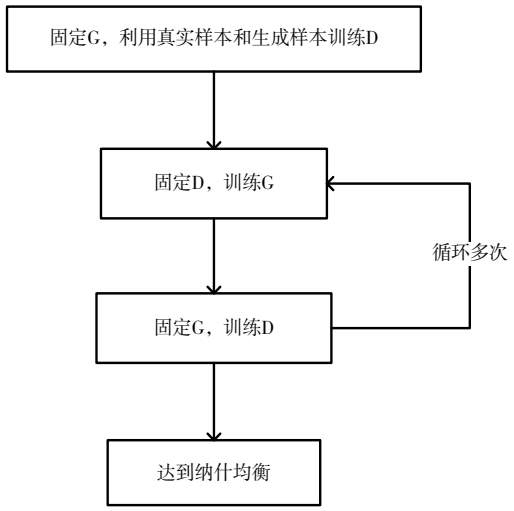

Fig. 2　Training process of GAN

图 2　GAN训练步骤

1.3　损失函数设计

由于GAN的训练需要G和D交替进行，如何设计损失函数指导这一过程十分关键。最初的GAN中主要通过引入对抗性损失实现这一目标，具体形式为：

$$\min_G \max_D V(D,G)=\begin{aligned}&E_{x\sim p_{data}(y)}\left[\log D(y)\right]\\&+E_{z\sim p_z(z)}\left[\log\left(1-D(G(z))\right)\right]\end{aligned} \tag{3}$$

式中$V(D,G)$为一个二分类的交叉熵函数，可以获得生成数据和真实数据分布情况的KL（Kullback-Leibler）散度；P_{data}和P_z分别为真实数据分布和生成数据分布。根据对抗性损失函数可知，GAN的训练目标为实现生成网络目标函数最小化，而判别网络目标函数最大化。

现有研究已提出一些新的损失函数设计思路。例如Zhu等[14]提出的CycleGAN模型使用循环一致性损失，模型中存在两组对称的G和D，输入一个图像可以按需求生成一个新的图像，也可以将新的图像输入返回成一个复原的原图像，通过衡量真实原图像与生成图像之间的差异来指导模型训练；再如CGAN[15]（Conditional Generative Adversarial Nets）、StarGAN[16-17]、ACGAN[18]（Auxiliary Classifier GANs）等模型中增加了分类器，将生成图像输入分类器并获得预测结果，利用预测结果与真实值的差异来优化模型。

1.4　GAN的发展

自GAN诞生后，相关研究十分丰富。原始GAN模型存在训练不稳定、模式崩溃、梯度消失等问题，研究人员设计了许多方法解决这些问题，逐步提高GAN的性能。此外，原始GAN模型依赖一组随机噪声作为输入，后续研究逐渐拓展了初始输入的范畴，包括由向量生成图像、由图像生成图像、由文本生成图像、由图像生成视频、由视频生

成视频等[6]。GAN模型亦在自然语言领域得到应用，如文本建模、对话生成等。

许多研究者针对GAN的局限性进行了改进。例如，Mirza等[15]提出的CGAN将条件向量c加入随机噪声，使得生成过程变得可控，c可以是目标数据的类别、属性等。然而，CGAN需要在有监督的条件下进行训练，而且训练难度大，图像生成质量较低；Mao等[19]分析了GAN训练不稳定的原因，采用最小二乘损失函数替代传统GAN中的交叉熵损失函数，从而得到LSGAN（Least Squares Generative Adversarial Networks）模型，在缓解训练不稳定问题的同时提高了图像生成质量，但是由于过于关注离散群点，使得生成数据的多样性降低，容易发生模式崩溃或梯度消失现象；Arijovsky等[20]提出的WGAN（Wasserstein GANs）模型使用EM距离（Earth-Mover Distance）取代JS散度（Jensen-Shannon Divergence），提高了GAN模型的训练稳定性，亦改善了模式崩塌现象，同时克服了真实数据与生成数据缺少交集的问题，但其存在难以收敛的问题；Salimans等[21]提出的WGAN-GP（WGAN-Gradient Penalty）模型设计了梯度惩罚损失，使WGAN可以更稳定地进行训练，缺点是增加了训练的计算复杂度，需要耗费更长时间，且生成样本缺乏多样性。

