GAN 生成对抗网络研究综述

20121034 胡才郁

(计算机工程与科学学院)

摘　要　本文首先介绍了生成对抗网络的基本结构，列举了 GAN 模型优缺点。其次介绍了生成模型的评价标准以及基于 GAN 的各种衍生模型。通过介绍近年来 GAN 在计算机视觉方向的贡献，如图像超分辨率重构、风格迁移等应用，分析其研究现状和发展趋势，展开说明了每种应用的理论改进之处、优点及使用场景进行了总结，最后对于 GAN 的未来发展进行探讨。

关键词 生成对抗网络；生成模型；生成器；判别器；计算机视觉

1 基本 GAN 模型简介

1.1 基本GAN模型的结构

生成对抗网络（GAN）可以拆分为两个模块：一个是判别网络，另一个是生成网络。简单来说就是，有一些真实数据，同时也有一些随机生成的假数据。生成网络 G 生成假数据，并将其模仿成真实数据，混合到真实数据里。而辨别网络 D 的任务为把真实数据和假数据区分开。

用形象的说法为，艺术家(生成网络 G)学习创造看起来真实的图像，而艺术评论家(判别网络 D)学习区分真假图像。训练过程中，生成网络在生成逼真图像方面逐渐变强，而判别网络在辨别这些图像的能力上逐渐变强。当判别器不再能够区分真实图片和伪造图片时，训练过程达到平衡。而生成网络(艺术家)即为可以完成生成任务的生成模型。

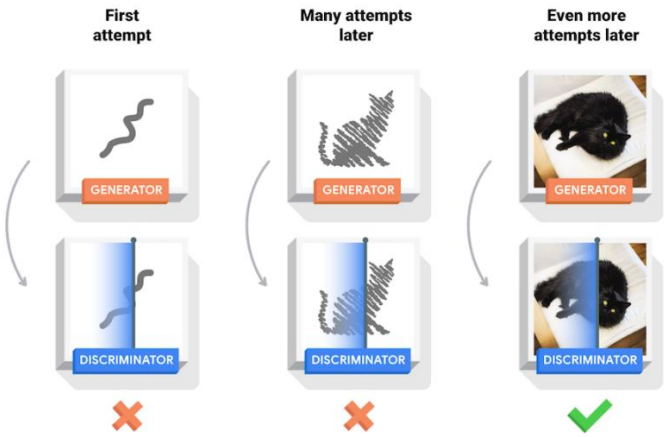

图 1 GAN 示意图

生成对抗网络由两个重要部分组成，即生成网络 G 与判别网络 D。生成网络 G 通过机器生成数据，目的是学习真实样本的分布，生成相似度逼近真实样本的伪样本，而判别网络 D 的作用是区分从数据获取的真实样本和由生成网络 G 生成的伪样本。两个模型通过不断地对抗训练来迭代优化，使生成网络 G 生成的数据分布最大可能接近于真实数据分布，当判别网络 D 中，每次输出的概率基本为 1/2，即判别器相当于随机猜测样本是真是假，此时说明模型达到了博弈论中的纳什均衡，即最优状态，这就是 GAN 的对抗性思想。

GAN 模型的训练过程可以分为 3 个阶段，第一阶段：固定判别器 D，训练生成器 G，第二阶段：固定生成器 G，训练判别器 D，然后循环阶段一和阶段二，不断进行训练。基本 GAN 模型结构如下图所示：

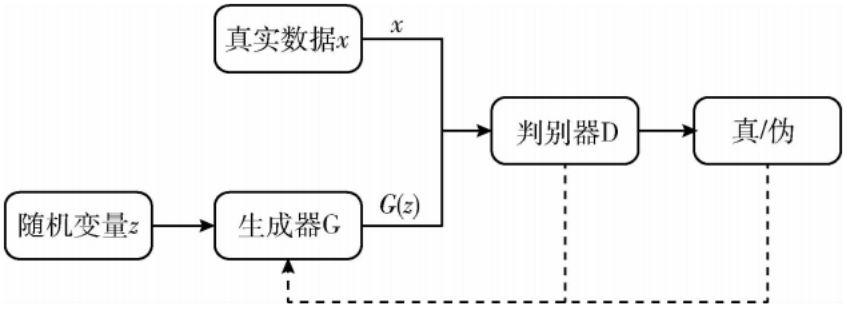

图 2 基本 GAN 模型结构

而在形式上，生成器 G 和判别器 D 的优化过程可以定义为二元极小极大博弈问题，其目标函数如下所示：

$$\min_G \max_D V(D,G) = E_{x\sim p_{\mathrm{dt}}(x)}[\log D(x)] + E_{z\sim p_z(x)}\left[\log\left(1 - D\left(G(z)\right)\right)\right] \tag{1}$$

