中图法分类号:TP391.7　　**文献标识码**:A　　**文章编号**:1006-8961(2021)12-2751-16

论文引用格式:Tan M K, Xu S K, Zhang S H and Chen Q. 2021. A review on deep adversarial visual generation. Journal of Image and Graphics, 26(12):2751-2766(谭明奎,许守恺,张书海,陈奇. 2021. 深度对抗视觉生成综述. 中国图象图形学报,26(12):2751-2766)[DOI:10.11834/jig.210252]

深度对抗视觉生成综述

谭明奎*,许守恺,张书海,陈奇

华南理工大学软件学院, 广州　510000

摘　要:　深度视觉生成是计算机视觉领域的热门方向,旨在使计算机能够根据输入数据自动生成预期的视觉内容。深度视觉生成使用人工智能技术赋能相关产业,推动产业自动化、智能化改革与转型。生成对抗网络(generative adversarial networks,GANs)是深度视觉生成的有效工具,近年来受到极大关注,成为快速发展的研究方向。GANs 能够接收多种模态的输入数据,包括噪声、图像、文本和视频,以对抗博弈的模式进行图像生成和视频生成,已成功应用于多项视觉生成任务。利用 GANs 实现真实的、多样化和可控的视觉生成具有重要的研究意义。本文对近年来深度对抗视觉生成的相关工作进行综述。首先介绍深度视觉生成背景及典型生成模型,然后根据深度对抗视觉生成的主流任务概述相关算法,总结深度对抗视觉生成目前面临的痛点问题,在此基础上分析深度对抗视觉生成的未来发展趋势。

关键词:深度学习;视觉生成;生成对抗网络(GANs);图像生成;视频生成;3 维深度图像生成;风格迁移;可控生成

A review on deep adversarial visual generation

Tan Mingkui*, Xu Shoukai, Zhang Shuhai, Chen Qi

School of Software Engineering, South China University of Technology, Guangzhou　510000, China

Abstract: Deep visual generation has aimed to create synthetic photo-realistic visual contents (such as images and videos) that could fool or please human perceptions according to some specific requirements. In fact, many human activities belong to the field of visual generation, e. g. , advertisement making, house designing and film making. However, these tasks normally can only be done by experts with professional skills gained through long-term training and the help of professional software such as Adobe Photoshop. Besides, it may also take a very long time to produce photo-realistic contents since the process can be very tedious and cumbersome. Thus, how to make these processes automated is a very important yet non-trivial problem. Nowadays, deep visual generation has become a significant research direction in computer vision and machine learning, and has been applied in many tasks, such as automatic content generation, beautification, rendering and data augmentation. Thanks to the current deep generative methods can be categorized into two groups: variational auto-encoder (VAE) based methods and generative adversarial networks (GANs) based methods. Based on encoder-decoder architecture, VAE methods first map input data into a latent distribution, and then minimize the distance between the latent distribution and some prior distribution, e. g. , Gaussian distribution. A well-trained VAE model could be used in the tasks of dimensionality reduction and image generation. However, an inevitable gap between the latent distribution and prior dis-

收稿日期:2021-04-07;**修回日期**:2021-07-16;**预印本日期**:2021-07-23

*通信作者:谭明奎　mingkuitan@scut.edu.cn

基金项目:国家自然科学基金项目(62072190);珠江人才计划项目(2017ZT07X183)

Supported by:National Natural Science Foundation of China (62072190)

tribution would make the generated images/videos blurred. Unlike the VAE model, GAN has learned a mapping between input and output distributions to synthesize sharper images/videos. A GAN model has contained two major modules. A generator has aimed to generate the fake data and a discriminator has distinguished whether a sample is fake or not. To produce plausible fake data, the generator has been matched the distribution of real data and synthesized fake data that would fulfill the requirements of reality and diversity. The optimization problem of learning the generator and discriminator has been formulated into a two-player minimax game. During the training, the two modules have been optimized alternately using stochastic gradient methods. At the end of the training, the generator and discriminator have been supposed to reach a Nash Equilibria of the minimax game. Due to the development of GAN model, more deep visual generation applications and tasks have occurred based on GAN model. The six typical tasks for deep visual generation have been presented as follows: 1) Image generation from noises: it is the earliest task of deep visual generation in which GAN model seeks to generate an image (e. g. , face image) from random noises. 2) Image generation from images: it tries to transform a given image into a new one (e. g. , from black-and-white image to color image). This task can be applied to applications like style transfer and image reconstruction. 3) Image generation from texts: it is a very natural task just like that humans describe the content of a painting and then the painters draw the corresponding images based on the texts. 4) Video generation from images: it aims to turn a static image into a dynamic video, which can be used in time-lapse photography, making animated videos from pictures, etc. 5) Video generation from videos: it is mainly used for video style transfer, video super-resolution and so on. 6) Video generation from texts: it is more difficult than image generation from texts since it needs the generated videos focusing on both semantical alignments with text and consistency among video frames. The challenges in deep visual generation have been analyzed and discussed. First, rather than 2D data, we should try to generate high-quality 3D data, which contains more information and details. Second, we could pay more attention to video generation instead of only image generation. Third, we could conduct some researches on controllable deep visual generation methods, which are more practical in real-world applications. Finally, we could try to expand the style transfer methods from two domains to multiple domains. In this review, we have summarized very recent works on deep adversarial visual generation through a systematic investigation. The review has mainly included an introduction of deep visual generation background, typical generation models, an overview of mainstream deep visual generation tasks and related algorithms. The deep adversarial visual generation research has been conducted further.

Key words: deep learning; visual generation; generative adversarial networks (GANs); image generation; video generation; 3D-depth image generation; style transfer; controllable generation

0 引 言

深度视觉生成是计算机视觉领域的一个重要研究方向,任务是根据特定的输入(随机噪声、文本、图像和视频等)生成与目标分布相匹配的图像或视频,可以实现对图像和视频的生成、美化、渲染和重建等操作。视觉生成任务在实际生活中并不陌生,很多场景的本质都是某种程度上的视觉生成,如艺术家进行绘画、电影工作者制作电影等。这类场景中,目标分布就是创作者脑海中构思的场景或视觉效果,生成过程就是人为地将其呈现出来。而深度视觉生成技术试图将人工生成的过程转化为智能生成的过程,以大幅减少重复性的人工劳动,甚至可以进行创造性的智能创作。深度视觉生成技术在视觉设计、图像/视频制作、艺术创作和电商广告等众多领域有广泛应用。具体任务包括老电影着色(Vondrick 等,2018)、破损照片修复(Wan 等,2020)、人体姿态估计(Wandt 和 Rosenhahn,2019)、动漫形象生成(Chen 等,2020e)、时尚设计(Dong 等,2020)、虚拟现实(Weng 等,2019)、广告生成(Zhu 等,2017)等。此外,深度视觉生成技术在医疗图像分析领域中有着至关重要的作用,可以用于医疗图像的生成(Frid-Adar 等,2018)、分割(Zhang 等,2018)、重构(Chen 等,2018)、检测(Baumgartner 等,2018)、去噪(Wang 等,2018a)、配准(Fan 等,2018)和分类(Ren 等,2018)等。目前深度视觉生成技术已实际服务于上述产业,并取得较大成功。从研究角度来看,深度视觉生成经过多年发展已成为机器学习热门的方向之一。该领域的文献数量

增长十分迅速，仅 2019—2020 年相关文献便超过 5 000 篇。

本文针对深度对抗视觉生成经典工作与最新工作进行概述性总结。介绍了两种经典的深度视觉生成算法：变分自编码器（variational auto-encoder，VAE）与生成对抗网络（generative adversarial networks，GANs）。针对深度对抗视觉生成的典型任务进行概括与总结，包括从噪声生成图像、从图像生成图像、从文本生成图像、从图像生成视频、从视频生成视频、从文本生成视频等。分析总结了目前深度对抗视觉生成的关键问题及挑战，并以此引出深度对抗视觉生成的未来发展趋势。

1　深度视觉生成概述

深度视觉生成的目标是生成尽可能真实的数据，其关键在于构造生成模型。典型的生成模型包括变分自编码器和生成对抗网络。

1.1　变分自编码器

变分自编码器是 Kingma 和 Welling（2014）基于编码器（encoder）和解码器（decoder）结构提出的一种经典深度视觉生成模型。如图 1 所示，在视觉生成过程中，VAE 先使用自动编码器将原始图像编码成潜变量（latent variable），并假设该变量符合正态分布，因而计算其平均值和标准偏差。然后从正态分布中采样并使用解码器生成图像。在参数优化过程中，VAE 使用交叉熵作为重构损失函数，使得生成数据与原始数据尽可能相近以保证数据生成质量；使用 KL（Kullback-Leibler）散度使得编码器的输出分布尽可能接近给定分布（即正态分布）。VAE 广泛应用于数据降维和数据生成（Zhu 等，2020a；Zhu 等，2020b）等方面，具有训练快、稳定等优势。然而，VAE强制性地将数据拟合到有限维度的预设分布上，两个分布的不匹配会导致 VAE 生成的图像不够清晰，限制了其应用范围。

