图像处理与分析调研报告

学号：162140117

姓名：陈思远

目录

写在前面：

撰写该调研报告时，我并没有完全按照原报告中陈列的 8 点进行分点陈述，而是依照相对通畅的思路进行分点叙述，但完整包含了文档中 8 点要求的内容。

1 前言

本次大作业对图像处理与分析展开相关调研。在第二部分中列举并简要介绍了图像处理与分析的若干应用方向。在第三部分中针对**图像生成方向**进行深入调研，简要粗浅地介绍了该方向自出现以来的经典生成方法、目前流行方法，总结了国内外的研究现状。在第四部分中选择最近发表于 SIGGRAPH '23 的论文**《Drag Your GAN: Interactive Point-based Manipulation on the Generative Image Manifold》**，从研究动机、研究方法和模型优点、局限性等方面较为深入地阐述和分析。最后，我阐述了我在学习该课程后的收获。

2 图像处理与分析应用方向

近年来,图像处理与分析技术被广泛应用到计算机视觉领域，用于图像预处理。具体来说，在计算机视觉任务中，图像数据往往需要进行预处理，以提升后续模型的性能。常见的图像预处理技术包括:

- 去噪技术：可以使用均值滤波、中值滤波、高斯滤波等方法去除图像中的噪点和噪声;
- 影像增强：通过调整对比度、亮度等参数，可以增强图像的质量，使目标对象更加醒目，更易于识别;
- 图像裁剪：自动检测图像的主要区域，将图像裁剪为感兴趣的区域，去除不相关的背景，减少计算量;
- 尺度归一化：调整图像到统一的尺度，利于模型特征提取和分类;
- 图像增广：利用旋转、翻转、颜色干扰等操作，对图像数据集进行扩增，增强模型的鲁棒性。

通过图像预处理，可以减少图像数据中的噪声，突出主要信息，使计算机视觉模型更准确地理解图像内容和细节，从而提高后续任务的性能。由于图像处理与分析技术的重要性，图像处理与分析的应用方向涵盖了几乎所有计算机视觉研究领域，主要包括：图像分类、目标检测、图像分割、图像增强、图像生成、图像理解等。

2.1 图像分类

图像分类的核心是从给定的分类集合中给图像分配一个标签的任务，即分析一个输入图像并输出一个或者多个对应图像中主体元素的标签，分别对应单标签图像分类与多标签图像分类，一般情况下，标签来自预先定义的类别集。单标签图像分类基本上可以分为三大类别：跨物种语义级别的图像分类（object 属于不同物种）、子类细粒度图像分类（object 属于同一物种的不同子类）、实例级图像分类。

具体说来，图像分类可以是对不同类物体的分类，如图 1 中展现的 10 个种类，也可以是对同一类物体进一步细粒度的分类，比如对狗的品种的分类、手写体数字的分类。传统的图像分类算法主要通过对图像像素值进行统计进行检测，如边缘检测，形态分析，霍夫变换，斑点检测，角点检测，各种图像阈值化技术等，结合如 SVM、决策树这类传统的机器学习分类器进行图像分类，这类方法十分依赖于特征工程，分类性能因此受到较大限

制。

2012 年，AlexNet[2]将卷积神经网络(CNN)引入图像分类，使得人们重新开始关注 CNN。而后 ImageNet 大型数据集的兴起、VGGNet[3]、ResNet[4]等深度卷积神经网络的出现，大幅提升了图像分类模型的性能，将该任务带向高潮。

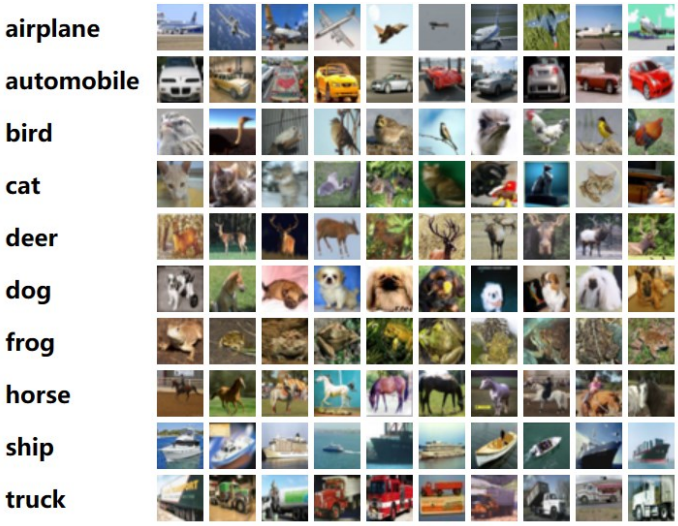


图 1 图像分类领域广泛使用的 CIFAR-10 数据集

2.2 目标检测

目标检测任务本质也是对图像中的物体进行识别，而这个任务与图像分类任务的区别在于前者输出长度不固定，即图像中物体的个数并不确定，包含多类别的主体而非一张图片即为一个主体。通常来说，目标检测任务要求输入一张图像，模型能给出图像中包含物体的类别，并在图像中使用线框标注出物体所在位置。这要求模型同时考虑语义信息和空间信息，难度相对来说高于图像分类。