式中：p_{dt} 代表真实的数据分布，$p_z(x)$代表生成的数据分布，训练生成器以最小化目标函数，而训练判别器以使其最大化，最终使生成器生成类似于真实数据的数据。

1.2 生成网络

生成器本质上是一个可微分函数，生成器接收随机变量z的输入，经生成器G后生成假样本$G(z)$。在 GAN 中，生成器对输入变量：基本没有限制，z通常是一个 100 维的随机编码向量，z可以是随机噪声或者符合某种分布的变量。生成器理论上可以逐渐学习任何概率分布，经训练后的生成网络可以生成逼真图像，但又不是和真实图像完全一样，即生成网络实际上是学习了训练数据的一个近似分布，这在数据增强应用方面尤为重要。

1.3 辨别网络

判别器同生成器一样，其本质上也是可微分函数，在 GAN 中，判别器的主要目的是判断输入是否为真实样本，并提供反馈以指导生成器训练。判别器和生成器组成零和游戏的两个玩家，为取得游戏的胜利，判别器和生成器通过训练不断提高自己的判别能力和生成能力，游戏最终会达到一个纳什均衡状态，此时生成器学习到了与真实样本近似的概率分布，判别器已经

不能正确判断输入的数据是来自真实样本还是来自生成器生成的假样本$G(x)$。即判别器每次输出的概率值都是 1 / 2。

2 GAN 模型的优势

由于生成器是一个简单的、确定的前馈网络，因此 GAN 可以用简单的方式对生成的数据进行采样，无需在学习过程中进行推断，因此在采样上计算较快且准确率较高。此外，GAN 模型可以并行生成样本，而无需利用与输入维数成比例的运行时间。再者，GAN 能够更好建立数据分布的模型，且已广泛应用于图像处理领域，因为事实证明能够很好的与图像配合使用，能生成更清晰逼真锐利的图像。理论上，GAN 由于也是神经网络，可以拟合任意一种函数，训练任何一种生成器网络，具有更加灵活的框架，与传统神经网络需要构建一个损失函数相比，GAN 可以学习损失函数。

GAN 具备以下优势：

（1）能学习真实样本的分布，探索样本的真实结构，且具有更强大的预测能力；

（2）不同于一般的机器学习模型，GAN 对生成样本非常的稳定；

（3）通过 GAN 生成以假乱真的样本，缓解了小样本机器学习的困难；

（4）为指导人工智能系统完成复杂任务提供了一种全新的思路；

（5）GAN 与传统神经网络的一个重要区别是传统神经网络需要人工精心设计和建构一个损失函数，而 GAN 可以学习损失函数。

3 基本 GAN 模型的问题与挑战

尽管 GAN 有许多优点，但是在实际训练 GAN 模型时，如何寻找纳什平衡的解是一个挑战和难点，在实际训练中容易出现判别器 D 收敛、生成器 G 发散的现象，两个网络之间很难出现很好的同步，因此 GAN 面临着难训练、不稳定的问题。

3.1 训练集样本的影响

神经网络的表现主要取决于模型自身的特点，以及训练使用的真实样本集。同样，GAN 模型的训练学习的质量也受制于训练样本集的影响。

一方面，样本集的自身内在数据分布情况可能会影响 GAN 的训练效率和生成质量。例如，文献在样本集上定义了类内距离集与类间距离集，并依此提出基于距离的可分性指数，用于量化样本可分性，并指出当不同种类样本按相同分布混合时最难以区分，使用这种样本集进行有监督学习时很难使模型有较好表现。这对于 GAN 的样本生成质量评价指标设计具有借鉴意义。

另一方面，GAN 模型的一大特点是学习真实样本分布，因此需要足够多真实样本进行训练才能有较好表现，研究如何使用小规模训练集得到较好的 GAN 模型是具有挑战和意义的。GAN 模型对训练集质量也有较高要求，而高质量的数据集往往难以获得，因此研究哪些数据

会影响模型表现，如何规避低质量样本带来的负面影响，以降低对训练集质量的高要求，也是目前热门的研究方向。

3.2 模式崩溃问题

模式崩溃是指，生成器生成不了多样性的样本，而是生成了与真实样本相同的样本，这种缺陷对于数据增强而言是致命的。例如在生成人脸图片的实验中，无法生成多种风格的人脸，只能生成某一种风格的人脸。