以上方法通过不断改进GAN模型的损失函数和训练方法优化了G与D之间的对抗博弈，使模型训练更加稳定，但仍无法生成高分辨率图像，且生成图像的真实性和多样性均较低。

2 包含特定对象的图像生成方法

基于生成过程中对输入噪声的不同处理方式，包含特定对象的图像生成方法可分为直接法、迭代法、分层法、解耦法和3D建模法五大类。各类方法的图像生成流程如图3所示。

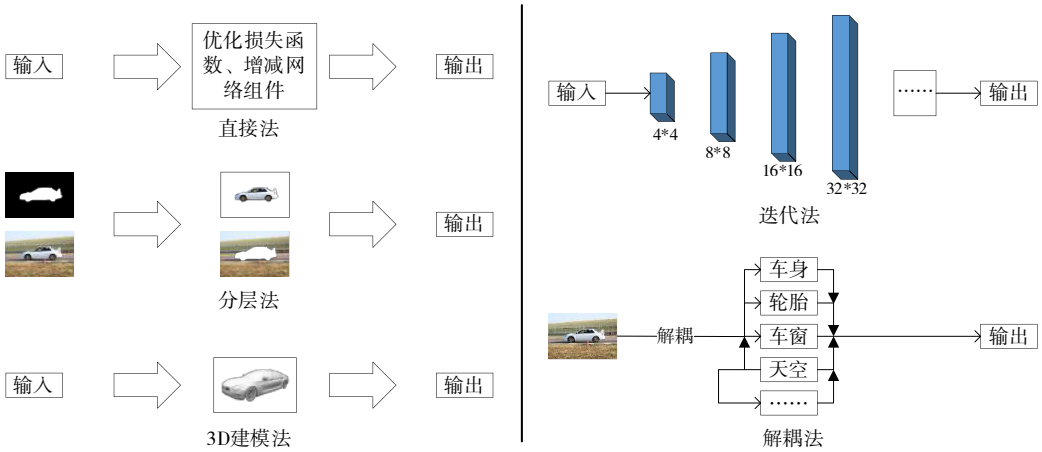

Fig. 3 Image generation process for various methods

图3 各类方法图像生成流程

2.1 直接法

如果一个模型在生成过程中仅依靠一组输入，然后通过G将输入直接映射到目标图像中，那么这种模型属于直接法。早期的GAN模型多基于一组G和D构造，有些对损失函数进行优化调整，有些引入深度学习中的一些模型结构，这些方法均属于直接法，其改进方式多为优化损失函数、调整个别组件功能。例如，Choi等[16-17]提出的StarGAN仅采用一对G和D即可在不同数据集上学习多种特征，例如从CelebA数据集中学习人脸的外表特征，从RaFD数据集中学习表情特征，最终可以生成具有目标外观和表情的人脸图像。StarGAN在传统D的基础上增加了一个分类器，使模型在训练中不仅能学到特征，还能将特征与对应标签联系起来，在生成时可以有目的地选择使用。其同时还将分类预测设计为损失函数，使用真实图像的分布分类损失约束D，虚假图像的分布分类损失约束G；Odena等[18]提出的ACGAN通过对G添加辅助分类标签的方式提升生成图像质量，不仅可以提高生成图像的分辨率，而且能减轻模式崩溃问题；Radford等[22]提出的DCGAN（Deep Convolutional GAN）将卷积神经网络引入GAN模型，鉴于卷积神经网络强大的图像处理能力，DCGAN得以生成更好的图像数据，并且在一定程度上缓解了训练不稳定的情况，因而成为图像生成领域的一个经典基本模型。原始GAN模型无法对生成过程进行控制，其输入是随机的，为更精细地控制输入条件，需要对目标属性进行多维描述。为此，Chen等[23]提出的InfoGAN（Information Maximizing GAN）采用无监督方式学习，基于信息论原理，通过最大化输入噪声与观察值之间互信息的方式对网络模型进行优化。以MNIST数据集为例，InfoGAN可通过输入隐式信息控制数字旋转与字符宽度，生成包含隐式信息的新数据集；Reed等[24]将卷积神经网络与GAN结合起来进行无监督学习，采用一组文本描述生成目标图像（Text to Image）。为了得到视觉上可以判别的文本表示，该模型利用卷积神经网络和循环文本编码器根据图像集学习一个对应的函数，文本分类器通过以下结构损失函数进行训练：

$$\frac{1}{N}\sum_{n=1}^{N}\Delta\left(y_n,f_v v_n\right)+\Delta\left(y_n,f_t\left(t_n\right)\right) \tag{4}$$

式中：v_n、t_n、y_n分别为图像、文本、描述类标签；f_v和f_t分别为图像和文本分类器。

Zhang等[25]提出的StackGAN++可根据一组文字描述生成符合要求的图像。该研究认为生成过程中不同分辨率图像的细节部分会存在差异，因此将不同分辨率下图像结构和颜色的差异信息纳入损失函数；Brock等[26]提出的BigGAN使图像生成精度得到极大提升，在128、256、512分辨率下均能生成高质量的自然场景图片。BigGAN对G进行正交正则化，使其服从于一个简单的截断技巧（Truncation Trick），通过减少G输入的方差来精细控制样本保真度与多样性之间的平衡。

此外，还有一些研究拓展了GAN模型的功能，使其除可以生成高质量图像外，还可以进行风格迁移。例如，Zhu等[14]提出的CycleGAN模型利用循环一致性损失将照片转换为油画风格的图像，或将冬季场景下的照片转换为夏季照片，还可以将照片中的斑马转换为马、苹果转换为橙子；Isola等[27]提出的Pix2pix GAN模型可将草图生成为逼真图像，非常适用于产品设计工作。该研究认为L1正则可使生成图像更清晰，而L2正则会导致图像模糊，因此在对抗损失的基础上添加了L1损失。表示为：

$$\mathcal{L}=\arg\min_{G}\max_{D}\mathcal{L}_{cGAN}(D,G)+\lambda\mathcal{L}_{L1}(G) \tag{5}$$

直接法仅利用一对G和D，将输入数据作为一个整体直接映射到目标图像，因而具有模型结构简单、训练速度快等优点。然而其较难生成高分辨率图像，且无法控制生成过程，容易出现训练不稳定的问题。

2.2　迭代法

早期的GAN网络大多只能生成一些低分辨率的图像[15]，如果想提高图像分辨率，则意味着扩大网络结构，模型将更难训练。因此，研究人员设想可以首先生成一个低分辨率图像，然后逐步提升其分辨率，这种方法称为迭代法。例如，Denton等[28]提出一种金字塔结构的LAPGAN（Laplacian Pyramid GAN）模型，其在生成图像时将随机噪声输入最底层生成器，对输出图像进行上采样可获得一个更高分辨率的图像；然后加入一组新的噪声后再输入下一层生成器，重复以上过程最终获得较高分辨率的图像。LAPGAN的训练过程则相反，首先将高分辨率的真实图像或生成图像输入上层D，然后逐步进行下采样获得更低分辨率的图像输入下层D。该方法既加快了训练速度，又极大稳定了训练过程；Huang等[29]提出的StackGAN是一种简单的双层迭代方法，第一层G的作用是生成外观轮廓和语义信息；第二层G需要第一层G的输入和输出，对图像细节信息进行完善，并获得最终图像；Karras等[30]提出一种渐进性提升分辨率的GAN模型，由低分辨率开始逐步增加G和D的层数，随着训练进展，模型获得越来越丰富的细节信息，最终生成高清图；Ma等[31]提出基于人物图像和姿势特征合成任意姿势人物图像的课题，其将人体特征解耦为姿势和纹理，生成过程分为位姿集成和图像细化两个关键阶段，使用两个串联的G，第一个生成一个较为模糊的图像，第二个生成一个差异图，两幅图像合并为最终结果。该方法实现了人体姿势的变换，但幅度不能太大，否则会导致图像生成失败；Abdal等[32]提出的StyleGAN使用渐进式G将人像特征映射到潜编码空间中，包括姿势、发型等粗糙特征，面部组成等中等特征以及颜色、纹理等细节特征，然后将这些潜编码输入生成网络以实现更精确的控制。鉴于StyleGAN优异的生成效果，许多研究人员在其基础上设计新的GAN网络[33-39]，优化其原始图像映射到潜空间的编码方式。为解决使用一张图像训练一个生成模型的问题，SinGAN[40]、ExSinGAN[41]借鉴金字塔式GAN模型结构，每一层由一对G和D组成，生成图像的分辨率逐层增加。每一层G的输入为上一层输出的上采样和一组随机噪声，其输出由同层的D进行判定，可以生成高质量、多样化的图像。