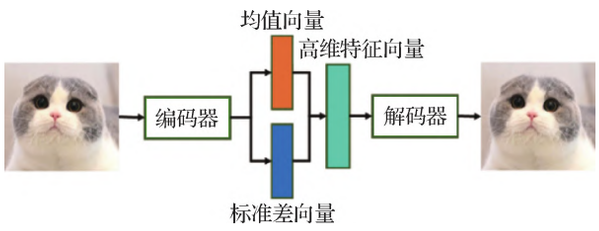

图 1　VAE 原理示意图

Fig. 1　Overview of VAE

1.2　生成对抗网络

相比变分自编码器而言，生成对抗网络（Goodfellow 等，2014）使用神经网络学习输入和输出分布之间的映射，能够生成更逼真、质量更高的图像/视频，应用范围更广。GAN 及其变体的本质是解决一个分布匹配问题（distribution matching problem）。模型对已有数据进行学习，获得匹配已有数据分布（该分布通常很难直接描述）的能力，进而生成符合目标分布的图像或视频。如图 2 所示，GAN 模型由生成器（generator）和判别器（discriminator）构成。生成器用来生成伪造数据（fake data），判别器用来区分伪造数据和真实数据（real data）。为分别提高二者的生成能力和判别能力，GAN 利用对抗博弈的思想进行优化。更好的生成器可以促使判别器优化，而更强的判别器则能促使生成器优化，二者博弈直至生成器能生成满足要求的数据，即具备真实性（reality）与多样性（diversity）的数据。其中，真实性指生成的数据要足够真实，至少使人无法分辨真假；多样性指生成的数据需与训练数据不完全一致，即能够生成新数据，否则相同的数据对任务没有任何帮助。

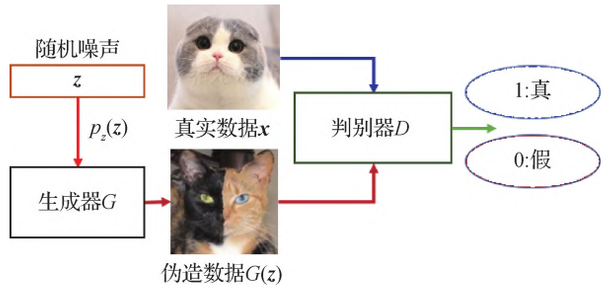

图 2　GAN 原理示意图

Fig. 2　Overview of GAN

原始生成对抗网络根据输入噪声 z 生成图像数据。此处噪声可以看做高维图像数据在低维空间上的一个投影 $p_z(z)$，生成器可看做一个映射函数。生成数据时，从低维空间随机采样噪声数据并输入到生成器中，生成器能够将数据从低维空间映射到高维空间，在高维空间中的对应点就是满足需要的生成图像。然而作为第 1 个提出的生成对抗网络，原始 GAN 存在一些缺陷：生成图像清晰度有限，多样性不足导致模式崩塌（mode collapse），难以解耦

隐空间(latent space)特征导致生成可控性差等。

2 典型对抗视觉生成任务及进展

原始 GAN 生成的图像有很多缺陷,而且仅能从噪声生成图像,在应用上有较大局限性。针对上述问题,Gui 等人(2020)对生成对抗网络提出了改进方案。GAN 发展历程如图 3 所示,这些工作一方面从不同角度改进生成和对抗算法,提高视觉生成的质量,使 GAN 能够生成真实、多样和可控的图像;另一方面提出适用于多模态数据的生成对抗网络,使 GAN 能够实现更多的视觉生成任务,扩展视觉生成的应用领域。相比于 VAE,GAN 的使用更灵活,应用范围更广,因此本文主要对 GAN 相关算法进行综述。

经过多年发展,对抗视觉生成已覆盖众多任务。如图 4 所示,典型任务包括从噪声生成图像、从图像生成图像、从文本生成图像、从图像生成视频、从视频生成视频和从文本生成视频。

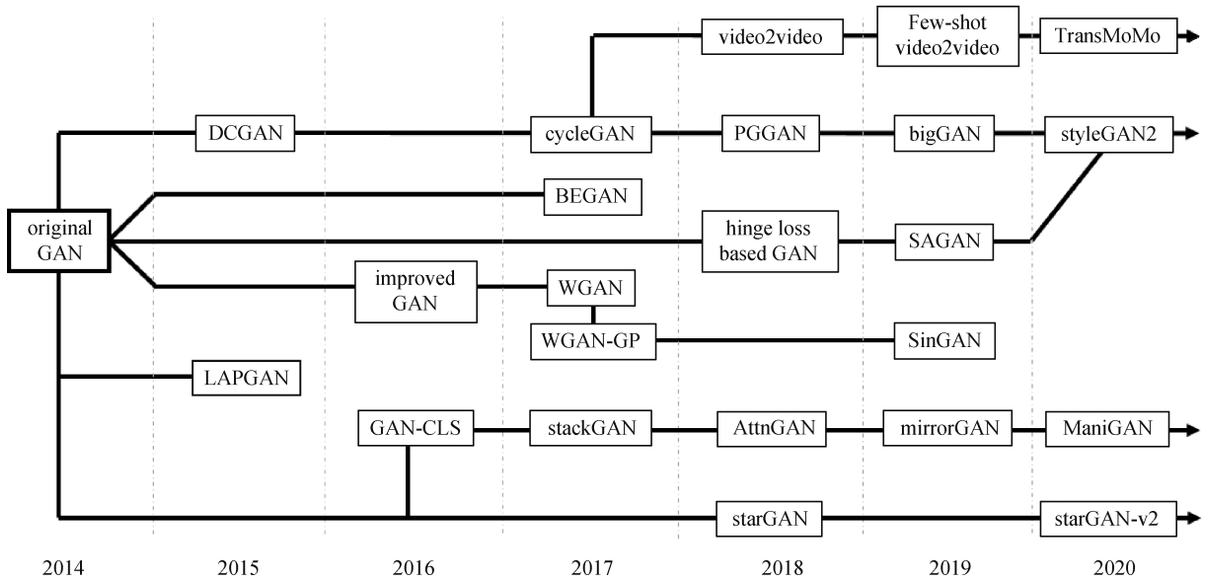

图 3 GAN 发展历程

Fig. 3 A road map of GANs

图 4 对抗视觉生成任务分类

Fig. 4 Task classifications for adversarial visual generation

2.1 从噪声生成图像

从噪声生成图像是深度对抗视觉生成最早出现的任务,可以用于数据增广、动画生成和人脸生成等。目前已有很多经典方法(Brock 等,2019;Gao

等，2020；Deng 等，2020；Kaneko 和 Harada，2020；Karras 等，2020）。

由于最原始的生成对抗网络（Vanilla GAN）（Goodfellow 等，2014）难以训练，且稳定性差，DCGAN（deep convolutional GAN）（Radford 等，2016）将卷积神经网络与 GAN 相结合，设计了一个较好的网络架构，使 GAN 在多数情况下能稳定训练。但是，该方法并没有完全解决 GAN 训练不稳定这一难题。针对此问题，WGAN（Wasserstein GAN）（Arjovsky 等，2017）使用 Wasserstein 距离衡量生成数据与真实数据的分布距离，该距离能够准确反映生成器生成样本的质量。WGAN 不仅提高了 GAN 训练过程的稳定性，使其不再受限于生成器和判别器的训练程度，而且一定程度解决了模式崩塌问题，使生成的图像更具多样性。虽然 WGAN 能够改善 GAN 的训练稳定性，但还是会出现不收敛的情形，导致生成的样本质量很差。于是，WGAN-GP（improved training of WGAN）（Gulrajani 等，2017）提出一种梯度惩罚策略，使 WGAN 的训练更加稳定，且能够生成更高质量的图像。随后，Karras 等人（2018）提出 PGGAN（progressive GAN），其关键思想是渐进式地增加生成器和判别器的规模，不断向网络中添加新层使网络模型逐渐复杂化，从而学习到更细化的特征。这种方法既可以提高 GAN 的训练速度，又能够使训练过程更加稳定，从而提升生成图像的质量。尽管如此，利用 ImageNet 这类复杂数据生成高分辨图像仍是一个难题。为解决这个问题，如图 5 所示，BigGAN（Brock 等，2019）通过减少生成器输入方差来精确控制样本保真度和多样性之间的平衡，提升大规模 GAN 训练过程的相对稳定性，最终达到提高图像生成质量的目的。但是，这样的网络仍然很难处理好生成图像的细节和整体的权衡。为此，Zhang 等人（2019）提出自我注意力生成对抗网络（self-attention GAN，SAGAN），将注意力机制引入 GAN，可以很好地发现图像中的依赖关系，从而更好地处理生成图像每个位置的细节，提升图像质量。

此外，为更好地控制生成图像特定特征，StyleGAN（Karras 等，2019）提出基于风格的生成器（style-based generator），通过分别修改每层输入来控制每层的视觉特征，大幅提升了 GAN 的生成可控性，同时提升了使用 GAN 进行视觉生成的可解释性。由于实例归一化与渐进式生成框架会导致