GAN 存在模式崩溃现象，当生成器 G 学习到一个参数设置，可以生成对判别器 D 而言特别逼真的样本，由于此样本很容易骗过判别器 D，所以生成器 G 可能会一次又一次的生成相同的伪样本，最终始终生成同样的样本点，出现模式缺失，无法继续学习。与普通神经网络训练过程相比，GAN 模型中存在生成器 G 与判别器 D 之间的博弈机制，这使得 GAN 模式崩溃问题变得复杂。

3.3 梯度消失问题

梯度消失也是 GAN 存在的问题之一，如果判别器 D 始终能够正确判断真实样本为真和生成样本为假，那么无论生成器 G 生成的样本多么好，判别器 D 都可以把它们分类为假样本，此时损失降为零，导致生成器 G 没有学习，就产生了梯度消失现象。

真实分布是一个高维分布，而生成分布来自于一个低维分布，所以其实很有可能生成分布与真实分布之间就没有重叠的部分。除此之外，不可能真正去计算两个分布，只能近似采样，所以也导致了两个分布没有重叠部分。如果辨别器 D 训练得太好，那么生成分布和原来分布基本没有重叠部分，这就导致了梯度消失；如果判别器 D 训练得不好，这样生成器得梯度又不准，就会出现错误得优化方向,因此 GAN 难以训练。

4 GAN 评估方法

对于生成问题的评估并不像评估分类、回归问题一样简单，有准确率、查准率等明确的指标。目前的评估方法主要是从定性评估与定量评估两个方面进行。

4.1 定性评估

定性评估的方法主要还是靠人的眼睛来进行判断.一般的做法是将真实图像和生成图像直接由让人来判断图像的真假，并且给出两者的相似程度，最后根据打分的结果统计一个最终的指标。

在实践中由于人的主观性是很强的，每个人的标准是不一致的，导致定性评估不是一个通用的标准。视觉检查在评估一个模型对数据的拟合程度时，在低维度数据的情况下可以工作得很好，但是在高维度数据的情况下，这种直觉性可能会导致误导。

4.2 定量评估

对于 GAN 这类生成模型而言，好的评价指标应偏向如下特点：生成样本与真实样本相似；生成样本在类内、类间保持多样化；模型在隐空间采样可控；对改变样本语义的失真和变化敏感。同时指标自身应该具有的特点包括：有明确的值域且值的大小能够反映对模型较好或较差的评价、对样本数量的需求低、计算复杂度低等。以下为两种常用于 GAN 生成图像的模型定量评估指标。

（1）IS 指标

质量较高的生成样本更容易被明确地分类，且生成样本在各个类中均匀分布时样本多样性较好，这就是 IS 评价指标的核心思想。以使用分类任务中，性能较强的 Inception-v3 模型作为分类器为例，预训练好的模型接收一幅图像作为输入，输出一个 1000 维向量，每一维值为输入属于某一类别的概率。样本质量越高，输出向量中的某一个值越趋近于 1，而其他值越趋近于 0。这是一个二元分类任务，样本多样性越好，样本在不同分类上的分布越趋向均匀分布。

（2）FID 指标

FID 的主要思想是：既然预训练网络模型可以提取样本特征信息，那么分别提取真实样本与生成样本特征信息，假设特征符合多元高斯分布，再计算分布间的弗雷歇距离，距离近则生成图片质量较高，多性较好。弗雷歇距离值为下式：

$$|\mu_{data} - \mu_g| + \left(\Sigma_{data} + \Sigma_g - 2\left(\Sigma_{data}\Sigma_g\right)^{\frac{1}{2}}\right) \tag{2}$$

FID 度量方式的思想和人类判断是一致的，该评价指标值越小，表示生成的图像越接近真实图像，生成的图片质量越好。FID 和生成图像的质量之间有很强的负相关性；该度量方式的优势在于其对噪声不是很敏感，而且可以检测出类内的模式崩溃的问题。

5 GAN 的发展及其衍生模型

研究者们针对 GAN 存在训练困难等问题，通过不断探索，最终提出了很多基于 GAN 的变体。下面主要对一些典型的 GAN 变体进行介绍。

5.1 DCGAN

深度卷积生成对抗网络（DCGAN），将卷积神经网络（CNN）应用到生成对抗网络中，通过对 GAN 的体系结构更改，提高了 GAN 的训练的稳定性。在 DCGAN 中，对 GAN 的体系结构进行了一些修改：将空间池化层函数替换为跨卷积；去除了完全连接层，能提高模型的稳定性；除了生成器的输出层和判别器的输入层之外，对每个单元的输入进行批归一化操作；在生成器中使用 ReLU 激活函数，在其输出层使用 Tanh 函数；在判别器中使用 ReLU 的变体 LeakyReLU 激活函数。DCGAN 具有更强大的生成能力，训练也更稳定，生成的样本具有更多