迭代法通常需要设计一系列结构相似的G和D，由粗到细地生成图像。该类方法可以生成分辨率更高的图像，但多层网络也使得模型结构更加复杂，需要更高的训练成本。

2.3　分层法

一些研究人员认为直接生成一整张图像的难度较大，可以将图像分割成两个甚至更多小的部分分别生成，然后将各个部分拼结到一起，作为最终生成图像，这种方法称为分层法。例如，Singh等[42]提出的FineGAN将图像分成背景、轮廓和细节3层，首先利用检测器提取出背景信息，基于此生成一张背景图；然后生成轮廓图像及对应的蒙版，将对象轮廓加入背景图中；最后生成色彩纹理及对应的蒙版，并缝合到图像中得到最终生成图像；Abdal等[43]提出Labels4Free技术，其认为图像的前景和背景是独立的，可以分别进行生产然后组合在一起。该方法在StyleGAN2[16]的基础上增加了一个蒙版生成器，蒙版包含区分前景与背景的信息；Yeo等[44]借鉴区域自适应归一化的方法提出自像素归一化，并运用于GAN模型中。自像素归一化可以生成一个自隐蒙版，将输入特征映射划分为前景和背景区域，然后从自隐蒙版中推断出每个区域的固有变换参数。还有一种思路是将图像分割成不同部分，针对每一部分分别进行生成训练。这种方法借鉴了人的视觉特性，即同一时刻人眼只能关注一小块区域，然后在大脑中将各个区域组合成一个整体。例如，Lin等[45]提出的COCO-GAN将人脸分成多个微块分别进行生成，然后拼接成一副完整的图像。其借鉴了ACGAN[18]的结构，在鉴别中增加了一个分类器，用于识别每个微块的位置，从而确保生成的每一微块都可以与其相邻微块无缝拼接；Mu等[46]提出的CoordGAN引入一个新的坐标空间，将原图中规范的二维坐标数据通过密集对应的方式转换为扭曲的坐标，

并将结构与纹理分开进行处理。

分层法可以分别生成目标图像的不同区域，其难点在于如何准确找到分层的边界，以及如何分辨前景与后景的边界。解决方法之一是将原图像及其蒙版信息一并输入，该法需要额外信息，增加了人工成本；另一种方法是通过训练使网络自行学得分层方法，该法需要带有蒙版标注的训练集，而且分层的边界难以做到十分精确。分层法可使模型更好地学习不同部分的特征，适用于生成前景与后景差异较大的图像。

2.4 解耦法

为提高神经网络学习的可解释性，使其能充分挖掘出数据中所蕴含的逻辑关系，研究者们尝试对数据中不同类别的信息进行解耦，包括图像的形状、纹理、外观、类别等，依据这种思想设计的GAN方法称为解耦法[47-48]。解耦法起源于独立成分分析（Independent Component Algorithm，ICA）法，其将一个信号分解为多个独立信号的叠加[48]。表示为：

$$x = w_1 z_1 + w_2 z_2 + L + w_n z_n = W_z \tag{6}$$

式中：x为实际信号，w_i为权重，z_i为独立信号。

解耦思想被引入到深度学习中，称为解耦表征学习[48]，旨在按照人类能够理解的方式从真实数据中对具有明确物理含义的生成因子进行解耦，并给出其所对应的独立潜在表示。

例如，前文所述的InfoGAN[23]最先引入解耦思想；Ma等[49]为实现对人体姿势转换生成过程的灵活控制，将人物图像简化成前景、背景与姿势的组合，分两阶段进行生成。阶段一从原始图像中获取前景、背景和姿势3类特征，阶段二提出一种新颖的两步映射技术用于对抗嵌入特征学习；Men等[50]提出一种属性分解方法AD-GAN（Attribute-Decomposed GAN），其以Unet为基本框架，在潜层特征空间中加入不同属性，以生成具有不同属性的人物图像。该模型首先通过源域语义图的语义布局自动将组件属性从源域人物图像中分离出来，该语义布局由预先训练的人物解析器提取；组件布局通过多分支嵌入到全局纹理编码器中，其潜在编码按照特定顺序重新组合以构造样式代码；级联样式块充当两个路径的连接，通过控制AdaIN层的仿射变换参数将样式编码表示的组件属性注入到姿势代码中；Zhang等[51]提出的PISE（Person Image Synthesis and Editing）模型基于解耦GAN将人体图像分解成姿势特征和样式特征，在生成过程中首先根据目标人体姿势生成一幅特征贴图，该贴图显示了目标姿势人体不同部位所需的样式信息；然后从原始图像中提取不同样式，将其填入特征图中后得到目标图像。该方法可灵活控制人体的不同姿势、不同部位，这也使其不利于自动生成大量数据；Liu等[52]提出的DynaST模型结合Transformer网络，利用其分层的结构特点设计了新的动态注意模块，以级联的方式确定每一个位置的tokens数量，即动态调整注意力分配，在生成图像细节方面取得了较好效果；Zhou等[53]提出一种将多模态信息映射到预训练样式的公共潜空间GAN反转技术Ti-GAN，该技术基于一个统一的框架将目标的不同属性解耦到不同大小的层中，在生成过程中将文本和图像引导到统一框架下，以实现对它们的同步操作，同时允许用户通过调整文本控制图像生成；Revanur等[54]提出CoralStyleCLIP模型，其在StyleGAN中融合多层注意力，兼顾了编辑保真度和简单性。该方法通过适当编辑共同学习W+空间中的全局方向，将每一层的预测兴趣区域限制为预训练分割网络中的空间段，大大降低了学习的复杂性。然而该方法将每一层的兴趣区域都传入网络预测的潜编码空间中，带来了更大的训练量。