在目标检测领域，传统方法基于手工设计特征+浅层分类器/回归器，如 Haar 特征+级联分类器(Viola Jones)、HOG 特征+SVM。R-CNN 系列算法将深度学习引入目标检测，大幅提高了检测准确度，但计算代价大且不是端到端模型，应用困难。而后的 YOLO 系列模型很好地解决了这些问题，它提出了 darknet-19 网络作为特征提取器，使用深度学习框架 PyTorch 进行模块化设计，同时使得更新迭代速度提升。

图 2 目标识别结果展示图

2.3 图像分割

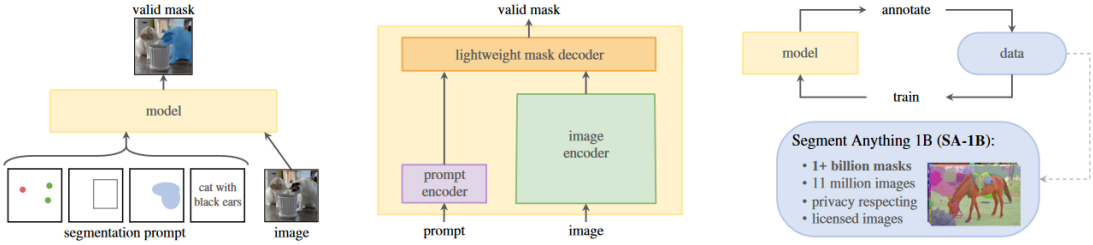
图像分割是计算机视觉研究中的一个经典难题，而且由于图像分割是图像分析的第一步，是计算机视觉的基础，是图像理解的重要组成部分，已经成为图像理解领域关注的一个热点。所谓图像分割是指根据灰度、彩色、空间纹理、几何形状等特征把图像划分成若干个互不相交的区域，使得这些特征在同一区域内表现出一致性或相似性，而在不同区域间表现出明显的不同。简单的说就是在一副图像中，把目标从背景中分离出来。

传统的分割方法例如基于阈值的分割方法，即通过对图像进行阈值化将图像中主体与背景分离，这种方法显然只适用于主体部分与背景灰度有较大差异的情况；基于边缘检测的图像分割算法试图通过检测包含不同区域的边缘来解决分割问题，主要利用了不同区域的边界上像素的灰度值变化比较剧烈这一先验知识，但是这种先验知识在现实世界的局限性也直接决定了这种方法的局限性。

图 3 图像分割结果展示图

Unet[5]发表于 2015 年，初衷是为了解决生物医学图像方面的问题，后被广泛应用于卫星图像分割，工业瑕疵检测等，其采用 Encoder-Decoder 结构，即需要保证从 Encoder 中提取出的图像信息，能够用于在 Decoder 中还原出原始图像信息，确保前者提取出的图像特征能够包含所有重要信息。

而后的 Segment Anything[6]通过 Prompt+基础大模型的套路来解决目标分割的问题。经过实测，在大多数场景中 SAM 的表现都比较好。



(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**) (c) **Data**: data engine (top) & dataset (bottom)

图 4 SAM 论文中给出的图像分割任务、模型、数据示意图

2.4 图像增强

图像增强技术早期是处理图像的重要技术，主要包括去噪、增加清晰度等作用，近年来更多的实际上是为了丰富数据集的多样性，增强模型的鲁棒性而使用。在早期阶段，主

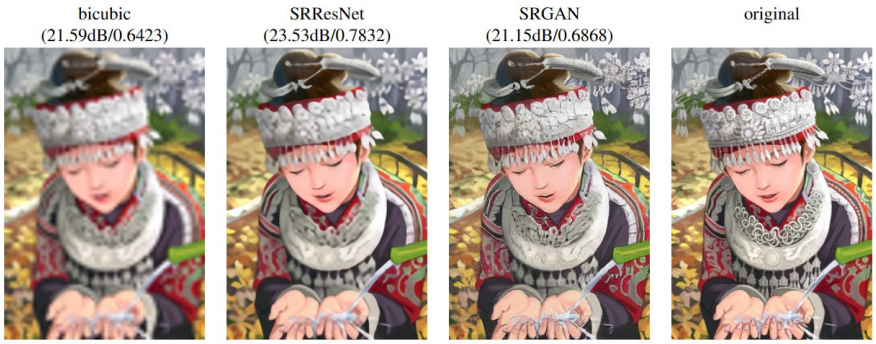


Figure 2: From left to right: bicubic interpolation, deep residual network optimized for MSE, deep residual generative adversarial network optimized for a loss more sensitive to human perception, original HR image. Corresponding PSNR and SSIM are shown in brackets. [4× upscaling]

图 5 SRGAN 论文中效果展示图

要通过人工构建简单的图像变换来丰富训练数据，如图像裁剪、旋转、平移、缩放、颜色抖动等基础变换，而后较深入的包括直方图均衡化、一系列的微分算子、灰度形态学操作、空间域与频域变换等等。

之后随着机器学习和深度学习的兴起，图像增强技术取得了显著的进展。卷积神经网络（CNN）等深度学习模型被广泛应用于图像增强任务，例如超分辨率、去噪等。生成式模型代表 GAN 作为一种强大的深度学习模型，由于可以在更高的层次上理解图像的语义信息，也被用于根据原始图像生成高质量的图像，从而提高增强效果。如下图 5 为 SRGAN[7] 对图像进行的超分辨率复原。