的多样性，因此，很多对于 GAN 的改进都是基于 DCGAN 的结构。DCGAN 只是对 GAN 模型的结构进行了改进，对生成器和判别器进一步的细化，并没有对优化方法进行改进。

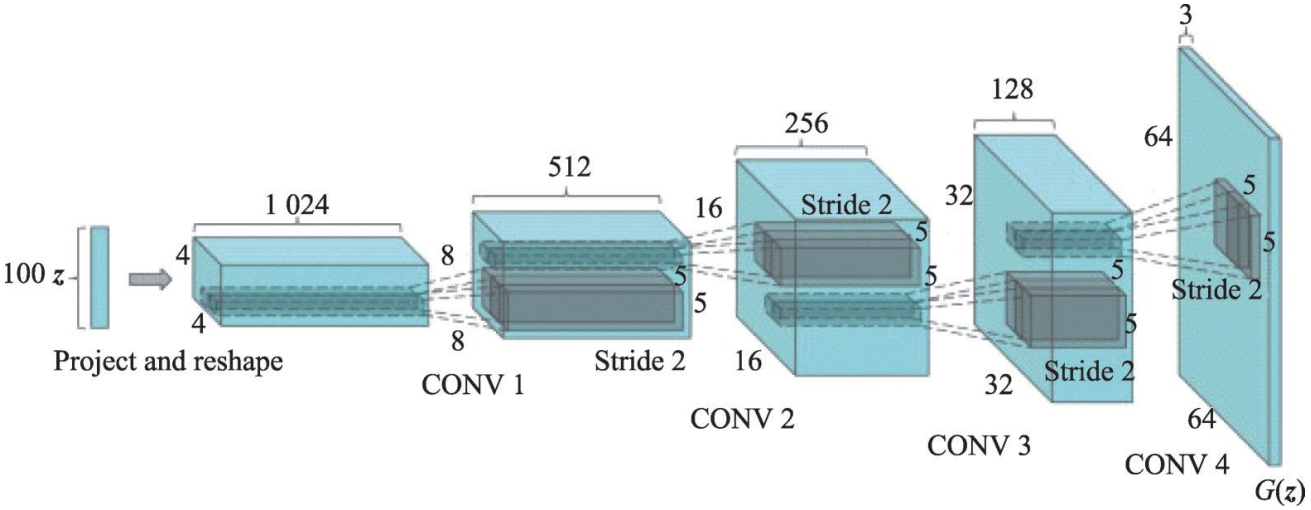

图 3 DCGAN 生成模型架构图

5.2 INFOGAN

GAN 强大的学习能力最终可以学习到真实样本的分布，但对输入噪声信号z和数据的语义特征之间的对应关系不清楚。一个理想的情况是清楚它们之间的对应关系，这样就能通过控制对应的维度变量来达到相应的变化。比如对于 MNIST 手写数字识别项目，在知道其对应关系的情况下，可以控制输出图像的光照、笔画粗细、字体倾斜度等。INFOGAN 解决了这个问题，它将输入噪声z分成两部分，一部分是噪声信号z，另一部分是可解释的有隐含意义的信号c。INFOGAN 模型结构图如下图所示。

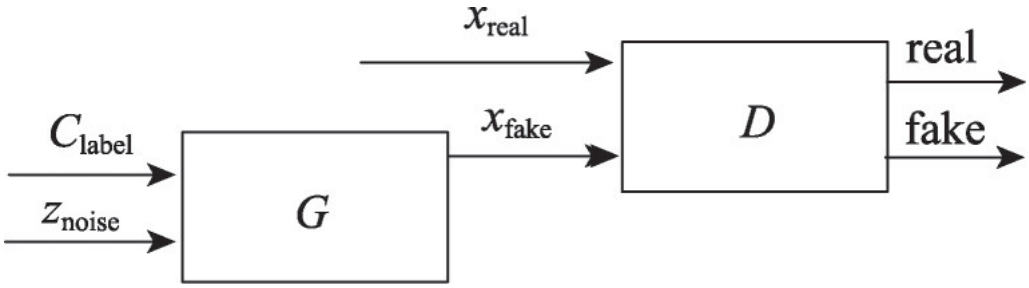

图 4 INFOGAN 生成模型架构图

生成器的输入多了一个隐含变量$C(c_1, c_2, c_3, \ldots, c_L)$，它代表的是如上面提到的图像的光照、笔画粗细、字体倾斜度等图像的语义特征信息。INFOGAN 对目标函数进行了约束，即c和$G(z, c)$之间的互信息，如以下公式：