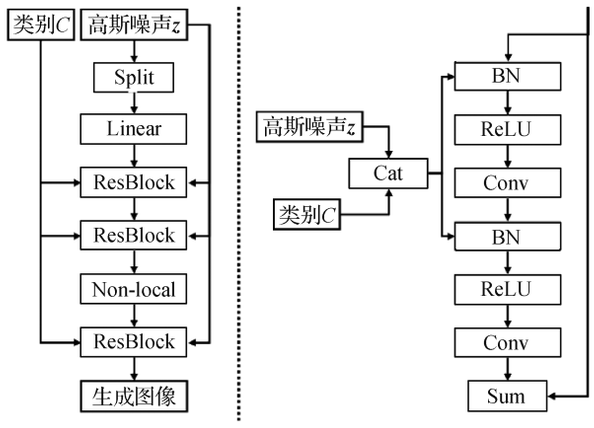

图 5　BigGAN 生成网络结构（Brock 等，2019）

Fig. 5　The architecture of BigGAN（Brock et al. , 2019）

（（a）network architecture；（b）ResBlock architecture）

StyleGAN 生成的图像出现斑点状特征伪影，对此，StyleGAN2（Karras 等，2020）提出以下解决方案：改进生成器中归一化方法，以移除图像斑点状特征伪影；调整训练框架，以保证训练过程中不改变网络的拓扑结构。该方法极大提升了生成图像的质量和效果。

2.2　从图像生成图像

从图像生成图像也是深度对抗视觉生成一个重要方向，具体任务包括风格迁移、图像重建和图像超分辨率等（Huang 等，2018；Choi 等，2020；Kim 等，2020a）。

对于风格迁移任务，构建内容对齐的样本对是极其重要的，然而实际中很难得到对齐的样本对。为解决这一难题，CycleGAN（Zhu 等，2017）通过训练两个 GAN 模型实现在没有成对数据情况下将图像从源域转换到目标域。该方法在训练时只需将源域和目标域的图像作为输入即可，不要求其图像内容匹配。然而，该方法无法从给定源域图像生成多种风格图像。针对该问题，MUNIT（multimodal unsupervised image-to-image translation）（Huang 等，2018）提出一个多模态的图像到图像转换框架，将图像空间拆解成内容空间与风格空间，通过输入不同的风格编码与内容编码组合将二者结合进行图像重构，以产生不同的多模态输出图像，从而允许用户可控地进行图像内容及风格转换。大多数图像迁移方法在训练时都需要大量的源域和目标域数据，很大程度上限制了这些方法的使用。为解决该问题，FUNIT（few-shot unsupervised image-to-image transla-

tion)(Liu 等,2019a)提出少样本无监督的图像到图像迁移方法,将对抗训练网络与一个新颖的网络相结合,以实现利用少数样本来迭代生成图像的目的。但该方法在训练时高度依赖大量的人工标注,这些标注的数据在训练中起着关键作用。为解决该问题,Wang 等人(2020c)提出使用抗噪声的伪标签进行半监督学习,使用循环一致性约束(cycle consistency constraint)来利用未标记数据中的信息,并对模型结构进行相应修改,以实现半监督的图像迁移任务。

在图像生成任务中,大多生成对抗方法通过学习数据分布来生成与训练图像类别相同的图像,这就要求训练集的类别是明确具体的,且每一类需要足够数量的图像。SinGAN(learning a generative model from a single natural image)(Shaham 等,2019)打破了这些限制,提出一种非条件的生成模型,仅利用一幅自然的图像,学习不同尺度的图像块间的关系,使用多尺度的对抗学习模型,如图 6 所示,最后得到更高质量且多样化的图像。

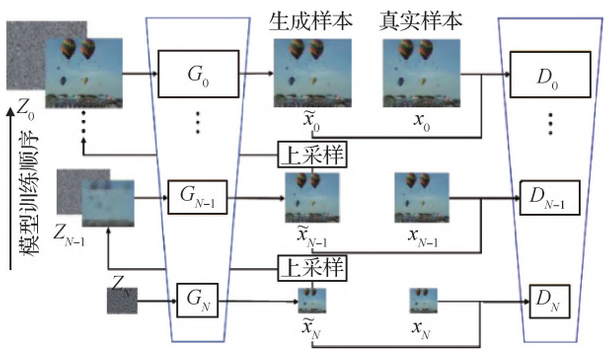

图 6 SinGAN 多尺度对抗学习模型(Shaham 等,2019)

Fig. 6 Multi-scale pipeline for SinGAN(Shaham et al., 2019)

超分辨率生成对抗网络(super-resolution GAN, SRGAN)(Ledig 等,2017)是一种提高图像分辨率的技术。然而这种方法在图像细节处理方面仍有不足,如生成的图像有伪影。为进一步提高图像超分辨率的质量, Wang 等人(2018d)对 SRGAN 进行改进并提出 ESRGAN(enhanced super-resolution GAN)方法。ESRGAN 去除所有批量归一化层以减少伪影,并引入新的具有更高容量的残差块来改进生成器,同时使用激活前的特征来改善感知损失(perceptual loss)。该方法能够生成更逼真且自然的图像。目前大多数单图像超分辨率方法通常是通过最小化超分辨率图像和低分辨率图像间的像素距离进行优化,这往往导致生成的图像模糊。为改进图像模糊问题,最新提出的 PULSE(self-supervised photo upsampling via latent space exploration of generative models)方法(Menon 等,2020)使用自监督学习方法,遍历高分辨率图像空间来匹配原始低分辨率图像,并利用高维高斯函数性质限制搜索空间来确保输出图像真实性,取得了更好的视觉效果。

2.3 从文本生成图像

从文本生成图像是一个非常自然的任务,就像人类使用语言对一幅画的内容进行描述,画家根据文本描述画出相应的图像。目前,该方向已有许多经典工作(Qiao 等,2019;Cha 等,2019;Li 等,2020a;Plumerault 等,2020;Mathew 等,2020)。

Mao 等人(2017)提出一种能够将一段描述性文本直接转换成图像的方法,基于 DCGAN(Radford 等,2016)向判别器额外增加真实图像和错误的文本描述,使判别器学习到文本与图像更匹配的对应关系,并通过插值的方法产生大量的文本以解决文本描述数量不足问题,最终实现由文本生成图像的任务。由文本特征直接生成图像很难得到较高分辨率的图像,为此,StackGAN(Zhang 等,2017)采用一种逐步递进的思想,先由文本生成图像的基本轮廓与颜色,然后再对此图像进行纠正并添加更多细节,从而生成高分辨率图像。但是,这类方法严重依赖初始生成图像的质量,且生成的图像与文本的关联性较弱。为解决这些问题,Zhu 等人(2019)提出 DM-GAN(dynamic memory GAN),设计了一个动态记忆模型(dynamic memory module)选择与生成图像相关的单词,使生成的图像很好地匹配文本描述。

为使生成图像更加细致,AttnGAN(attentional GAN)(Xu 等,2018)引入注意力机制,通过关注文本描述中的关键词,能够在图像的不同子区域生成更精细的信息。然而,由于文本和图像模式的多样性,仅使用单词级别的注意力机制并不能确保全局语义的一致性。因此, MirrorGAN(Qiao 等,2019)先根据文本生成图像,再将图像重新转换成文本,并与原始文本进行对比,从而更好地解决文本与图像间的一致性问题。

生成图像属性与给定文本中属性表述不一致是从文本生成图像普遍存在的问题。ManiGAN(text-guided image manipulation GAN)(Li 等,2020a)试图解决该问题,提出文本和图像的仿射结合模型(affine combination module, ACM),以融合图像特征与文本特征,并设计细节改正模型(defail cornection

module，DCM）来纠正不匹配属性，同时补全图片细节。ManiGAN 模型结构如图 7 所示。

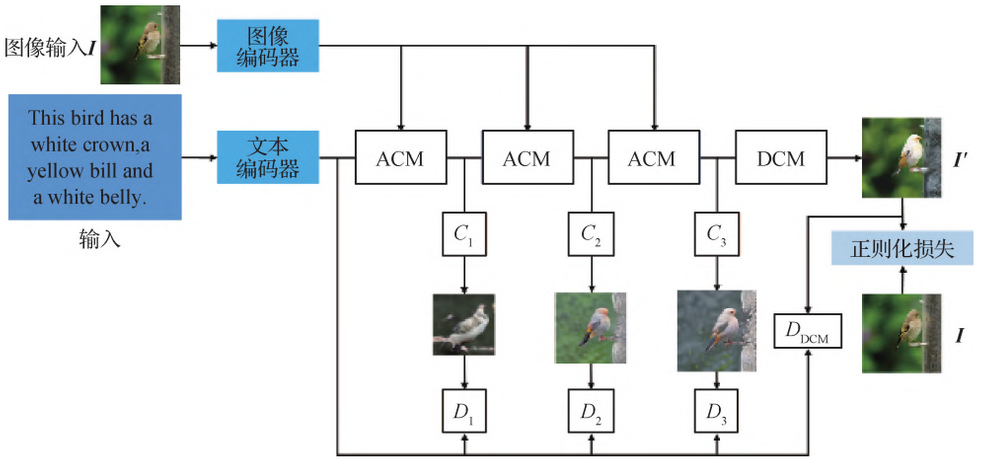

图 7　ManiGAN 模型结构（Li 等，2020a）

Fig. 7　The architecture of ManiGAN（Li et al.，2020a）

2.4　从图像生成视频

从图像生成视频是指将静态图像变成动态视频，具体可以用在延时摄影、从图像制作视频动画等任务。该方向已开展了许多研究工作（Shen 等，2019；Kim 等，2020b；Otberdout 等，2020；Zhao 等，2020；Maximov 等，2020）。