2.5 图像理解

从本质来说，图像理解要求模型能描述出图像中物体与物体之间的关系、图像中展现的情景等等具体信息，也就是说，图像理解要求模型能够识别图像中的事物主体、辨别事物主体所处的环境、理解事物主体之间的关系。要同时做到这几点，其关键就是提取图像中的高维语义特征和空间特征，而这无疑难于计算机视觉其他领域。

在上个世纪，图像理解主要依赖于手工设计的特征提取算法，如 SIFT（尺度不变特征变换）和 HOG（方向梯度直方图）。这些方法对于简单的图像识别任务效果不错，但由于这些方法提取出的图像信息过于浅层，并且在处理复杂场景和变化较大的图像时表现有限，因此应用十分有限。

图 6 左图为手工设计的特征提取算法结果示意，右图为 DALL·E 模型结果

随后出现的卷积神经网络，通过卷积核提取图像的细节特征，能够较好地提取出图像的高维特征。但是由于图像理解要求模型能够真正看懂一幅图像，是目标检测、图像分割等其他领域的集合，其难度明显高于其他领域，似乎直到最近几年，随着多模态任务的兴起，尤其是 image2text 与 text2image 任务的火热，使得研究者开始重新审视这个问题。早期如 Google 的 im2txt 模型，实现输入图像输出对应文本，但是效果并不好。最近 OpenAI 推出的 DALL·E 系列模型，则是一种基于语言的人工智能图像生成器，可以根据文本提示创建高质量的图像和艺术作品。

3 图像生成方向的调研

3.1 图像生成方向的背景

图像生成任务是指给定一段特定的输入（文字、图像、二者组合或者其他形式），输出一幅与输入相关（这种相关性由任务和模型决定）的图像。例如，在 text2image 任务中要求输入一段描述性文本，模型需要输出一幅符合文本要求的图像。图像生成任务的意义主要体现于两个方面：

- 有监督的深度学习需要依靠大量数据来训练网络，通常情况下，数据量越多，训练出来的网络具有更好的泛化性能。然而，收集数据需要耗费人力和财力，甚至有一部分特殊场景下的图像采集十分困难。因此，大部分视觉任务的训练数据目前依旧处于缺失或不足的状态。正是针对这样的问题，研究人员提出了基于无监督训练的图像生成方法。
- 图像生成的发展也带来了图像编辑、图像恢复、文本图像转化等有趣且重要的应用，使得人们可以根据自己的想法轻松地修改图像、生成特定图像，是计算机视觉落地的一个重要方向。

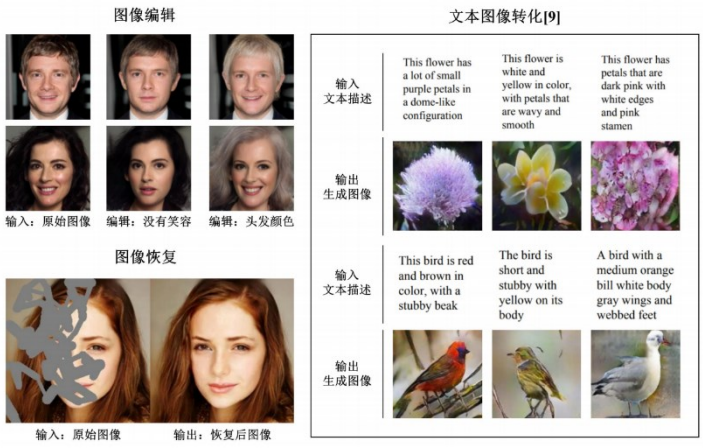


图 7 图像生成领域的多个研究方向

3.2 图像生成方向的子方向

正如上文提到图像生成任务是给定一段特定的输入，输出一幅与输入相关的图像，根

据输入形式、输入与输出关系的不同，图像生成又可以划分为多个子领域。

- **Text2Image**：从文本描述中生成图像，即输入一段关于图像内容的描述文本，模型根据其中的信息生成一幅包含所有重要信息的图像，这要求模型能够提取文本的重要信息，并将其编码输入到图像生成模块，而图像生成模块则要理解这段文本信息编码，并完成文本到像素的映射；
- **图像超分辨率**：输入输出形式为 Image2Image，其目的是将低分辨率图像转换为高分辨率图像，即在不改变原始图像的基本信息的情况下，恢复图像细节部分；
- **风格迁移**：输入输出形式也是 Image2Image，需要将一个图像的风格与另一个图像的内容相结合，生成具有前者风格的图像，例如输入一张现代照片，输出一幅以梵高风格作画的画作图像，即在不改变原始内容的主体结构的情况下，对其艺术形式进行转变；
- **交互式图像生成**：这一任务允许用户按照自己的意愿，通过与计算机系统进行实时的、双向的交互来生成图像。与传统的图像生成方法相比，交互式图像生成赋予用户更多的控制权和参与度，使其能够直接参与图像的创作过程。即在不改变原始图像风格、内容、主体框架的情况下，对图像内容进行细微的调整。