解耦法通过探索图像数据内部蕴含的逻辑关系掌握其具备的特征因子，更好地模拟了人类观察图像时的联系、推理、想象机制，使图像生成从潜空间学习走向显式学习，为灵活控制生成过程提供了条件。解耦法的难点为在无监督条件下从原始图像中学习不同的特征因子，并且在生成中更好地实现多样性，比较适用于具有显著结构化特征的图像生成。

2.5 3D建模法

照片是将三维场景映像为二维图像。对于图像生成来说，通过建立三维场景实现二维图像生成的方法称为3D建模法。例如，Saito等[55]提出PIFu（Pixel-aligned Implicit Function）模型基于端到端的学习方法输入人体图像及相应的蒙版图片、相机参数，可以预测出人体表面每个点的三维坐标以及该点的像素值，进而通过点云方式重建人体的三维模型及对应纹理。但该方法对蒙版和相机参数要求较高，对拍摄距离和视角亦有严格限制；Mildenhall等[56]在已知视角下对场景进行一系列捕获以合成新视角下的图像，通过神经辐射场（Neural Radiance Fields，NeRF）隐式表达三维场景，利用多层感知器进行学习，以空间中点的体积密度和有向颜色值表示场景并对其进行渲染，从而输出任何角度的照片。然而该方法中每个场景必须单独优化，场景之间没有知识共享，在单个或极稀疏视图的限制下，其无法利用任何先验知识加速图像重建；Wang等[57]在NormalGAN模型中引入法线贴图（Normal Maps），输入单个前视图RGB-D图像后，模型可以利用前视图推断出后视图，然后结合两个视图信息生成3D模型。该方法有较好的深度去噪能力，可以高精度地复原模型，但其对数据采集设备要求较高，需要单目深度相机；Aliakbarian等[58]设计了一种基于条件归一化流的方法FLAG（Flow-based 3D Avatar Generation），该方法可以学习给定头部和手部数据的全身姿势条件，实现三维分布到普通分布的逆映射，并且在普通分布中进一步学习更高相似区域的概率映射，通过这种概率映射，模型可以观察到潜在空间编码方式，从而合理预测生成人体的姿势；Foti等[59]采用自监督方法训练三维变分自动编码器，利用潜编码中的已知差异和相似性

定义了一个新的损失函数，通过在不同形状之间交换任意特征来管理小批量图像生成。该方法可以解耦面部和身体的身份特征，同时保持了良好的表示和重建能力。

3D建模法模拟了现实情况下获取同一目标不同图像的方式，其在完整复原3D模型的情况下可以获取任意视角下的2D图像。该方法的难度在于需要额外信息复原3D模型，如不同角度的图像，而且较难获取数据集用于模型训练，难以改变对象姿态。

以上5类方法的典型代表、发布年份、基本思想以及优缺点总结如表1所示。

Table 1　Comparison of 5 kinds of image generation methods

表 1　5类图像生成方法比较

类别	典型代表(按照发布年份排序)	基本思想	优点	缺点
直接法	DCGAN[22](2016)、InfoGAN[23](2016)、ACGAN[18](2017)、CycleGAN[14](2017)、StarGAN[16-17](2018)、BigGAN[26](2018)、Text-to-image[24](2023)	将G作为一个映射函数，直接将一组输入数据映射为目标图像	网络结构简单、模型训练相对容易	图像生成质量较低
迭代法	LAPGAN[28](2015)、StackGAN[29](2017)、ProGAN[30](2017)、PG2[31](2018)、SinGAN[40](2021)、ExSinGAN[41](2022)	先生成一个低分辨率的图像，然后逐步提升其分辨率	生成图像分辨率较高	网络结构较为复杂，训练成本较高
分层法	FineGAN[42](2019)、COCO-GAN[45](2020)、Labels4Free[43](2021)、SPNGAN[44](2021)、CoordGAN[46](2022)	将图像分成两个甚至更多部分分别进行生成，然后将各个部分拼合或融合到一起	模型可以针对不同区域进行差异化学习，使图像更逼真	缺乏带有蒙版的训练集
解耦法	Attribute-Decomposed GAN[50](2020)、PISE[51](2021)、TiGAN [53](2022)、DynaST[52](2022)、CoralStyleCLIP[54](2023)	将图像中的特征引子加以分解，生成过程中根据需要分别对其进行调整	便于对生成图像的细节进行控制；可以学到图像蕴含的深层特征信息	对对象进行特征解耦难度较大、生成图像多样性较低
3D建模法	PIFu[55](2019)、NeRF[56](2020)、NormalGAN[57](2020)、FLAG[58] (2022)、VALDGAN[59](2022)	将2D图像还原为3D表示，然后根据需要获取2D图像	真实性较好，最符合现实场景中不同视角照片的拍摄方式	需要输入2.5D图像或同一物体不同视角的一组图像；难以改变对象形状

3　包含对象的数据集

用于图像生成的数据集数量众多，以下介绍几个较为典型的包含特定对象的图像数据集：①LSUN[60]。其是一个大规模的图像数据集，用于训练模型进行场景理解。该数据集包含超过900万张图像，按照场景划分为人物、车辆和骑行车等类别；②Imagenet[61]。其是一个巨大的可供图像/视觉训练的图片库，是评估图像分类算法性能的基准。其按照WorldNet结构组织，带有标签，包括2.2万个类别、约1 500万张图片；③CUB[62]。其由加州理工学院创建，主要用于细粒度分类识别研究，包含200种不同类别、11 788张鸟类图像数据；④CIFAR-10[63]。其是一个用于识别普适物体的小型数据集，包含飞机、汽车、鸟类、猫类、鹿类、狗类、蛙类、马类、船舶和卡车等10个类别的60 000张RGB彩色图片，图片尺寸为32×32。由于数据集中包含现实世界中的物体，不仅噪声很大，而且比例、特征不尽相同；⑤Stanford-cars[64]。其是覆盖多种规格（轿车、轿跑车、敞篷车、两厢车和货车等）、不同拍摄角度、不同分辨率等细粒度的车辆数据集，包含196类、16 185张汽车图像，分为8 144张训练图像和8 041张测试图像。该数据集主要基于汽车品牌、车型以及年份进行划分。