恢复沿维度折叠的视频任务具有很大挑战性，其中运动模糊的图像就是通过沿时间维度投影运动轨迹生成的。为解决这个问题，Balakrishnan 等人（2019）提出一种针对运动视频恢复的网络架构，首先使用一个概率模型解决任务的不确定性，然后使用卷积神经网络学习各域的图像结构，以生成准确的信号。该方法可以从空间投影中恢复行人走路的姿态和人脸图像，也可以恢复 MNIST（Mixed National Institute of Standards and Technology）数据集手写数字的动作方向。

在生成视频时，一个重要问题是如何获取时序信息，多数方法需借助参考视频提供的颜色外观变化信息，然而寻找与输入图像具有相似语义信息的参考视频十分困难。针对这个问题，Time-lapse 方法（Nam 等，2019）通过使用多帧联合的条件网络，学习室外场景光照变化与时间之间的相关性，并引入时间戳作为控制变量，从而避免使用参考视频。该模型可将一个单一的户外图像生成一个具有时序信息的连续视频，实现延时摄影效果。

从图像生成视频技术推动了虚拟试衣系统的发展。Dong 等人（2019）提出一种流指导变换的生成对抗网络（flow-navigated warping GAN，FW-GAN）。该模型旨在通过一幅人物图像、一幅服装图像，以及一些目标姿势学习生成一段虚拟试衣视频。FW-GAN 提出一种流嵌入判别器（flow-embedding discriminator），即通过在鉴别器中加入有效的流输入来改善时空平滑性，同时使用语法一致性损失函数（parsing constraint loss）作为结构约束的一种形式，改善模型在不同姿态和不同服装下的生成结果。该模型还能够缓解因人体姿势不同导致的严重遮挡问题。除此之外，Zhao 等人（2020）提出一个根据已创建的绘画生成一段绘画视频的方法，由于画家使用的绘画技巧独特，且色彩组合丰富，该方法旨在学习这种绘画技巧的随机决策。

2.5　从视频生成视频

从视频生成视频主要包括视频的风格转换、迁移和超分辨率等工作（Wang 等，2018b，c；Chan 等，2019；Xu 等，2019；Yang 等，2020；Maximov 等，2020）。

视频到视频的视觉生成任务旨在将输入的语义视频转换为具有真实感的视频，但生成的视频很难保证前后帧的一致性。针对此问题，Wang 等人（2018b）提出一种基于对抗学习框架的视频生成模型 Vid2Vid（video-to-video synthesis），将前后帧的光流信息作为约束，对 pix2pixHD（pixel to pixel high definition）（Wang 等，2018c）进行改进，进而生成连贯且高质量的视频。但是，这些视频生成模型存在数据匮乏及模型泛化能力有限等问题。针对该问题，Few-shot Vid2Vid（Wang 等，2019c）提出一种基

于小样本学习的视频生成框架,如图 8 所示。在生成人物视频时,使用注意力机制捕捉身体局部区域,以生成视频中未曾见过的信息。尽管这些 Vid2Vid 的方法可以实现短时间内的时序一致性,但不能保证长期的时序一致性。为解决这一局限性,Mallya 等人(2020)引入了一个新的 Vid2Vid 框架,可以在渲染过程中有效利用所有过去生成的视频帧。该方法不仅提高了视频生成的质量,而且使单个图像生成器移植到视频生成器成为可能。

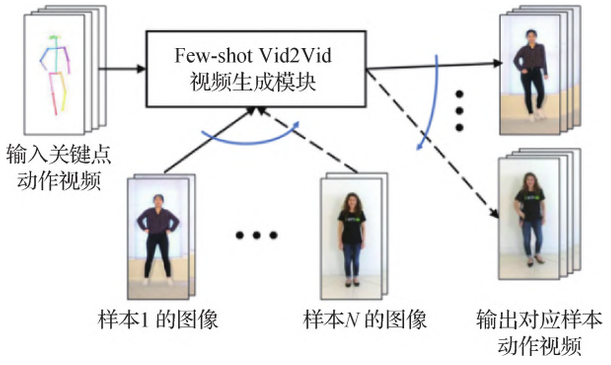

图 8　Few-shot Vid2Vid 视频生成框架(Wang 等,2019c)

Fig. 8　The architecture of Few-shot Vid2Vid (Wang et al., 2019c)

此外,视频修复可以应用到许多视频编辑和视频恢复任务中,目的是利用视频中已有的内容填补视频时空上的空洞。Kim 等人(2019b)提出一种深度神经网络的快速修复视频方法,基于编码器—解码器模型,旨在收集相邻帧的细节信息,并生成未知区域。同时,通过建立一个周期性反馈与一个时序模块,使输出保持时间的一致性。该方法能够生成更准确、更流畅的视频,将其应用到视频重定向任务中,能够取得良好的视觉效果。

在视频风格迁移方向,Chan 等人(2019)提出一个将源视频的舞姿转移给目标人物视频的模型,首先使用检测器创建输入视频的姿态估计模型,然后将姿态进行标准化,最后设计一个系统来学习从标准化姿势到目标人物的图像映射,从而生成一段目标人物和源视频同样动作的新视频。尽管如此,这些方法在处理人体动作迁移任务上,仍然很难建立准确的模型来刻画人体复杂的非线性动作,且该任务在真实世界中缺少有效的动作匹配数据。为此,Yang 等人(2020)提出 Trans-MoMo 模型,这是一种无监督的人体动作重定网络(motion retargeting network),可利用 2 维关键点信息,根据无标注的网络数据端到端地训练,从而更好地生成人体动作视频。

2.6　从文本生成视频

从文本生成视频也是深度对抗视觉生成的重要方向之一,其任务与从文本生成图像类似。从文本生成视频是一个重大挑战,需要满足以下几个特殊要求:1)整段文本与整个视频语义一致;2)文本中有实际意义的单词与视频中局部区域(如物体)语义一致;3)视频帧之间连贯。针对上述要求,已开展了许多相关研究工作(Gupta 等,2018;Lin 等,2018;Balaji 等,2019a,b;Chen 等,2020a)。

Mittal 等人(2017)最早提出利用文本生成视频,通过 VAE 与注意力机制结合来创建时序的帧序列。实验表明,Sync-DRAW(synchronized deep recurrent attentive writer)可以有效学习视频的时空信息。随后,Marwah 等人(2017)提出一种生成可变长度语义视频的网络结构,能够增量式地生成视频,还能进行时空风格转换。

此外,Li 等人(2018)通过训练一个条件生成模型解决文本生成视频问题。使用两个生成器分别用于生成背景颜色与获取文中的动态信息;同时开发了一种从公开的在线视频中自动创建匹配的文本到视频语料库的方法获取训练数据。该生成框架在准确反映输入文本信息的同时,生成的视频准确且多样化。尽管如此,这些使用 VAE 和 GAN 的方法可能会出现生成视频模糊或训练过程不稳定以及难以收敛的问题。为此,Liu 等人(2019b)提出一个跨模态的对偶学习方法(cross-modal dual learning),通过对偶学习机制,同时学习句子和视频之间的双向映射,从而生成更真实的视频,并能够与相应的文本描述较好地保持语义的一致性。

这些方法在生成特定的视频帧时,大都未能充分利用之前生成的帧信息,且视频与文本信息一致性的衡量指标没能很好地建立。为解决这些问题,Deng 等人(2019)提出一个内省的循环卷积 GAN(introspective recurrent convolutional GAN),该模型生成器既考虑了每一帧视频的信息,又考虑了整个视频的时间连贯性,同时利用互信息(mutual information)来具体衡量语义一致性,使模型生成的视频与对应的文本之间的语义距离不断地进行对比,从而使生成的视频具有更好的视觉质量。Chen 等人(2020a)在此方面提出了一个自底向上的生成对抗

网络模型 BoGAN(bottom-up GAN)，如图 9 所示，对整体文本与视频、单词与视频局部区域之间进行语义对齐，同时使生成视频帧间的变化与真实视频帧变化一致，以保证视频连贯性。