3.3 图像生成技术

图像生成技术可以粗略地分为传统方法与基于深度学习的方法，其分界点大致地可以定于 2012 年 AlexNet[2]将深度学习引燃的时刻。

3.3.1 基于规则的图像生成方法

使用预定义的规则和算法来生成图像，这些规则包括几何变换（如平移、旋转、缩放）、颜色调整、纹理重复等。实际上，这些规则就是出现在图像编辑软件中的图像修改工具，例如可以使用图层混合模式、蒙版、滤镜等功能来改变图像。基于规则的合成方法适用于简单的合成任务，可以在不需要大量数据和复杂模型的情况下进行图像转变（或者说生成）。

后来引入纹理映射技术和复杂的光照模型，将图像映射到 3D 模型的表面上，并模拟光的反射和折射，使得计算机生成的图像更具真实感。

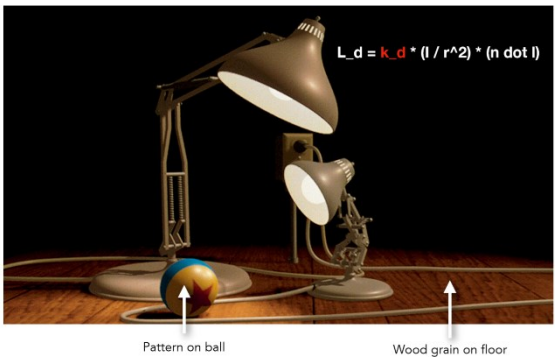


图 8 纹理映射技术和光照技术应用示意图

3.3.2 马尔可夫随机场

马尔可夫随机场是一种数学模型，描述的是在一系列状态中转移的概率过程。在图像生成领域，马尔可夫随机场通过建模像素与像素之间的关系，学习不同位置像素之间的依存关系，从已知的像素点按顺序预测下一个未知的像素点，从而生成新的图像。

这种马尔可夫随机场的图像生成方法通常需要选择适当的概率分布和能量函数，以确保生成的图像符合特定的要求，比如保持图像的平滑性、真实性和清晰度。同时，一些改进的方法也结合了深度学习技术，使得生成的图像更加逼真和多样，例如 Li 等人[1]就将 CNN 与马尔可夫随机场结合。马尔可夫随机场的使用不仅使得图像生成模型在静态图像生成上取得了显著的进展，也在动态场景生成、超分辨率图像合成等方面展现了强大的潜力。

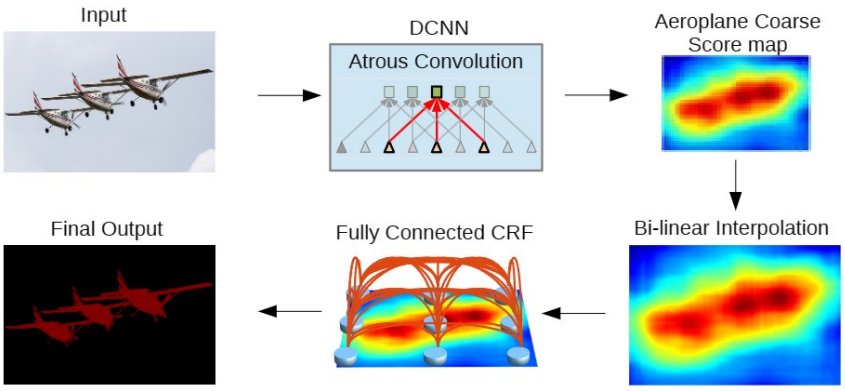


图 9 DeepLab v2 模型示意图

3.3.3 自回归模型

自回归模型是指模型后一时间步的输出依赖于过去时间步的观测值，这与马尔可夫随机场十分相似，但两者的建模方式及变量意义从本质上来说是不同的。自回归模型可形式化表示为：

$$X_t = c + \sum_{i=1}^{p} w_i X_{t-i} + \epsilon_t$$

X_t为时间步t的观测值，w_i为模型权重，c为模型偏置，ϵ_t为模型误差。将图像像素理解为拉直的序列，考虑每个像素都依赖于前面的像素，便可通过这种递归方法进行逐像素预测，即当前预测值可以作为下一个预测值的输入。

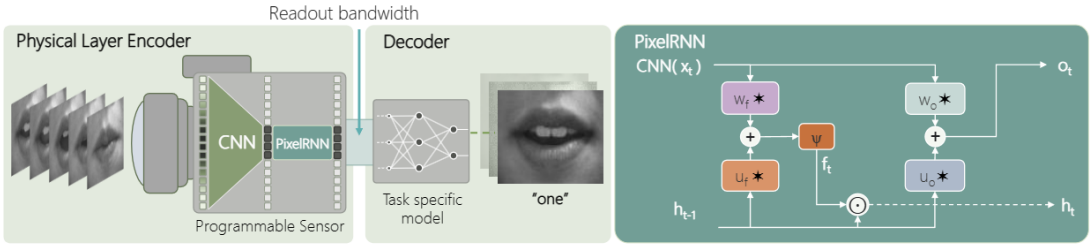


图 10 PixelRNN 模型示意图

这种自回归模型如 PixelRNN[8]（使用循环神经网络 RNN 来捕捉图像中像素的空间依

赖关系）和 PixelCNN[9]（通过卷积神经网络 CNN 捕捉像素之间的互相依存关系）都展示了很不错的实验效果，但是按照像素点去生成图像，且考虑到对数似然的最大化涉及到对整个图像的概率分布进行建模，导致计算成本高，在可并行性上受限，在处理大型数据如大型图像或视频时有一定困难。