4　图像生成评价方法

图像生成模型性能评价是一项复杂的工作。一个简单的GAN可以看作一个映射函数，其输入为一组低维空间随机噪声，输出为在高维空间中均匀分布的数据。该研究领域较为关注生成图像的多样性、真实性以及可控度。多样性是指生成的图像均匀分布，而不是生成很多非常相似的图像。真实性是指生成的图像与现实场景一样，尤其是要使人眼无法分辨真假。在实际研究过程中经常会出现生成图像分布较为均匀，多样性很好，但图像缺乏语义信息，看上去很假；或是图像非常逼真，但却高度相似的现象。可控度是指可以灵活控制图像生成过程。总体来说，一个好的评价指标给出的结果应与人类感知一致，因此很难用一个合适的指标评价所有图像生成模型。目前常用评估方法包括定性分析（Qualitative Evaluation）和定量分析两种（Quantitative Evaluation）[1]。

4.1　定性分析

定性分析是对模型生成图像的真实性、多样性等进行人工评价，这种方法往往依赖于人的主观感受，常用方法包括比较真实图片与生成图片、比较不同模型生成的相同目标图片等。一些研究在进行定性分析时通过第三方平台（如Amazon Mechanical Turk论坛）将评测任务外包出去，将生成图像与真实图像混合在一起，由评测人员对照片的多样性和真实性进行打分，研究人员可以根据分数定性判断生成模型的性能，一些经典研究均采用了该种方式。根据定性分析情况来看，图像生成模型的性能正在逐渐提高[14,17,40,65]。

4.2　定量分析

定量分析常用指标包括IS（Inception Score）和FID（Fréchet Inception Distance）等[66]。

4.2.1 IS

IS被广泛应用于评价生成模型质量，其可同时衡量生成图像的真实性和多样性。IS基于类概率分布，利用预训练模型计算生成图像的分数。以分类网络Inception V3为例，其在ImageNet上经过预训练，将输入的生成图像x输出为$p(y_i|x)$。这是一个多维向量，该向量的不同维度代表其属于某一类别的概率。当生成的单个图像质量越高时，分类器将以较高的置信度对其进行分类，即在对应维度具有较高分数，而在其他维度具有较低分数。因此，度量生成图像真实性的公式为：

$$H(y|x) = -\sum_{i=1} p(y_i|x)\log\left[p(y_i|x)\right] \tag{7}$$

假设向分类网络输入一组数据$\{x_1, x_2, x_3\}$，它们具有较好的多样性，则其输出$\{y_1, y_2, y_3\}$应当均匀分布，即不同的y应在不同维度上均取得较高分数。则多样性的衡量公式为：

$$H(y) = -\sum_{i=1}^{n} p(y_i)\log\left[p(y_i)\right] \tag{8}$$

综上所述，IS的计算公式为：

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

IS可以有效反映出生成图片的质量。其缺点是分数计算容易受到数值样本选取的影响，当数据集的内部差异性较大时，不建议选择该标准。此外，当模型出现过拟合情况时，该指标无法对图像进行区分。目前，一些研究对IS进行了改进，如Mode Score[67]，其在IS的基础上结合数据集的标签信息，尽可能地降低生成图片与数据集中图片标签y分布的KL散度；又如M-IS（Modified Inception Score）解决了类内模式崩溃问题，将类内交叉熵引入IS，可以评价生成图像的质量和类内多样性；再如AMS（AM Score）考虑到数据在各类别中的分布并不均匀，采用训练数据集类别分布与生成数据类别分布的KL散度评价，AMS越小，模型性能越好。

4.2.2 FID

FID关注生成图像与真实图像之间的联系，亦通过Inception模型进行计算。其在Inception模型的基础上去掉最后一个全连接层，从而得到一组D维向量。当输入N张生成图像和N张真实图像后可得到一组$N\times D$维的向量。计算两个$N\times D$维向量的距离便可得到FID值。公式为：

$$FID(g,r) = \left\|\mu_g - \mu_r\right\|_2^2 + T_r\left(\sum g + \sum r - 2\sum g\sum r^{\frac{1}{2}}\right)) \tag{10}$$

式中：g和r分别表示生成图像和真实图像，μ表示均值，T_r表示矩阵的迹。

FID表示生成图像与真实图像特征向量之间的距离，该距离越近，则模型性能越好。FID具有较高的鲁棒性，即使存在一些噪声也可有效评价生成效果，且其数值接近于人类的主观评价。

5 困难与挑战

基于GAN的图像生成方法通常会面临训练不稳定、模式崩溃、梯度消失等问题，且生成包含特定对象的图像比生成风景、纹理图像面临更大挑战：一是如何生成有充足语义信息的图像，即对象应是真实的、符合常理的；二是如何既使对象与背景的边界清晰明确，又使两者恰当地融合。

5.1 训练不稳定

GAN的训练目的是使模型通过对抗博弈达到纳什均衡状态。在这种博弈状态下，D性能的上升将导致G性能下降，反之亦然，因而基于梯度下降的优化算法不一定能获得收敛结果，博弈容易出现一边倒的情况。例如一个过于强大的D总能准确分辨出G生成的数据，这使得G得不到正反馈，性能便无法得到提高。此外，即使两者达到均衡状态，也可能出现G生成的数据没有现实意义的情况。