$$\min_G \max_D V_I(D, G) = V(D, G) - \lambda I\big(c; G(z, c)\big)$$

实验证明了 INFOGAN 确实学到了一些可解释的语义特征，通过控制这些特征可以生成想要的数据。如下图，通过控制图像的角度、宽度，生成形状不一样的数据。

(a) Result of INFOGAN after adding width and inclination information to c

(b) INFOGAN generation result after adding data type information to c

图 5 INFOGAN 在 MNIST 数据集上实验结果

图(a)在c中加入了宽度与倾斜度的信息，INFOGAN 中生成的数字中也对应宽度与倾斜度改变；图(b)在c中加入了除 MNIST 数据集手写数字外的其他数据信息，则 INFOGAN 中生成的数字中展现出了除原始 MNITST 数据集外的其他特征。INFOGAN 具有特征学习的能力，能够提前数据集中的特征。

5.3 LSGAN

LSGAN，又叫最小二乘生成对抗网络，LSGAN 指出使用 JS 散度并不能拉近真实分布和生成分布之间的距离，使用最小二乘可以将图像的分布尽可能的接近决策边界，因此 LSGAN 使用了最小二乘损失函数代替了 GAN 的损失函数。

该做法有效改善了传统 GAN 生成图片质量不高，训练不稳定的问题。最小二乘损失与交叉熵损失相比，优势在于生成样本在欺骗判别器的前提下同时让生成器把距离决策边界比较远的生成图片拉向决策边界，这样保证了生成高质量的样本。LSGAN 以交叉熵作为损失，会使得生成器不会再优化那些被判别器识别为真实图片的生成图片，即使这些生成图片距离判别器的决策边界仍然很远，也就是距离真实数据比较远，因为此时的交叉熵损失已经很小，生成器完成了为它设计的目标。

6 GAN 在计算机视觉领域的应用

计算机视觉是目前人工智能研究的重要领域，而 GAN 在计算机视觉的许多方面都表现非凡，从最初的图像生成，到后面的一系列应用，越来越多新的 GAN 框架被提出并应用到新的领域，由于 GAN 自身的对抗特性它能不断地自我提升，在生成样本领域取得了比传统方法更显著的效果。下面将介绍 GAN 在计算机视觉上的应用及为了实现目标任务在结构上做出的改变。

6.1 图像超分辨率

图像超分辨率是指由一幅低分辨率图像或图像序列恢复出高分辨率图像，此技术分为超分辨率复原和超分辨率重建。

图像超分辨率一直是计算机视觉领域的一个研究热点，SRGAN 是 GAN 的一个变体，也是 GAN 在图像超分辨率应用上的一个成功案例。SRGAN 基于相似性感知方法提出了一种新的损失函数，有效解决了恢复后图像丢失高频细节的问题，并使人能有良好视觉感受。SRGAN 从特征上定义损失，它将生成样本和真实样本分别输入 VGG-19 网络，生成网络包含 5 个残差块,每个残差块包含两个 3x3、64 特征的卷积层，后接 BN 层，激活函数选择 PReLU。残差块后面两个 2 倍的分步幅卷积层用来增大特征尺寸。判别网络部分包含 8 个卷积层，随着网络层数加深，特征个数不断增加，特征尺寸不断减小，最终通过两个全连接层和一个 sigmoid 激活函数得到判断为自然图像的概率。

SRGAN 的损失函数为感知损失函数，它包括加权的两部分，一部分是内容损失，另一部分是对抗损失函数。内容损失函数采用特征空间的最小均方差来表示。对抗损失函数和判别器输出的概率有关。