3.3.4 自编码器

自编码器一般由编码器 Encoder 和解码器 Decoder 组成，在图像处理领域中，编码器负责将输入的图像编码为低维特征向量，而解码器则通过低维特征向量恢复出原始图像，这样的巧妙结构迫使编码器生成的图像特征向量尽可能多地包含图像信息，以便解码器能够最小化复原图像与原始图像的差异。

在图像生成领域，通常采用自编码器中的变分自编码器模型 VAE[10]。这种模型在 2014 年由 Diederik Kingma 和 Max Welling 提出，本质上就是在常规的自编码器的基础上，在 Encoder 的结果中加入了“高斯噪声”，使得后面的 Decoder 能够对噪声有鲁棒性；并且在损失函数中加入额外的 KL loss，目的是让 Encoder 的输出均值为 0，方差为 1，事实上就是相当于在 encoder 的损失函数中加入一个正则项，希望 Encoder 的输出具有零均值的特性。

总的说来，变分自编码器是在自编码器的基础上让图像编码的潜在向量服从高斯分布，从而实现图像的生成，优化了数据对数似然的下界，虽然 VAE 在图像生成上是可并行的，较自回归模型提升了速度、降低了计算复杂度，但是 VAE 存在着生成图像模糊的问题，我认为这是因为在 Encoder 中压缩数据维数，而这部分压缩的特征并不能表征出原图像的细粒度信息。

随后针对 VAE 的改进工作，如 IntroVAE[11]、NVAE[12]等，则通过引入 GAN、多尺度设计等方法对效果进行了提升，基本上解决了图像模糊的问题。

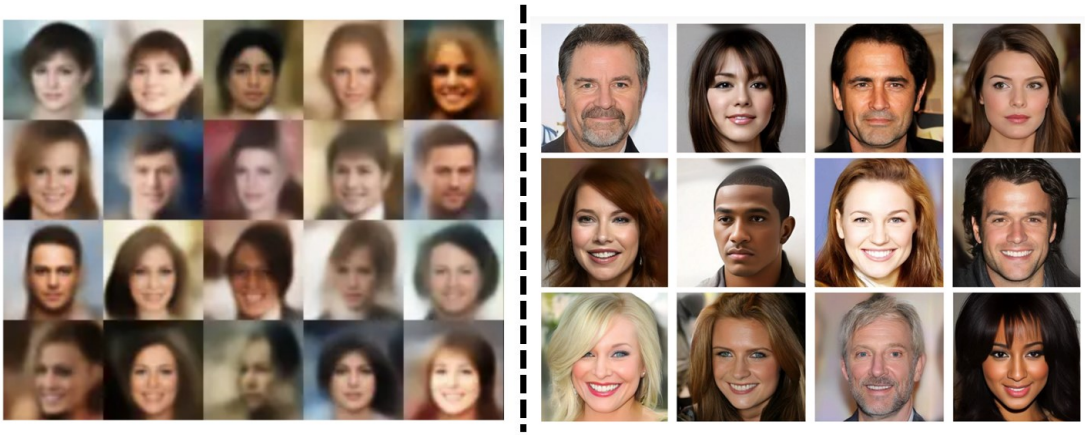


图 11 左图为 VAE 生成的较为模糊的图像，右图为 NVAE 生成的清晰图像

3.3.5 生成对抗网络

生成对抗网络（Generative adversarial network, GAN）自 2014 年由 Ian Goodfellow 等人提出后，就越来越受到学术界和工业界的重视。由于后文主要针对 DragGAN 一文进行分析，故在此详细阐述 GAN 系列模型的发展历程

从其基本思想来看，GAN 出发于博弈论中的零和博弈，将生成问题视作判别器和生成

器这两个网络的对抗和博弈：生成器从输入噪声中（一般是指均匀分布或者正态分布）输出一幅图像，判别器需要分辨出生成器的输出与真实图像。前者试图产生更接近真实的数据，相反，后者试图准确地分辨真实图像与生成器生成的虚假图像。由此，两个网络在对抗中进步，在进步后继续对抗，由生成网络得到的图像也随着辨别器性能的增强越来越完美，逼近真实数据，从而可以生成接近真实数据分布的图像。

早期的 GAN 虽然效果优于 VAE，但其本身具有两个较大的问题——一是如果出现判别器分辨能力很强的情况，生成器的梯度消失会较为严重，这样会导致在网络训练上很多时候生成器的参数基本上不会发生改变；二是由于网络是对抗式的训练范式，常常会造成训练时模型的崩溃（collapse mode，即无论输入什么，生成器都输出一张接近真实的图像，而判别器一直判断该图像为真实图像），需要精心设计网络。

WGAN[14]通过对生成样本和真实样本加噪声使得两个分布产生重叠，理论上可以解决训练不稳定；WGAN-GP 通过引入梯度惩罚，使得 GAN 训练更加稳定，收敛更快，同时能够生成更高质量的样本。后续产生了非常多的 GAN 变种，通过不同的技巧方法提升 GAN 生成图像的能力。