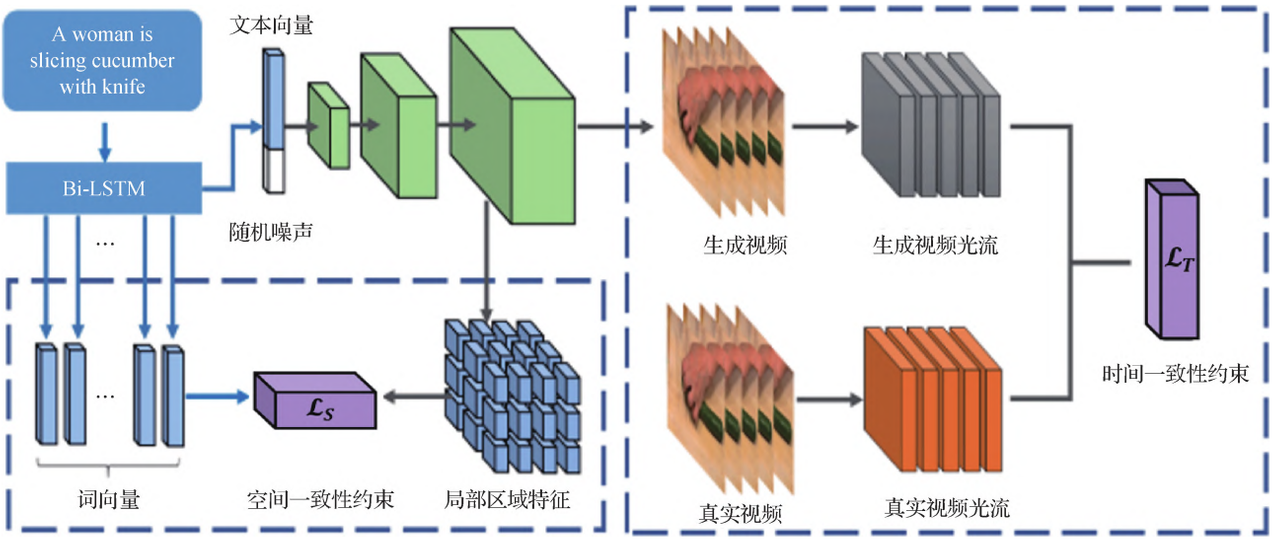

图 9　BoGAN 网络框架(Chen 等，2020a)

Fig. 9　The architecture of BoGAN(Chen et al. , 2020a)

3　深度对抗视觉生成未来趋势

深度对抗视觉生成发展至今，已在多个领域实现落地应用，并创造较大的实际价值。然而深度对抗视觉生成依然面临诸多挑战。总体来说，深度对抗视觉生成的最大挑战在于生成的图像或视频无法与人类的视觉认知达成高度一致。具体来说，首先，现在大多深度对抗视觉生成方法只能生成 2D 图像，而 2D 图像难以完全反映现实 3D 世界；其次，在视频生成任务中时间维度信息是必不可少的重要特征，然而输入数据往往缺乏该信息，如图像、文本等；再次，深度对抗视觉生成难以可控地生成图像和视频，生成的样本随机性较大，只能实现一些简单任务；最后，深度对抗视觉生成用于风格迁移等任务上时仅能较好地实现两个域之间的迁移，多域迁移的效果很差。

深度对抗视觉生成的未来发展趋势就是针对上述挑战改进生成算法，弥补目前生成算法的不足，加强深度视觉生成的真实性、可控性和多样性，进一步扩大应用范围。

3.1　3 维深度图像生成

现实世界是 3D 的，3 维深度图像无疑能够更加真实地反映现实世界和人类视觉感受。2D 视觉生成不可避免地在许多领域的实际应用中受到限制，如人脸 3D 建模、机器人学习、虚拟现实、游戏行业和设计行业等。3 维深度图像生成的关键问题在于如何从 2D 图像或文本等数据中构建出深度信息并进行真实准确的 3 维建模。

视角合成是实现 3 维深度图像生成的一种重要方法，其通过已有的一组视角生成未知的目标视角。基于多视角图像可以进行 3 维建模。DVS(deep view synthesis)(Liu 等，2021)针对视角合成中像素匹配难度大、生成图像质量差的问题，提出自洽机制与已有视角进行一致性约束，进而结合生成模型生成高质量的目标视角图像。

Wu 等人(2020a)实现了从单幅 2D 图像重构 3 维深度图像并取得了很好的效果，仅使用一幅单视图图像而无需额外监督信息来生成高质量的 3 维物体模型。该方法基于对称性假设，采用多个 encoder-decoder 网络将一幅物体图像分解为深度、光照和视角等多个维度，组合渲染，重构出 3 维物体模型。PIFu(pixel-aligned implicit function)(Saito 等，2019，2020)由一幅人物的高清图像进行高保真 3 维重建，人物细节可以获得高精度还原。该方法提出了端到端训练的多级结构，粗糙模型观察低分辨率图像，专注于整体推理；精细模型观察更高分辨率的图像获取更多细节信息实现高精度还原。

NormalGAN(Wang 等,2020a)实现了从单个 RGB-D 图像,重建出 3 维人体。该方法提出法线贴图(normal mapping)约束的对抗学习框架,对前视深度图进行有效去噪校正,并推断具有几何细节的后视图深度图像。最终结合前视图和后视图的 RGB-D 信息生成完整和详细的人体 3 维模型。

Mildenhall 等人(2020)提出一个 MLP(multilayer perceptron)网络 NeRF(representing scenes as neural radiance fields for view synthesis)非显式地学习静态 3 维场景。针对一个静态场景使用神经网络建模,使用大量已知相机参数的图像进行训练,训练完成后可以从任意角度渲染出清晰的场景图像。Chen 等人(2020b)针对房屋设计专业性要求高、过程复杂且烦琐问题,提出 HPGM(house plan generative model)方法,使用图卷积布局预测网络(graph conditioned layout prediction network,GC-LPN)构建房屋布局,并通过语言引导生成对抗网络(language conditioned texture GAN,LCT-GAN)生成房间内部纹理,最后使用 3 维渲染技术生成 3 维房屋模型图,如图 10 所示。

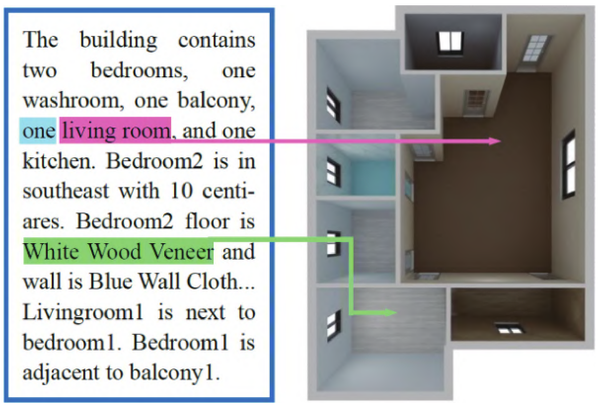

图 10　HPGM 方法 3 维房屋模型生成效果图

Fig. 10　An example of generated 3D house using HPGM

3.2　高质量视频生成

视频是反映动态世界最直观手段,已成为信息的主要载体,也是目前最主要的大数据来源。视频生成可以用于延时摄影、视频预测和视频超分辨率等任务。相对于文本、语音和图像等数据,视频数据维度更高、内容更复杂,尤其是在连续多帧预测中会产生累积误差问题。因此,视频生成的技术挑战很大。从图像或文本生成视频的关键问题在于如何获取时间维度信息,让视频能够流畅且真实。

生成外观和动作逼真的人类视频是一项十分有挑战性的任务。为解决这一难题,Wang 等人(2020b)提出一种新型时空生成模型 G3AN(three-stream generator-architecture),旨在捕获高维视频数据的分布,并且通过分解多尺度的外观和运动特征以实现时空一致性融合,同时提出一种新的自我注意力模块用于视频生成,大幅提升了视频生成的质量。

视频超分辨率(video super-resolution,VSR)(Liao 等,2017)旨在从低分辨率当前帧和相邻帧中恢复逼真的高分辨率视频帧。VSR 通常利用当前帧坐标系与相邻帧坐标系之间的光流实现时域对齐(temporal alignment)。因此,不准确的光流会导致视频生成效果较差。Tian 等人(2020)提出 TDAN(temporally-deformable alignment network)方法首次将可变卷积(deformable convolution)引入 VSR,使用可变卷积组成网络,自适应地使当前帧与相邻帧对齐,然后利用重构网络融合对齐后的视频帧。该方法避免使用光流进行计算,取得了较好的视频生成结果。

3.3　随机生成到可控生成

在现实世界中,人类艺术家都是随心所欲的进行绘画创作、电影动画设计和特效制作等视觉相关工作。作者往往是在心中有整体构思和预想之后,通过各种方法将其构建出来,形成最终成果。由此可见,现实世界中的视觉生成过程是可控的,能够达到期望的视觉效果。然而现有的深度视觉生成技术可控性差,大部分只能进行随机视觉生成,无法用于对生成可控性和生成细节要求高的任务中。因此,可控的深度视觉生成十分具有挑战性。

StyleGAN 能够生成具有眼睛、牙齿、头发和上下文(脖子、肩膀、背景)的逼真面部图像,但缺乏对面部姿势、表情和场景照明等语义参数的控制。为解决该问题,StyleRig(Tewari 等,2020a)提出基于 3DMM(3D morphable face model)和 StyleGAN 的人脸重建模型,利用 3DMM 参数进行人脸面部的可解释性编辑。该算法以自监督方式训练,无需人工标注。