5.2 模式崩溃

在GAN中，G的目的是生成更加接近真实的数据，而D的目的则是评判输入数据的真假，G的优化主要来源于D的判断，这导致在训练中容易出现这样一种情况，即G发现某一类型的输出易于被D判断为真，从而强化了G生成这一类型输出的能力，最终导致G生成的数据集中于某一类型，缺乏多样性，即模式崩溃现象。这个问题可以归因于KL散度的不对称性。

5.3 梯度消失

在训练GAN的过程中，G的迭代提升依赖于D的反馈信息。训练所用真实数据集往往呈现高维分布，而G容易生成低维分布的数据，这导致输入D的数据中真实数据与生成数据差异过大，D每次都可以准确区分出真假，从而不能反馈有效信息，导致G的梯度消失。另一个导致梯度消失的原因为D一开始并不能准确判断输入数据的真假，因而反馈给G无效甚至是错误的信息，使得G得不到有效训练。

5.4 生成充足的语义信息

生成包含特定对象图像难度大的一个重要原因是其包含更多语义信息。在风景图像中，即使出现一些形变、色差也不会引起人眼的明显感觉，但如果图像中含有人物、车辆或其他对象，即使是微小的瑕疵都容易被人察觉，例如出现了人们无法做出的动作，或汽车外观不符合常识。因此，如何确保生成图像中包含足够的语义信息给图像生成模型提出了更高要求。

5.5 合理区分对象与背景

如果是纯自然风光的图像，不同区域之间（如草原和山川、沙滩和海水）并不需要有清晰明确的分界线，它们之间可以有一些融合，从一个区域的边缘逐渐过渡到另一个区域。然而对于包含特定对象的图像，对象的边缘部分不

应该与背景有融合，两者有清晰的界线。此外，对象应当很好地融入场景之中，例如一个人站在户外，那他的身上应当有阳光照射；如果一个人站在雨中，而他的身上却很明亮，就会使图像的真实性大大降低。

6　趋势与展望

经过多年的发展，GAN在图像生成领域取得了显著进展，相关研究集中在优化损失函数、设计网络结构等方面，这些改进使模型表现出更好的性能，但仍未能从根本上解决模式崩溃、梯度消失等问题。结合目前GAN的发展趋势对包含特定对象的图像生成研究工作作出以下展望。

6.1　深入理论研究

制约图像生成方法发展的一个关键就是可解释性问题。当前图像生成研究主要基于深度网络理论，关于GAN的基础理论研究尚不完善。GAN的思想来源为博弈论中的纳什均衡，而目前并没有从理论上证明平衡点的存在，导致GAN系列模型模式崩溃、训练不稳定、梯度消失等问题迟迟不能解决。如果不能从理论上解决可解释性问题，并建立有效的数学模型，GAN将难以稳定地应用于实际场景中。今后可从GAN的博弈机制出发，将G和D的所有相关因素综合起来，形成一个统一的理论框架解决该问题。同时，GAN的训练依赖于G和D的交替优化，这种对抗性给训练带来很大困难。如何从理论层面分析这一训练过程也是今后研究的关键方向之一。

6.2　完善评估体系

评价生成模型的指标有很多，但目前为止缺乏科学、统一的标准。有研究[68]认为使用不同方法评估GAN模型可能会产生互相矛盾的结论，应根据应用场景选择不同的评估方法。设计一套从性能、真实性、多样性以及可控性等多方面综合评估GAN模型性能的指标体系是今后研究的重点方向。

6.3　优化模型结构

目前，GAN的一些研究趋向于多模态融合，如文本—图像、图像—视频、文本—视频等，这些方向对于开拓GAN的应用场景、实现落地很有帮助，但其多基于有监督或半监督的训练，需要成对的数据。例如，对于文本转图像工作来说，其需要每一幅图像带有一组文字描述的训练集，这是很难获得的。因此，探索仅需少量标签的训练方式或无监督的训练方式是一项有意义的工作。同时，一个可以应用于实际工作的GAN模型需要能灵活控制生成过程，现有模型多为一次性输入数据后直接获得生成图像，难以灵活编辑，因此设计可控的图像生成模型也是值得研究的问题。

6.4　开拓应用领域

目前，GAN在图像生成领域中的应用较多，今后可进一步拓展其应用范围。例如利用强化学习中的策略梯度算法，在离散场景下提升GAN的生成性能；利用GAN生成对抗样本，以解决深度学习系统的安全性问题。在计算机建模领域，GAN也有着广泛的应用前景，例如在虚拟现实领域，GAN可以生成特定的逼真场景；在游戏设计领域，GAN可以生成高分辨率的游戏场景、道具、任务形象等。

7　结语

自2014年被提出以来，GAN便成为人工智能领域的研究热点，在理论基础、实践方法和应用场景等方面成果丰硕。本文通过分析GAN的基本思想，对直接法、迭代法、分层法、解耦法、3D建模法5类基于GAN的图像生成方法的经典模型进行总结，并在此基础上对包含特定对象的图像生成发展方向进行展望，以期帮助相关研究者不断改进GAN的模型结构和训练方法，提高生成样本的质量，更好地应用于各行各业。

参考文献：

[1] MA D A，TANG P，ZHAO L J，et al. Review of data augmentation for image in deep learning [J]. Journal of Image and Graphics，2021，26(3)：487-502.
马崇昇，唐娉，赵理君，等. 深度学习图像数据增广方法研究综述[J]. 中国图象图形学报，2021，26(3)：487-502.

[2] ZHUO L，WANG G C，LI S K，et al. Fast-Vid2Vid：spatial-temporal compression for video-to-video synthesis [C]// Israel：European Conference Computer Vision，2022.

[3] HU X T，HUANG Z W，HUANG A L，et al. A dynamic multi-scale voxel flow network for video prediction[DB/OL]. https://doi. org/10. 48550/arXiv. 2303. 09875.

[4] TAN M K，XU S K，ZHANG S H，et al. A review on deep adversarial visual generation [J]. Journal of Image and Graphics，2021，26(12)：2751-2766.
谭明奎，许守恺，张书海，等. 深度对抗视觉生成综述[J]. 中国图象图形学报，2021，26(12)：2751-2766.