Mirza 等人在 2014 年提出 Conditional GAN (CGAN)[13]，在原有的 GAN 模型中加入了条件信息，使得生成器可以根据输入的条件信息生成对应的图像，并且在判别器中对图像是否真实、图像是否与条件信息匹配做出判别。

Radford 等人在 2015 年提出 DCGAN[15]，主要是在网络架构上改进了原始 GAN，利用 CNN 架构替换了原始 GAN 生成器与判别器的全连接网络，使得生成器可以生成更加逼真的图像。

Liu 等人在 2016 年提出了 CoGAN[16],该模型由两个生成器和两个判别器组成，每个生成器和判别器都对应一个域（可以是不同的图像集，例如黑白照片和彩色照片，或者艺术品和自然景观），可以生成高质量、多样性和语义一致性的图像。

Phillip 等人在 2017 年提出的 pix2pix[17]将成对的图像作为模型的输入，期望在在两个图像集之间学习得到潜在的关系。这个模型能够较好地实现图像转换任务，例如黑白图像到彩色图像、线条图像到真实图像。但是 pix2pix 模型需要大量的训练数据才能获得良好的性能，且训练数据的质量对模型的性能有很大影响，而成对图像的数据集非常少，标注成本高，使得该模型应用受限。

朱俊彦等人提出了 CycleGAN[18]，目的在于无需成对的数据集，对任意两个假设存在潜在关系的图像数据集的映射关系进行学习，解决了 pix2pix 等模型需要在成对数据集上训练的限制。

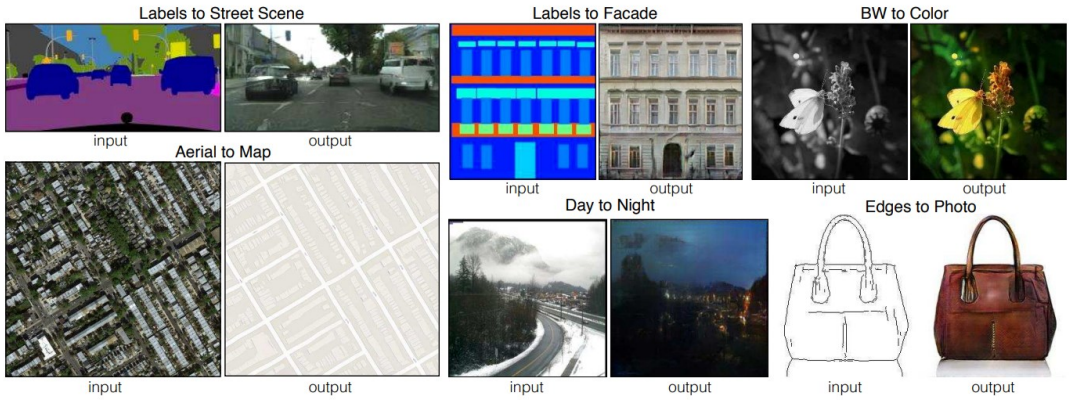


图 12 pix2pix **生成的图像**

3.3.6 扩散模型

之前提到的 VAE 的解码器（即生成器），是通过学习将标准高斯映射到数据样本，编码器采用的后验分布是自定义地将数据样本映射到标准高斯，而扩散模型正好相反，它采用基于马尔可夫链的前向过程将数据样本映射到标准高斯分布，再通过学习训练一个生成器，将标准高斯分布映射到数据样本，从而生成接近真实的图像。这样的训练方式和模型

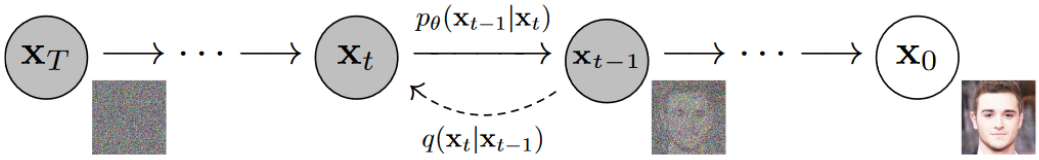


图 13 扩散模型前向及反向过程基本框架

架构使得扩散模型无论是图像生成效果，还是运行效率都高于 VAE。该模型在 2020 年被提出，后续 Stability AI、Google Brain 等公司相继基于扩散模型开发了自己的文生图、图像生成视频等模型，展现了该模型的有效性。

具体说来，给定一张原始图像X_0，前向阶段在X_0上逐步增加噪声，每一步得到的图像X_t只与上一步的结果X_{t-1}相关，直至第T步得到的图像X_T满足高斯分布，
而逆向阶段即解码器（或者生成器）阶段则是不断去除噪声的过程，直至最终生成一幅较为真实的图像。

图 14 Scalable Diffusion Models with Transformers 模型生成的图像

4 文献阅读与分析：《Drag Your GAN: Interactive Point-based Manipulation on the Generative Image Manifold》