大多图像转换框架缺乏对图像变化因子(如物体形状、纹理和背景等)进行单独学习的能力。为此,Li 等人(2020b)提出 MixNMatch(mix-and-match image generation method)条件生成模型,可同时将物体背景、形状等因子编码到潜在空间,利用联合图像编码分布匹配来学习潜在因子编码器,通过控制这

些因子组合生成逼真图像。PIE(portrait image embedding)(Tewari 等,2020b)提出了一种将真实人像图像嵌入 StyleGAN 潜在空间的方法,允许对图像中的头部姿态、面部表情和场景照明直观地进行编辑。将 3 维人脸模型的控制空间映射到 GAN 的潜在空间,最终实现了对人脸参数空间的语义编辑。该方法能够可控生成更高质量的肖像照片。

Men 等人(2020)提出一种新的可控的人物图像生成模型(attribute-decomposed GAN),如图 11 所示,该模型可以生成具有所需人物属性(如衣服、裤子和姿态等)的真实人物图像。其核心思想是将人的属性特征作为独立编码嵌入潜在空间,通过建模固有姿态和人物属性之间复杂的相互作用,实现对人物属性的灵活控制,大幅提高了人物图像生成的质量。StyleFlow(Abdal 等,2021)研究了属性条件采样和属性控制编辑两个子问题,在隐空间控制隐藏特征来控制图像的属性,在人脸和汽车属性编辑上取得了很好效果。

图 11　基于 attribute-decomposed 的可控生成模型(Men 等,2020)

Fig. 11　Controllable generative model based on attribute-decomposed GAN(Men et al., 2020)

3.4　二域迁移到多域迁移

风格迁移是视觉生成中的一个重要研究方向,包含画风迁移、人脸迁移和动作迁移等,具有众多应用场景和重要实际价值。现有深度视觉生成方法无法解决多领域迁移问题,只能进行二域之间的互相转化。但实际应用中,多个数据集或多种属性等多域互相转化的要求是很常见的。此外,多域迁移不仅对机器视觉很重要,对机器学习中的迁移学习、半监督学习和统计学都是十分重要的课题。

为降低多域生成难度,AEGAN(auto-embedding)(Guo 等,2019)通过自动编码器学习图像内在高维结构信息,并将其作为跳板进一步生成高分辨率图像,同时设计降噪网络去除生成图像中的噪点并填补细节信息。

StarGAN(Choi 等,2018)通过跨领域和数据集的训练方式解决传统 GAN 在多领域之间风格迁移低效问题。该方法在生成器的输入中添加目标域信息,并改变判别器结构使其不仅能判别真伪,还能判断图像类别,在面部属性转移和面部表情生成任务中取得了更好的效果。StarGAN-V2(Choi 等,2020)是 StarGAN 的升级版本,主要解决图像风格迁移过程中多样性不足、多领域可扩展性有限等问题,设计了一个映射网络用于生成风格编码,然后用风格编码指导生成器进行目标风格学习,从而实现目标域下多风格图像的转换。但对于生成图像任务,直接生成高分辨率图像十分困难,且容易导致生成图像包含噪点及结构不完整的物体。

此外,现有方法常常忽略联合优化域间多个边缘分布距离,这可能导致分布不匹配问题。基于最优传输理论,Cao 等人(2019)提出 MWGAN(multi-marginal Wasserstein GAN),在不同域(P)之间联合优化多个边缘分布距离,如图 12 所示,利用跨域相关性并缓解分布不匹配问题。

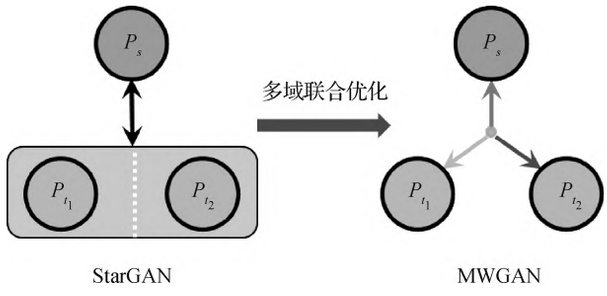

图 12　MWGAN 多域联合优化示意图(Cao 等,2019)

Fig. 12　Multi-domain joint optimization using MWGAN (Cao et al., 2019)

4　结　语

深度视觉生成是计算机视觉中的热门领域,经过多年高速发展已在基础理论、技术方法和落地应

用等方面取得了重要突破。本文对深度对抗视觉生成相关研究进展进行概述,将深度对抗视觉生成经典任务分为从噪声、图像、文本生成图像和从图像、视频、文本生成视频6类,并从深度对抗视觉生成技术现存挑战出发,对其未来发展趋势进行预测。希望通过对深度对抗视觉生成的分类总结,帮助相关研究人员了解该领域的发展现状,并对发展趋势进行启发性预测,期望能够促进对抗视觉生成领域发展,拓宽深度对抗视觉生成技术的应用范围。

参考文献(References)

Abdal R, Zhu P H, Mitra N J and Wonka P. 2021. StyleFlow: attribute-conditioned exploration of StyleGAN-generated images using conditional continuous normalizing flows. ACM Transactions on Graphics, 40(3): #21 [DOI: 10.1145/3447648]

Arjovsky M, Chintala S and Bottou L. 2017. Wasserstein GAN//Proceedings of 2017 International Conference on Machine Learning. Sydney, Australia: [s. n.]: 214-223

Balaji Y, Min M R, Bai B, Chellappa R and Graf H P. 2019a. Conditional GAN with discriminative filter generation for text-to-video synthesis//Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China: AAAI Press: 1995-2001 [DOI: 10.24963/IJCAI.2019/276]

Balaji Y, Min M R, Bai B, Chellappa R and Graf H P. 2019b. TFGAN: improving conditioning for text-to-video synthesis//Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: [s. n.]

Balakrishnan G, Dalca A V, Zhao A, Guttag J V, Durand F and Freeman W T. 2019. Visual deprojection: probabilistic recovery of collapsed dimensions//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 171-180 [DOI: 10.1109/ICCV.2019.00026]

Baumgartner C F, Koch L M, Tezcan K C, Ang J X and Konukoglu E. 2018. Visual feature attribution using Wasserstein GANs//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 8309-8319 [DOI: 10.1109/CVPR.2018.00867]

Brock A, Donahue J and Simonyan K. 2019. Large scale GAN training for high fidelity natural image synthesis//Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: [s. n.]

Cao J Z, Mo L Y, Zhang Y F, Jia K, Shen C H and Tan M K. 2019. Multi-marginal Wasserstein GAN//Proceedings of the 33rd Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates, Inc.: 1774-1784

Cha M, Gown Y L and Kung H T. 2019. Adversarial learning of semantic relevance in text to image synthesis//Proceedings of 2019 AAAI Conference on Artificial Intelligence, 33: 3272-3279 [DOI: 10.1609/aaai.v33i01.33013272]

Chan C, Ginosar S, Zhou T H and Efros A. 2019. Everybody dance now//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 5932-5941 [DOI: 10.1109/ICCV.2019.00603]

Chen Q, Wu Q, Chen J, Wu Q Y, van den Hengel A and Tan M K. 2020a. Scripted video generation with a bottom-up generative adversarial network. IEEE Transactions on Image Processing, 29: 7454-7467

Chen Q, Wu Q, Tang R, Wang Y H, Wang S and Tan M K. 2020b. Intelligent home 3D: automatic 3D-house design from linguistic descriptions only//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 12622-12631 [DOI: 10.1109/CVPR42600.2020.01264]

Chen Y H, Shi F, Christodoulou A G, Xie Y B, Zhou Z W and Li D B. 2018. Efficient and accurate MRI super-resolution using a generative adversarial network and 3D multi-level densely connected network//Proceedings of 2018 International Conference on Medical Image Computing and Computer-Assisted Intervention. Granada, Spain: Springer: 91-99 [DOI: 10.1007/978-3-030-00928-1_11]

Chen Z, Wang C Y, Yuan B and Tao D C. 2020e. PuppeteerGAN: arbitrary portrait animation with semantic-aware appearance transformation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 13515-13524 [DOI: 10.1109/CVPR42600.2020.01353]

Choi Y, Choi M, Kim M, Ha J W, Kim S and Choo J. 2018. StarGAN: unified generative adversarial networks for multi-domain image-to-image translation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 8789-8797 [DOI: 10.1109/CVPR.2018.00916]

Choi Y, Uh Y, Yoo J and Ha J W. 2020. StarGAN v2: diverse image synthesis for multiple domains//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 8185-8194 [DOI: 10.1109/CVPR42600.2020.00821]

Deng K L, Fei T Y, Huang X and Peng Y X. 2019. IRC-GAN: introspective recurrent convolutional GAN for text-to-video generation//Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China: AAAI Press: 2216-2222 [DOI: 10.24963/ijcai.2019/307]