[5] QIU H B，YU B S，GONG D H，et al. SynFace：face recognition with synthetic data [C]//Montreal：International Conference on Computer Vision，2021.

[6] CHEN F J，ZHU F，WU Q X，et al. A survey about lmage generation with generative adversarial nets [J]. Chinese Journal of Computers，2021，44(2)：347-369.
陈佛计，朱枫，吴清潇，等. 生成对抗网络及其在图像生成中的应用研究综述[J]. 计算机学报，2021，44(2)：347-369.

[7] LARSEN A B L，SØNDERBY S K，LAROCHELLE H，et al. Autoencoding beyond pixels using a learned similarity metric [C]//New York：Proceedings of the 33nd International Conference on Machine Learning，2016：1558-1566.

[8] KINGMA D P，WELLING M. Auto-encoding variational Bayes [C]// Banff：2nd International Conference on Learning Representations，2014.

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

[10] YAO K，GAO P L，YANG X，et al. Outpainting by queries[C]//Israel：European Conference Computer Vision，2022.

[11] YU Y C, ZHAN F N, LU S J, et al. WaveFill: a wavelet-based generation network for image inpainting[C]// Montreal: International Conference on Computer Vision, 2021.

[12] WAN Z Y, ZHANG J B, CHEN D D, et al. High-fidelity pluralistic image completion with Transformers[C]// Montreal: International Conference on Computer Vision, 2021.

[13] XU W J, LONG C J, WANG R S, et al. Dynamic ResBlock generative adversarial network for artistic style transfer[C]// Montreal: International Conference on Computer Vision, 2021.

[14] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Venice: IEEE International Conference on Computer Vision, 2017: 2242-2251.

[15] MIRZA M, OSINDERO S. Conditional generative adversarial nets[DB/OL]. https://arxiv. org/abs/1411. 1784.

[16] CHOI Y, UH Y J, YOO J, et al. StarGAN v2: diverse image synthesis for multiple domains[C]//Seattle: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 8188-8197.

[17] CHOI Y, CHOI M, KIM M, et al. StarGAN: unified generative adversarial networks for multi-domain image-to-image translation[C]//Salt Lake City: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 8789-8797.

[18] ODENA A, OLAH C, SHLENS J. Conditional image synthesis with auxiliary classifier GANs[C]// Sydney: Proceedings of the 34th International Conference on Machine Learning, 2017: 2642-2651.

[19] MAO X D, LI Q, XIE H R, et al. Least squares generative adversarial networks [C]//Venice: International Conference on Computer Vision, 2017: 2813-2821.

[20] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein GAN[DB/OL]. https://arxiv. org/abs/1701. 07875.

[21] SALIMANS T, GOODFELLOW I J, ZAREMBA W, et al. Improved techniques for training GANs[C]// Barcelona: Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems, 2016: 2226-2234.

[22] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[C]//San Juan: 4th International Conference on Learning Representations, 2016.

[23] CHEN X, DUAN Y, HOUTHOOFT R, et al. InfoGAN: interpretable representation learning by information maximizing generative adversarial nets [C]//NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems, 2016: 2180-2188.

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

[25] ZHANG H, XU T, LI H S, et al. StackGAN++: realistic image synthesis with stacked generative adversarial networks[J]. Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(8): 1947-1962.

[26] BROCK A, DONAHUE J, SIMONYAN K. Large scale GAN training for high fidelity natural image synthesis[C]//New Orleans: 7th International Conference on Learning Representations, 2019.

[27] ISOLA P, ZHU J Y, ZHOU T H, et al. Image-to-image translation with conditional adversarial networks[C]//Honolulu: Conference on Computer Vision and Pattern Recognition, 2017.

[28] DENTON E L, CHINTALA S, FERGUS R, et al. Deep generative image models using Alaplacian pyramid of adversarial networks[C]//NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing System, 2015: 1486-1494.

[29] HUANG X, LI Y X, POURSAEED O, et al. Stacked generative adversarial networks[C]//Honolulu: Conference on Computer Vision and Pattern Recognition, 2017.

[30] KARRAS T, AILA T, LAINE S, et al. Progressive growing of GANs for improved quality, stability, and variation[C]// Vancouver: 6th International Conference on Learning Representations, 2018.

[31] MA L Q, JIA X, SUN Q R, et al. Pose guided person image generation [C]//NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 405-415.

[32] ABDAL R, ZHU P, MITRA N J, et al. StyleFlow: attribute-conditioned exploration of styleGAN-generated images using conditional continuous normalizing flows[J]. ACM Transactions on Graphics, 2021, 40(3): 1-21.

[33] PATASHNIK O, WU Z Z, SHECHTMAN E, et al. StyleCLIP: text-driven manipulation of styleGAN imagery[C]//Montreal: International Conference on Computer Vision, 2021: 2065-2074.

[34] WU Z Z, LISCHINSKI D, SHECHTMAN E. StyleSpace analysis: disentangled controls for styleGAN image generation[C]//Conference on Computer Vision and Pattern Recognition, 2021: 12863-12872.

[35] ALALUF Y, TOV O, MOKADY R, et al. HyperStyle: styleGAN inversion with hypernetworks for real image editing[C]//New Orleans: Conference on Computer Vision and Pattern Recognition, 2022: 18490-18500.

[36] TOV O, ALALUF Y, NITZAN Y, et al. Designing an encoder for StyleGAN image manipulation[J]. ACM Transactions on Graphics, 2021, 40(4): 1-14.

[37] RICHARDSON E, ALALUF Y, PATASHNIK O, et al. Encoding in style: a styleGAN encoder for image-to-image translation[C]//Conference on Computer Vision and Pattern Recognition, 2021: 2287-2296.

[38] KWON G, YE J C. Diagonal attention and style-based GAN for content-style disentanglement in image generation and translation[C]//Montreal: International Conference on Computer Vision, 2021: 13960-13969.

[39] ENDO Y. User-controllable latent Transformer for StyleGAN image layout editing[J]. Computer Graphics Forum, 2022, 41(7): 395-406.