这篇论文采用对抗生成网络实现图像生成，具体说来，作者在这篇论文中提出了一个 GAN 的变体——DragGAN[19]，基于 StyleGAN2，加入运动监督和点跟踪两个模块，实现了通过鼠标取定锚点即可自动生成新图像的强大功能。采用这个模型，要想按照自己的想法编辑一幅图片，用户可以先设置一个起始点，一个目标点，和想要改变的区域（即设置 mask），然后模型就会迭代地执行运动监督和点跟踪这两个步骤，前者会控制起始点向目标点运动，后者则是确保起始点与视图修改的位置始终匹配，在图像更改过程中不产生偏移。

4.1 研究动机

在这篇论文提出之前，图像生成领域已经有大量的基于 GAN 的模型、基于扩散模型的图像生成方法，但对图像生成大都缺乏直观的、精准的控制，为了解决这一问题，DragGAN 需要面临以下困难：

- 灵活性：模型需要能够控制多种空间属性特征，包括位置、姿势、形状、表情、物体的空间布局；
- 精确性：模型需要能够精确地按照操作者的意愿，控制图像的调整和生成；
- 泛化性：模型需要能够适应多种不同的物体类别，而不仅仅限于特定的物体种类；

4.2 基础模型 StyleGAN

实际上，DragGAN 只是提出了一种生成更好的隐变量的方式，其根据输入生成图像的核心部件是 2021 年提出的 StyleGAN2[21]，而 StyleGAN2 是 StyleGAN[20]的部分修改版，故这里简要介绍 StyleGAN。

StyleGAN 中的“Style”一词不是风格转换中的图像风格，而是指数据集中人脸的主要属性，比如人物的姿态、人脸的风格，包括了脸型上面的表情、人脸朝向、发型等等，还包括纹理细节上的人脸肤色、人脸光照等等图像中主体的特征。

StyleGAN 的网络结构包含两个部分，第一个是 Mapping network，第二个是 Synthesis network，前者将隐藏变量z转变为中间隐藏变量w，以此作为整幅图像的风格特征并用于控制生成图像的风格与之大致相同；后者的作用是获取三个输入——中间隐藏变量w、上一层的输出、服从一定分布的随机噪声，通过这三个输入生成最终的图像。形式化语言如下：

从隐空间Z中获取一个隐变量z，并通过一个 8 层的感知机作为映射函数f，获得中间隐变量$w \in W$：

$$f: Z \to W$$

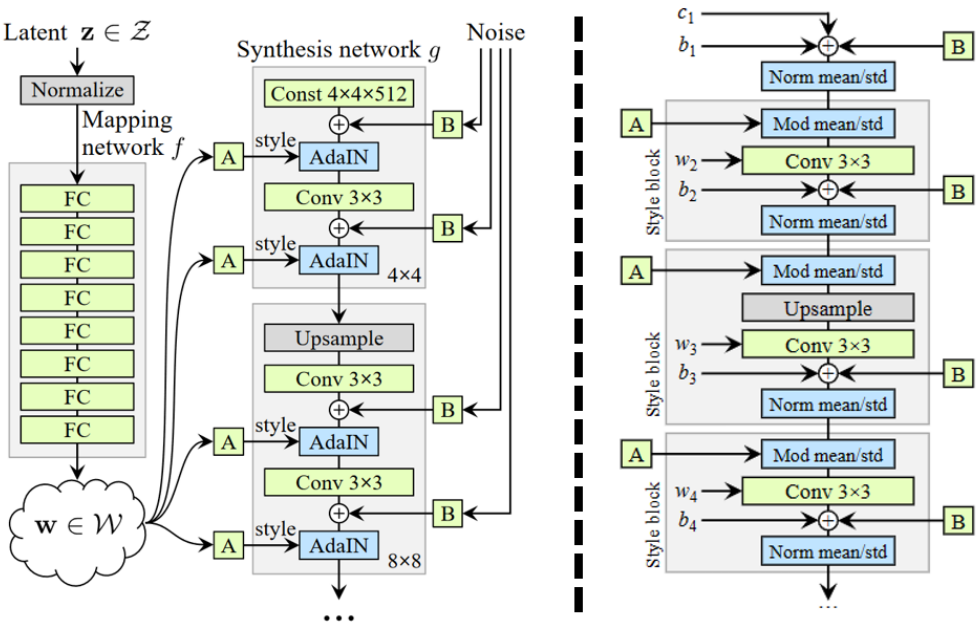


图 15 左图为 StyleGAN 模型结构示意图，右图为改进后的 StyleGAN2 模型示意图

通过仿射变换（Affine Transformation）将中间隐变量w转换为代表风格的变量$y = (y_s, y_b)$，利用图片风格迁移算法 AdaIN 将该风格y融入到一个混合了噪声的常量中，并通过卷积层，重复多次后，完成该层操作，将最终结果输入到下一层：

$$AdaIN(x_i, y) = y_{s,i}\frac{x_i - \mu(x_i)}{\sigma(x_i)} + y_{b,i}$$

在这样的多层这样的变换融合操作中，每一层将融合了图像特征的中间隐变量w通过图片风格迁移算法加入卷积层，使得模型能够逐步将图像的风格特征融入到生成的图像中。另外，额外增加的服从一定特征分布的噪声，增加了模型生成的图像的细节多样性、局部特征随机性。