Deng Y, Yang J L, Chen D, Wen F and Tong X. 2020. Disentangled and controllable face image generation via 3D imitative-contrastive learning//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 5153-5162 [DOI: 10.1109/CVPR42600.2020.00520]

Dong H Y, Liang X D, Shen X H, Wu B W, Chen B C and Yin J. 2019. FW-GAN: flow-navigated warping GAN for video virtual try-

on//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 1161-1170 [DOI: 10.1109/ICCV.2019.00125]

Dong H Y, Liang X D, Zhang Y X, Zhang X J, Shen X H, Xie Z Y, Wu B W and Yin J. 2020. Fashion editing with adversarial parsing learning//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 8117-8125 [DOI: 10.1109/CVPR42600.2020.00814]

Fan J F, Cao X H, Xue Z, Yap P T and Shen D G. 2018. Adversarial similarity network for evaluating image alignment in deep learning based registration//Proceedings of 2018 International Conference on Medical Image Computing and Computer-Assisted Intervention. Granada, Spain: Springer: 739-746 [DOI: 10.1007/978-3-030-00928-1_83]

Frid-Adar M, Klang E, Amitai M, Goldberger J and Greenspan H. 2018. Synthetic data augmentation using GAN for improved liver lesion classification//The 15th IEEE International Symposium on Biomedical Imaging (ISBI 2018). Washington, USA: IEEE: 289-293 [DOI: 10.1109/ISBI.2018.8363576]

Gao C Y, Liu Q, Xu Q, Wang L M, Liu J Z and Zou C Q. 2020. SketchyCOCO: image generation from freehand scene sketches//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 5173-5182 [DOI: 10.1109/CVPR42600.2020.00522]

Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y. 2014. Generative adversarial nets//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 2672-2680

Gui J, Sun Z, Wen Y G, Tao D C and Ye J P. 2020. A review on generative adversarial networks: algorithms, theory, and applications [EB/OL]. [2021-04-07]. https://arxiv.org/pdf/2001.06937.pdf

Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V and Courville A. 2017. Improved training of Wasserstein GANs//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 5769-5779

Guo Y, Chen Q, Chen J, Wu Q Y, Shi Q F and Tan M K. 2019. Auto-embedding generative adversarial networks for high resolution image synthesis. IEEE Transactions on Multimedia, 21(11): 2726-2737 [DOI: 10.1109/TMM.2019.2908352]

Gupta T, Schwenk D, Farhadi A, Hoiem D and Kembhavi A. 2018. Imagine this! Scripts to compositions to videos//Proceedings of 2018 European Conference on Computer Vision. Munich, Germany: Springer: 610-626 [DOI: 10.1007/978-3-030-01237-3_37]

Huang X, Liu M Y, Belongie S and Kautz J. 2018. Multimodal unsupervised image-to-image translation//Proceedings of 2018 European Conference on Computer Vision. Munich, Germany: Springer: 179-196 [DOI: 10.1007/978-3-030-01219-9_11]

Kaneko T and Harada T. 2020. Noise robust generative adversarial networks//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 8401-8411 [DOI: 10.1109/CVPR42600.2020.00843]

Karras T, Aila T, Laine S and Lehtinen J. 2018. Progressive growing of GANs for improved quality, stability, and variation//Proceedings of the 6th International Conference on Learning Representations. Vancouver BC, Canada: [s.n.]

Karras T, Laine S and Aila T. 2019. A style-based generator architecture for generative adversarial networks//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4396-4405 [DOI: 10.1109/CVPR.2019.00453]

Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J and Aila T. 2020. Analyzing and improving the image quality of StyleGAN//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 8107-8116 [DOI: 10.1109/CVPR42600.2020.00813]

Kim D, Woo S, Lee J Y and Kweon I S. 2019b. Deep video inpainting//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 5785-5794 [DOI: 10.1109/CVPR.2019.00594]

Kim J, Kim M, Kang H and Lee K H. 2020a. U-GAT-IT: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation//Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia: [s.n.]

Kim S W, Zhou Y H, Philion J, Torralba A and Fidler S. 2020b. Learning to simulate dynamic environments with GameGAN//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1228-1237 [DOI: 10.1109/CVPR42600.2020.00131]

Kingma D P and Welling M. 2014. Auto-encoding variational Bayes//Proceedings of the 2nd International Conference on Learning Representations. Banff, Canada: [s.n.]

Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z H and Shi W Z. 2017. Photo-realistic single image super-resolution using a generative adversarial network//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 105-114 [DOI: 10.1109/CVPR.2017.19]

Li B W, Qi X J, Lukasiewicz T and Torr P H S. 2020a. ManiGAN: text-guided image manipulation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 7877-7886 [DOI: 10.1109/CVPR42600.2020.00790]

Li Y H, Singh K K, Ojha U and Lee Y J. 2020b. MixNMatch: multifactor disentanglement and encoding for conditional image generation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and

Pattern Recognition. Seattle, USA: IEEE: 8036-8045 [DOI: 10.1109/CVPR42600.2020.00806]

Li Y T, Min M R, Shen D H, Carlson D E and Carin L. 2018. Video generation from text//Proceedings of the 32nd AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18). New Orleans, USA: AAAI: 7065-7072

Liao R J, Tao X, Li R Y, Ma Z Y and Jia J Y. 2017. Video super-resolution via deep draft-ensemble learning//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 531-539 [DOI: 10.1109/ICCV.2015.68]

Lin A S, Wu L M, Corona R, Tai K, Huang Q X and Mooney R J. 2018. Generating animated videos of human activities from natural language descriptions//Proceedings of the 32nd Conference on Neural Information Processing Systems. Montréal, Canada: Curran Associates Inc.

Liu M Y, Huang X, Mallya A, Karras T, Aila T, Lehtinen J and Kautz J. 2019a. Few-shot unsupervised image-to-image translation//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 10550-10559 [DOI: 10.1109/ICCV.2019.01065]

Liu Y, Wang X, Yuan Y T and Zhu W W. 2019b. Cross-modal dual learning for sentence-to-video generation//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France: ACM: 1239-1247 [DOI: 10.1145/3343031.3350986]

Liu Z M, Jia W, Yang M, Luo P Y, Guo Y and Tan M K. 2021. Deep view synthesis via self-consistent generative network [J/OL]. [2021-03-07]. IEEE Transactions on Multimedia, https://ieeexplore.ieee.org/document/9339999 [DOI: 10.1109/TMM.2021.3053401]

Mallya A, Wang T C, Sapra K and Liu M Y. 2020. World-consistent video-to-video synthesis//Proceedings of 2020 European Conference on Computer Vision. Glasgow, United Kingdom: Springer: 359-378 [DOI: 10.1007/978-3-030-58598-3_22]

Mao X D, Li Q, Xie H R, Lau R Y K, Wang Z and Smolley S P. 2017. Least squares generative adversarial networks//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2813-2821 [DOI: 10.1109/ICCV.2017.304]

Marwah T, Mittal G and Balasubramanian V N. 2017. Attentive semantic video generation using captions//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 1435-1443 [DOI: 10.1109/ICCV.2017.159]

Mathew S, Nadeem S, Kumari S and Kaufman A. 2020. Augmenting colonoscopy using extended and directional CycleGAN for lossy image translation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 4695-4704 [DOI: 10.1109/CVPR42600.2020.00475]

Maximov M, Elezi I and Leal-Taixé L. 2020. CIAGAN: conditional identity anonymization generative adversarial networks//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 5446-5455 [DOI: 10.1109/CVPR42600.2020.00549]

Men Y F, Mao Y M, Jiang Y N, Ma W Y and Lian Z H. 2020. Controllable person image synthesis with attribute-decomposed GAN//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 5083-5092 [DOI: 10.1109/CVPR42600.2020.00513]

Menon S, Damian A, Hu S J, Ravi N and Rudin C. 2020. PULSE: self-supervised photo upsampling via latent space exploration of generative models//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 2434-2442 [DOI: 10.1109/CVPR42600.2020.00251]

Mildenhall B, Srinivasan P P, Tancik M, Jonathan T, Barron J T, Ramamoorthi R and Ng R. 2020. Nerf: representing scenes as neural radiance fields for view synthesis//Proceedings of 2020 European Conference on Computer Vision. Glasgow, United Kingdom: Springer: 405-421 [DOI: 10.1007/978-3-030-58452-8-24]

Mittal G, Marwah T and Balasubramanian V N. 2017. Sync-DRAW: automatic video generation using deep recurrent attentive architectures//Proceedings of the 25th ACM International Conference on Multimedia. Mountain View, USA: ACM: 1096-1104 [DOI: 10.1145/3123266.3123309]

Nam S, Ma C, Chai M, Brendel W, Xu N and Kim S J. 2019. End-to-end time-lapse video synthesis from a single outdoor image//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 1409-1418 [DOI: 10.1109/CVPR.2019.00150]

Otberdout N, Daoudi M, Kacem A, Ballihi L and Berretti S. 2020. Dynamic facial expression generation on Hilbert Hypersphere with conditional Wasserstein generative adversarial nets [J/OL]. [2021-03-07]. IEEE Transactions on Pattern Analysis and Machine Intelligence, https://ieeexplore.ieee.org/document/9117185 [DOI: 10.1109/TPAMI.2020.3002500]

Plumerault A, Le Borgne H and Hudelot C. 2020. Controlling generative models with continuous factors of variations//Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia: [s. n.]