下图是应用在图像超分辨率上的网络模型实验的结果，由于 SRGAN 的网络架构与损失函数便于优化，生成的超分辨率图像明显优于其他模型。

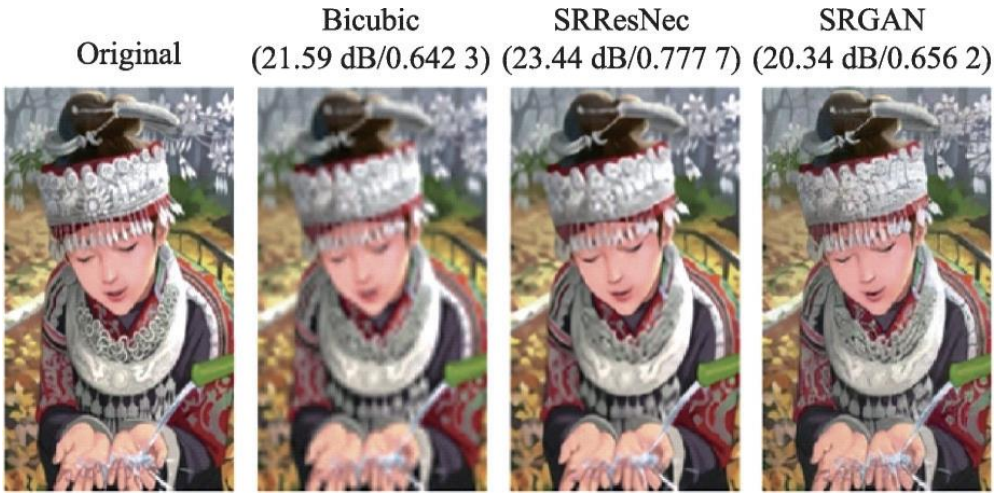

图 6 超分辨率图像生成

6.2 风格迁移

图像的风格迁移就是将一幅图像转换为另一种风格的图像，如将一张照片图像经过 GAN 处理，自动转化为油画风格的图像。深度学习最早是使用 CNN 框架来实现的，但这样的模型存在训练速度慢,对训练样本要求过高等问题。由于 GAN 的自主学习和生成随机样本的优势，以及降低了对训练样本的要求，使得 GAN 在图像风格迁移领域取得了丰硕的研究成果。

如果输入和输出的训练图像是同一场景，只是表现风格不同，这是一种“配对”训练。GAN 的变种 CycleGAN 利用循环 GAN 进行无监督地从一种风格图像转化到另一种风格的图像，是

一种无配对的训练方法：输入和输出的训练图像不仅风格不同，而且内容也不同。这种方法比配对训练方法难度更大，但训练数据的来源也大大扩展了。如下图中将照片风格转换为不同艺术家的油画风格。

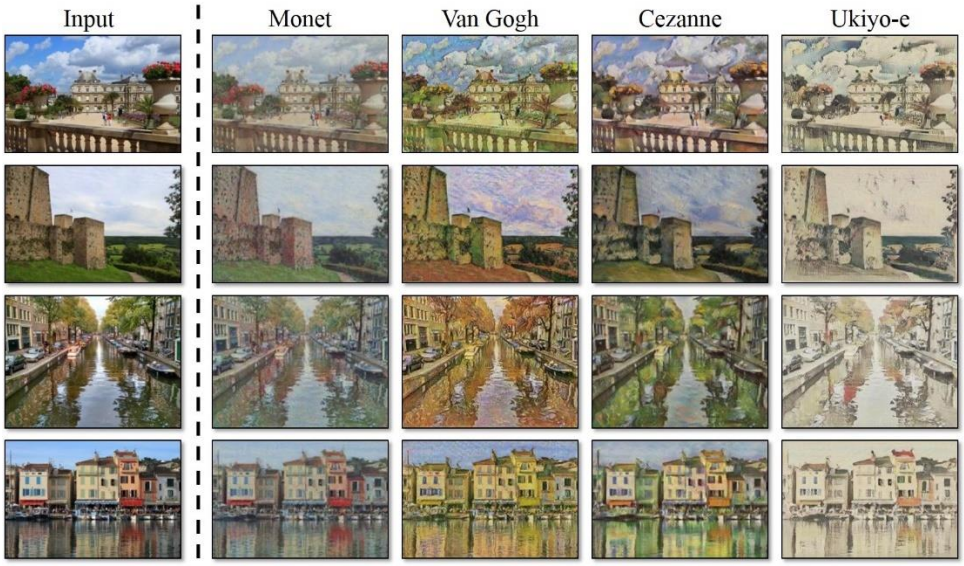

图 7 风格迁移

但是 CycleGAN 仍然存在一些问题，例如在将马转化为斑马的时候，很可能会将骑马的人也加上斑纹。因为原始训练集都是大自然中的马，因此模型学习不到人的相关知识。并且发现有些时候模型会将一些比较突出的背景也加上斑纹，如石头，醒目的植物等等，因此在稳定性方面还是有提升空间的。此外，CycleGAN 不易实现形状的改变，比如将猫改成狗，外表形状上必然要发生变化，并且 CycleGAN 不易训练，训练周期较长，这也是可以改进之处。

6.3 图像修复

图像修复是图像处理中的一个经典的应用领域，传统的图像修复方法通常是根据周围的像素点估计待修复的像素点，但是这种算法大多比较复杂，而且对于大面积的图像损毁很难修复。由于 GAN 是通过对抗博弈的方式来进行训练的，所以在图像修复方面不用受限于可用的图像统计信息，且能使得修复效果更加自然。