[40] SHAHAM T R, DEKEL T, MICHAELI T. SinGAN: learning a generative model from a single natural image[C]//Seoul: International Conference on Computer Vision, 2019: 4569-4579.

[41] ZHANG Z C, HAN C Y, GUO T D. ExSinGAN: learning an explainable generative model from a single image[DB/OL]. https://arxiv. org/abs/2105. 07350.

[42] SINGH K K, OJHA U, LEE Y J. FineGAN: unsupervised hierarchical disentanglement for fine-grained object generation and discovery[C]// Long Beach: Conference on Computer Vision and Pattern Recognition, 2019: 6490-6499.

[43] ABDAL R, ZHU P H, MITRA N J, et al. Labels4free: unsupervised segmentation using styleGAN [C]// Montreal: International Conference on Computer Vision, 2021: 13950-13959.

[44] YEO Y J, SAGONG M C , PARK S, et al. Image generation with self pixel-wise normalization [J]. Applied Intelligence, 2022, 53: 9409-9423.

[45] LIN C H, CHANG C C, CHEN Y S, et al. Coco-GAN: generation by parts via conditional coordinating[C]// Seoul: International Conference on Computer Vision, 2019: 4511-4520.

[46] MU J, MELLO S D, YU Z D, et al. CoordGAN: self-supervised dense correspondences emerge from GANs [DB/OL]. https://arxiv. org/abs/2203. 16521.

[47] CHENG K Y, MENG C Y, WANG W S, et al. Research advances in dis-

entangled representation learning[J]. Journal of Computer Applications, 2021, 41(12): 3409-3418.
成科扬,孟春运,王文杉,等. 解耦表征学习研究进展[J]. 计算机应用, 2021, 41(12): 3409-3418.

[48] WEN Z D, WANG J R, WANG X X, et al. A review of disentangled representation learning [J]. Acta Automatica Sinica, 2022, 48 (2): 351-374.
文载道,王佳蕊,王小旭,等. 解耦表征学习综述[J]. 自动化学报, 2022, 48(2): 351-374.

[49] MA L Q, SUN Q R, GEORGOULIS S, et al. Disentangled person image generation[C]// Salt Lake City : Conference on Computer Vision and Pattern Recognition, 2018: 99-108.

[50] MEN Y F, MAO Y M, JIANG Y N, et al. Controllable person image synthesis with attribute-decomposed GAN[C]// Seattle: Conference on Computer Vision and Pattern Recognition, 2020: 5083-5092.

[51] ZHANG J S, LI K, LAI Y K, et al. Pise: person image synthesis and editing with decoupled GAN[C]// Virtual : Conference on Computer Vision and Pattern Recognition, 2021: 7982-7990.

[52] LIU S H, YE J W, REN S C, et al. DynaST: dynamic sparse transformer for exemplar-guided image generation[C]//European Conference on Computer Vision, 2022: 72-90.

[53] ZHOU Y F, ZHANG R Y, GU J X, et al. TiGAN: text-based interactive image generation and manipulation[C]//Conference on Artificial Intelligence, 2022: 3580-3588.

[54] REVANUR A, BASU D, AGRAWAL S, et al. CoralStyleCLIP: co-optimized region and layer selection for image editing[DB/OL]. https://arxiv.org/abs/2303.05031.

[55] SAITO S, HUANG Z, NATSUME R, et al. Pifu: pixel-aligned implicit function for high-resolution clothed human digitization[C]//Seoul: International Conference on Computer Vision, 2019:2304-2314.

[56] MILDENHALL B, SRINIVASAN P P, TANCIK M, et al. Nerf: representing scenes as neural radiance fields for view synthesis[C]//European Conference on Computer Vision, 2020:405-421.

[57] WANG L Z, ZHAO X C, YU T, et al. NormalGAN: learning detailed 3d human from a single RGB-D image[C]// Glasgow: European Conference on Computer Vision, 2020:430-446.

[58] ALIAKBARIAN S, CAMERON P, BOGO F, et al. FLAG: flow-based 3D avatar generation from sparse observations[C]//New Orleans: Conference on Computer Vision and Pattern Recognition, 2022: 13243-13252.

[59] FOTI S, KOO B J, STOYANOV D, et al. 3D shape variational autoencoder latent disentanglement via mini-batch feature swapping for bodies and faces[C]// New Orleans: Conference on Computer Vision and Pattern Recognition, 2022: 18709-18718.

[60] YU F, ZHANG Y D, SONG S, et al. LSUN: construction of a large-scale image dataset using deep learning with humans in the loop[EB/OL]. https://www.yf.io/p/lsun.

[61] YANG K Y, QINAMI K, LI F F, et al. Towards fairer datasets: filtering and balancing the distribution of the people subtree in the ImageNet hierarchy[C]// Barcelona: Conference on Fairness, 2020:547-558.

[62] HORN G V, BRANSON S, FARRELL R, et al. Building a bird recognition App and large scale dataset with citizen scientists: the fine print in fine-grained dataset collection[C]//Boston: Conference on Computer Vision and Pattern Recognition, 2015: 595-604.

[63] KRIZHEVSKY A, HINTON G. Learning multiple layers of features from tiny images[R]. Toronto: University of Toronto, 2009.

[64] KRAUSE J, STARK M, DENG J, et al. 3D object representations for fine-grained categorization [C]// Sydney: International Conference on Computer Vision Workshops, 2013: 554-561.

[65] BARRATT S T, SHARMA R. A note on the inception score[DB/OL]. http://arxiv.org/abs/1801.01973.

[66] PARK T, LIU M Y, WANG T C, et al. Semantic image synthesis with spatially-adaptive normalization[C]// Long Beach: Conference on Computer Vision and Pattern Recognition, 2019: 2337-2346.

[67] CHE T, LI Y R, JACOB A P, et al. Mode regularized generative adversarial networks[C]//Toulon: International Conference on Learning Representations, 2017.

[68] THEIS L, OORD A V D, BETHGE M. A note on the evaluation of generative models[C]//San Juan: International Conference on Learning Representations, 2016.

(责任编辑:尹晨茹)