Qiao T T, Zhang J, Xu D Q and Tao D C. 2019. MirrorGAN: learning text-to-image generation by redescription//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 1505-1514 [DOI: 10.1109/CVPR.2019.00160]

Radford A, Metz L and Chintala S. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks//Proceedings of the 4th International Conference on Learning Representations. San Juan, USA: [s. n.]

Ren J, Hacihaliloglu I, Singer E A, Foran D J and Qi X. 2018. Adver-

sarial domain adaptation for classification of prostate histopathology whole-slide images//Proceedings of 2018 International Conference on Medical Image Computing and Computer-Assisted Intervention. Granada, Spain: Springer: 201-209 [DOI: 10.1007/978-3-030-00934-2_23]

Saito S, Huang Z, Natsume R, Morishima S, Li H and Kanazawa A. 2019. PIFu: pixel-aligned implicit function for high-resolution clothed human digitization//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 2304-2314 [DOI: 10.1109/ICCV.2019.00239]

Saito S, Simon T, Saragih J and Joo H. 2020. PIFuHD: multi-level pixel-aligned implicit function for high-resolution 3D human digitization//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 81-90 [DOI: 10.1109/CVPR42600.2020.00016]

Shaham T R, Dekel T and Michaeli T. 2019. SinGAN: learning a generative model from a single natural image//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 4569-4579 [DOI: 10.1109/ICCV.2019.00467]

Shen G Y, Huang W B, Gan C, Tan M K, Huang J Z, Zhu W W and Gong B Q. 2019. Facial image-to-video translation by a hidden affine transformation//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France: ACM: 2505-2513 [DOI: 10.1145/3343031.3350981]

Tewari A, Elgharib M, Bharaj G, Bernard F, Seidel H P, Pérez P, Zollhöfer M and Theobalt C. 2020a. StyleRig: rigging StyleGAN for 3D control over portrait images//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 6141-6150 [DOI: 10.1109/CVPR42600.2020.00618]

Tewari A, Elgharib M, Mallikarjun B R, Bernard F, Seidel H P, Pérez P, Zollhöfer M and Theobalt C. 2020b. PIE: portrait image embedding for semantic control. ACM Transactions on Graphics, 39(6): #223 [DOI: 10.1145/3414685.3417803]

Tian Y P, Zhang Y L, Fu Y and Xu C L. 2020. TDAN: temporally-deformable alignment network for video super-resolution//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 3357-3366 [DOI: 10.1109/CVPR42600.2020.00342]

Vondrick C, Shrivastava A, Fathi A, Guadarrama S and Murphy K. 2018. Tracking emerges by colorizing videos//Proceedings of 2018 European Conference on Computer Vision. Munich, Germany: Springer: 402-419 [10.1007/978-3-030-01261-8_24]

Wan Z Y, Zhang B, Chen D D, Zhang P, Chen D, Liao J and Wen F. 2020. Bringing old photos back to life//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 2744-2754 [DOI: 10.1109/CVPR42600.2020.00282]

Wandt B and Rosenhahn B. 2019. RepNet: weakly supervised training of an adversarial reprojection network for 3D human pose estimation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 7774-7783 [DOI: 10.1109/CVPR.2019.00797]

Wang J N, Zhao Y Y, Noble J H and Dawant B M. 2018a. Conditional generative adversarial networks for metal artifact reduction in CT images of the ear//Proceedings of 2018 International Conference on Medical Image Computing and Computer-Assisted Intervention. Granada, Spain: Springer: 3-11 [DOI: 10.1007/978-3-030-00928-1_1]

Wang L Z, Zhao X C, Yu T, Wang S T and Liu Y B. 2020a. NormalGAN: learning detailed 3D human from a single RGB-D image//Proceedings of 2020 European Conference on Computer Vision. Glasgow, United Kingdom: Springer: 430-446 [DOI: 10.1007/978-3-030-58565-5_26]

Wang T C, Liu M Y, Tao A, Liu G L, Kautz J and Catanzaro B. 2019c. Few-shot video-to-video synthesis//Proceedings of the 33rd Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates, Inc.: 5014-5025

Wang T C, Liu M Y, Zhu J Y, Liu G L, Tao A, Kautz J and Catanzaro B. 2018b. Video-to-video synthesis//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal Canada: Curran Associates Inc.: 1152-1164

Wang T C, Liu M Y, Zhu J Y, Tao A, Kautz J and Catanzaro B. 2018c. High-resolution image synthesis and semantic manipulation with conditional GANs//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 8798-8807 [DOI: 10.1109/CVPR.2018.00917]

Wang X T, Yu K, Wu S X, Gu J J, Liu Y H, Dong C, Qiao Y and Loy C C. 2018d. ESRGAN: enhanced super-resolution generative adversarial networks//Proceedings of 2018 European Conference on Computer Vision. Munich, Germany: Springer: 63-79 [10.1007/978-3-030-11021-5_5]

Wang Y H, Bilinski P, Bremond F and Dantcheva A. 2020b. G3AN: disentangling appearance and motion for video generation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 5263-5272 [DOI: 10.1109/CVPR42600.2020.00531]

Wang Y X, Khan S, Gonzalez-Garcia A, van de Weijer J and Khan F S. 2020c. Semi-supervised learning for few-shot image-to-image translation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 4452-4461 [DOI: 10.1109/CVPR42600.2020.00451]

Weng C Y, Curless B and Kemelmacher-Shlizerman I. 2019. Photo wake-up: 3D character animation from a single photo//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 5901-5910 [DOI: 10.1109/CVPR.2019.00606]

Wu S Z, Rupprecht C and Vedaldi A. 2020a. Unsupervised Learning of Probably Symmetric Deformable 3D Objects from Images in the

Wild//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1-10 [DOI: 10.1109/CVPR42600.2020.00008]

Xu R, Li X X, Zhou B L and Loy C C. 2019. Deep flow-guided video inpainting//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 3718-3727 [DOI: 10.1109/CVPR.2019.00384]

Xu T, Zhang P C, Huang Q Y, Zhang H, Gan Z, Huang X L and He X D. 2018. AttnGAN: fine-grained text to image generation with attentional generative adversarial networks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 1316-1324 [DOI: 10.1109/CVPR.2018.00143]

Yang Z Q, Zhu W T, Wu W, Qian C, Zhou Q, Zhou B L and Loy C C. 2020. TransMoMo: invariance-driven unsupervised video motion retargeting//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 5305-5314 [DOI: 10.1109/CVPR42600.2020.00535]

Zhang H, Goodfellow I, Metaxas D and Odena A. 2019. Self-attention generative adversarial networks//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR: 7354-7363

Zhang H, Xu T, Li H S, Zhang S T, Wang X G, Huang X L and Metaxas, D N. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks// Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 5907-5915[DOI:10.1109/ICCV.2017.629]

Zhang P Y, Wang F S, Xu W and Li Y. 2018. Multi-channel generative adversarial network for parallel magnetic resonance image reconstruction in K-space//Proceedings of 2018 International Conference on Medical Image Computing and Computer-Assisted Intervention. Granada, Spain: Springer: 180-188 [DOI: 10.1007/978-3-030-00928-1_21]

Zhao A, Balakrishnan G, Lewis K M, Durand F, Guttag J V and Dalca A V. 2020. Painting many pasts: synthesizing time lapse videos of paintings//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 8432-8442 [DOI: 10.1109/CVPR42600.2020.00846]

Zhu J Y, Park T, Isola P and Efros A A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2242-2251 [DOI: 10.1109/ICCV.2017.244]

Zhu M F, Pan P B, Chen W and Yang Y. 2019. DM-GAN: dynamic memory generative adversarial networks for text-to-image synthesis//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 5795-5803 [DOI: 10.1109/CVPR.2019.00595]

Zhu Q L, Bi W, Liu X J, Ma X Y, Li X L and Wu D P. 2020a. A batch normalized inference network keeps the KL vanishing away//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Seattle, USA: Association for Computational Linguistics: 2636-2649

Zhu Y Z, Min M R, Kadav A and Graf H P. 2020b. S3VAE: self-supervised sequential VAE for representation disentanglement and data generation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 6537-6546 [DOI: 10.1109/CVPR42600.2020.00657]

作者简介

谭明奎,1983 年生,男,教授,主要研究方向为机器学习、视觉生成、大数据分析和大规模凸优化。
E-mail: mingkuitan@scut.edu.cn

许守恺,男,博士研究生,主要研究方向为视觉生成和模型压缩。E-mail: sexsk@mail.scut.edu.cn
张书海,男,硕士研究生,主要研究方向为视觉生成和对抗攻击。E-mail: mszhangshuhai@mail.scut.edu.cn
陈奇,男,博士研究生,主要研究方向为深度学习与视觉生成。E-mail: sechenqi@mail.scut.edu.cn