SNGAN 提出了一种生成式图像修复系统，可以使用自由形式的掩模和输入来完成图像。该系统基于从数百万张图像中学习的卷积，无需额外的标记工作，解决了将所有输入像素都视为有效像素的问题，使训练快速稳定。但由于 GAN 模型本身存在的问题，导致在修复图像时，可能会出现过度平滑或模糊的情况。

GAN 技术的引入提高了修复后图像的质量。在各种有损图像的修复中，人脸图像的修复是一种相对比较困难，但有广泛应用需求的技术。GAN 用作人脸补全，主要由一个生成器、两个判别器和一个解析网络组成。生成器实际上是一个自编码器，输入为有损的整个脸部图像。一个判别器鉴别输入的整个真实图像和生成图像的真伪，另一个判别器鉴别输入的遮挡部分的

生成图像和真实图像的真伪。生成图像的非遮挡部分用真实图像的对应像素替代。解析网络将脸部分割为不同个语义部位（如嘴、鼻、眼、…），确定每个像素所属的部位。这种方式可以同时获取补全后的完整图像和补全部分的信息，避免模型出现仅仅关注补全那一部分时带来的误判。下图是人脸补全 GAN 网络结构：

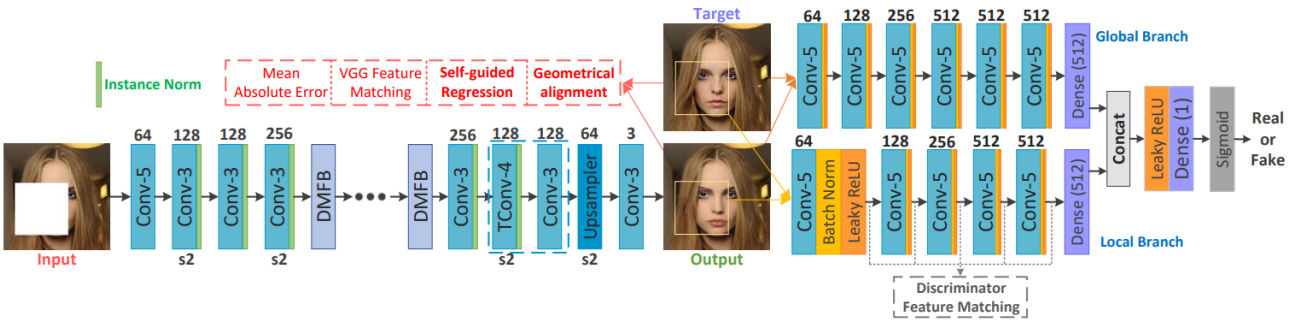

图 8 人脸补全 GAN 网络结构

它基于神经网络直接生成缺失区域的内容，通过引入重建损失，两个对抗性损失和语义解析损失的组合进行训练，确保了像素忠实度和局部全局内容的一致性。它能处理任意形状的大面积缺失像素，并产生逼真的面部完成结果。在人脸图像数据集 FFHQ 上的实验展示了这种方法对关键部位缺失的人脸图像的高质量补全的结果。

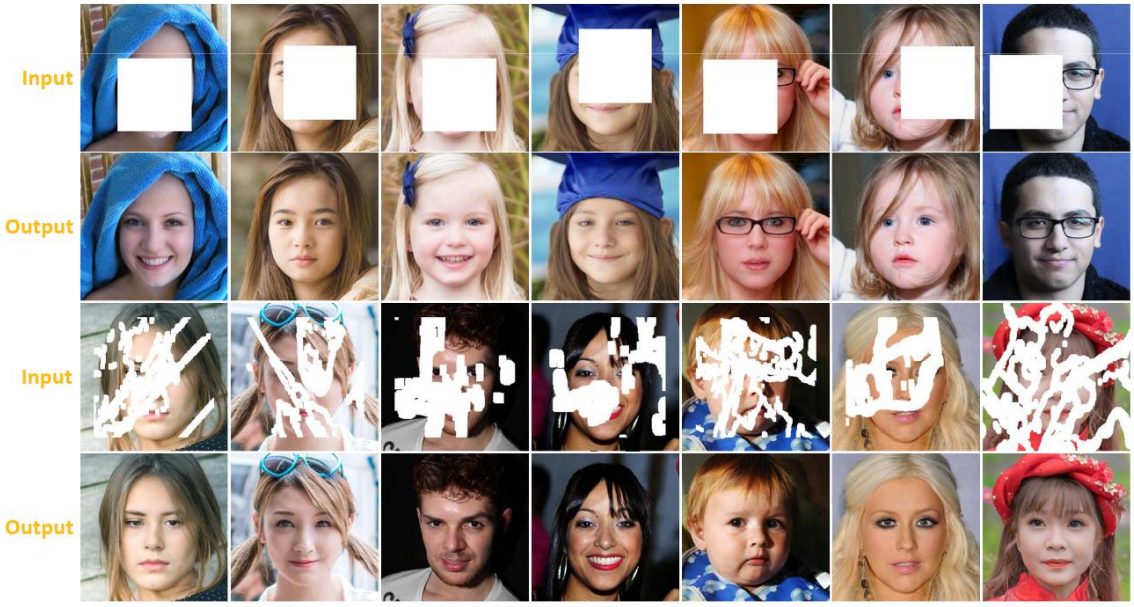

图 9 人脸补全效果

6.4 视频预测

由于基于生成对抗网络做图像生成可以保持图像的细节纹理特征。由于自然场景的内容复杂和运动变化，对视频中将来帧（future frame）的预测是一项具有挑战性的任务，也是无监督

视频表示学习的关键技术之一。现存的预测方法致力于对像素值的直接估计，和实际的将来帧有一定的差距，容易形成模糊的预测结果。

图像模糊是视觉任务中经常遇到的一个问题，比如：图像数据采集过程中由于物体运动导致的模糊，目标跟踪中相机的运动导致的模糊等。

dual motion GAN 是一种双重运动 GAN 结构。为了使合成的将来的视频帧与真实的视频帧没有区别，这种 GAN 通过双重对抗训练机制，迫使预测的将来帧和视频中像素流（保持一致。为了更好地 预测，最初的将来帧预测和双重的将来流预测形成一个闭环，相互之间传递反馈信号。

(a) 真实帧

(b) 预测帧

图 10 视频预测的关键帧处理

7 结语

本文对基本 GAN 模型的结构进行了介绍，并分析了 GAN 的优缺点与评价方法。其次，总结介绍了部分 GAN 的衍生模型。

生成对抗网络 GAN 由生成网络 G 与判别网络 D 组成，有很强的生成能力，主要应用在计算机视觉领域。现在 GAN 可以产生超分辨率图像、完成图像的风格迁移、进行图片补全、预测视频关键帧等等。GAN 为人工智能注入了活力，特别是计算机视觉领域，其无监督学习的生成对抗方法极大促进了该领域的发展。它作为一种生成模型，对于解决样本不足、生成质量差、提取特征难度大等问题提供了一种较好的解决方案。对基于深度学习的生成对抗网络在计算机视觉方面的应用进行了分析总结，不仅深入分析了 GAN 在理论模型方面的改进，而且重点介绍了 GAN 在视觉方面的几类突出的应用。

GAN 的训练类似于人类一理解周围复杂的世界的过程，其对抗生成的思想与人的思维模式异曲同工。在 GAN 这个方向继续探索，或与有可能创造出具有认知的机器学习模型，在其奥妙之中，有着许多机会进行理论探索和技术开发，其中也蕴藏着大量创新引用的机会。

致谢 十分感谢在百忙之中抽出时间审阅本文的老师。也因为《计算机科学进展》这门课，我有机会探索神秘而又有趣的 GAN。由于本人的学识和写作的水平有限，在本文的写作中难免有僻陋，恳请老师多指教。

春季学期注定对我们来说是个难以忘记的学期，它有着两个学期的考试都安排在一起的紧张刺激，也有着疫情突如其来的措手不及。不知不觉，这也是我足不出校的三个多月。查询了学校的健康之路我才发现，自 3 月 16 日以来，已经做过了 40 次核酸，以及不知道多少次抗原。 希望疫情早日结束，大家都能回到正常的生活轨迹之中。

参 考 文 献

[1] Goodfellow, Ian J., Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron C., and Bengio, Yoshua. Generative adversarial nets. NIPS, 2014.

[2] Zheng Z , Zheng L , Yang Y . Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in VitroC// 2017 IEEE International Conference on Computer Vision (ICCV). IEEE Computer Society, 2017.

[3] GIU-GANs Global Information Utilization for Generative Adversarial Networks[J/OL]

[4] S. Gao, Q. Han, D. Li, M. Cheng, P. Peng, Representative batch normalization with feature calibration, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2021, pp. 8669–8679.

[5] 高健鑫. 基于深度学习的移动应用界面色彩视觉风格迁移研究[D].四川大学,2021.DOI:10.27342/d.cnki.gscdu.2021.000768.

[6] 林野. 基于生成对抗网络的跨域人脸合成研究和应用[D].四川大学,2021.DOI:10.27342/d.cnki.gscdu.2021.000741.

[7] 耿鑫. 基于生成对抗网络的去雾算法研究[D].南京邮电大学,2021.DOI:10.27251/d.cnki.gnjdc.2021.001137.