4.3 基于交互式点的操作的分析

在之前提到 DragGAN 是在 StyleGAN2 的基础上，对隐变量w（实际上就是控制图像风格）进行细致地修改，使得图像生成能够按照用户的意愿进行。而用户的意愿是通过交互点的形式输入到模型中的，模型驱动控制点指向的物体部分向目标点方向移动，例如在嘴角处设置控制点p，在嘴角上方设置目标点q，则最终期望控制点指向的嘴角向目标点移动，而在一般情况（即一般分布中，由基础模型 StyleGAN2 的训练数据分布决定）下，嘴角上移到那个位置，大概率人物主体是处于笑的状态的，因此最终生成的图像中能够达到使照片中的人微笑的效果。

图 16 交互式点操作示意图（源自原论文）

4.4 运动监督

在 DragGAN 提出的基于交互点的操作中，需要控制设定的控制点不断向设定的目标点移动，这就要求需要一种监督手段，对偏离运动方向的生成像素进行相应的惩罚，以迫使模型在每一次迭代生成中，控制点附近的像素点都相对于之前的位置更靠近目标点。

为了实现这一想法，原作者在论文中提出了一个不依赖于任何附加的神经网络的运动监督损失。其关键思想或者说先验知识是，生成器产生的中间特征具有区分性，一个简单的损失就足以监督运动。具体说来，原作者考虑了 StyleGAN2 的第 6 块之后（即处于 StyleGAN2 中间）的特征映射f，在分辨性和区分性之间有很好的权衡，即处于整体结构和细节处理的平衡位置，因此根据该映射得到的隐变量进行运动监督损失计算，并在此处通过双线性插值调整f的大小以与最终图像拥有相同的分辨率。

运动监督损失的思想主要是迫使生成图像中目标点的像素值与原图像中控制点的像素值相似，同时为了提供遮挡区域（即 mask，尽量保证指定区域不发生大幅度变化），在该损失中加入重构损失，迫使模型生成的图像中无关区域与原始图像类似，其损失函数如下：

$$\mathcal{L}=\sum_{i=0}^{n}\sum_{q_i\in\Omega_1(p_i,r_1)}\|f(q_i)-f(q_i+d_i)\|_1+\lambda\|(f-f_0)\cdot(1-M)\|_1$$

其中p_i是控制点，t_i是目标点，$\Omega_1(p_i,r_1)$是指与控制点p_i距离不超过r_1的区域，q_i为区域中的任意一点，$f(q_i)$表示像素q_i处经过特征映射f后的特征值，$d_i=\frac{t_i-p_i}{\|t_i-p_i\|_2}$是从$p_i$指向$t_i$的归一化向量，$f_0$是原图像中的特征映射，$M$为操作者指定的掩码遮罩。$\|f(q_i)-f(q_i+d_i)\|_1$项要求控制点周围的像素与目标点附近对应点的相似度尽量高，差异尽量小，也就是驱动控制点向目标点移动。$\|(f-f_0)\cdot(1-M)\|_1$项则是要求遮罩M处外的区域中的像素与原像素值保持尽量一致，差异尽量小，也就是迫使模型不修改遮罩外的区域。

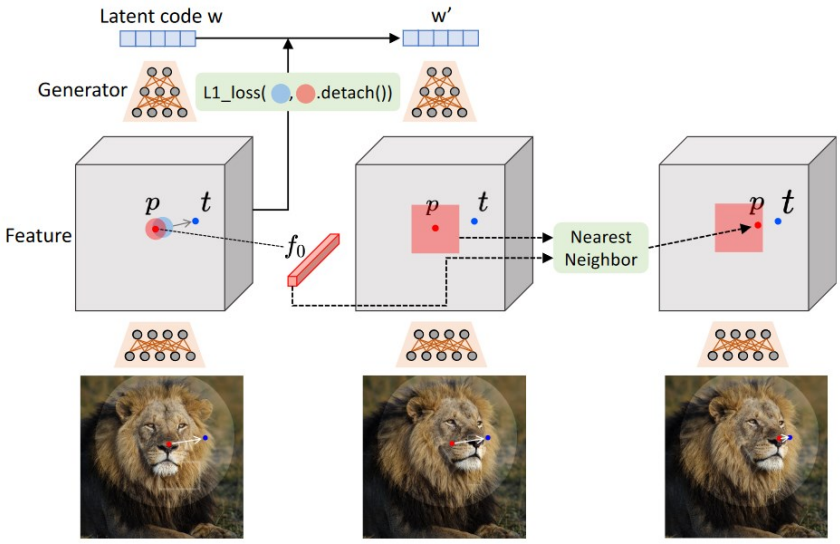


图 17 运动监督及点跟踪实现示意图（图自原论文）

通过这样一个精巧设计的损失函数，逐步学习修改输入 StyleGAN2 中的隐变量，使得最终输出的图像随着迭代次数的增多，不断接近期望的效果。在后续的实验中，原作者观察到图像的空间属性主要受前 6 层的影响，而剩下的网络层只影响外观。因此，受 Karras 等人“style-mixing”技术[20]的启发，StyleGAN 网络结构只更新前 6 层的，同时维持其他层

参数不变以保持最终生成图像的外观。

4.5 点跟踪

在上一节提到的运动监督步骤中，模型通过学习更新得到了一个新的隐变量w'，新的映射函数f'，以及一幅新生成的图像I'，而这副图像I'与原始图像I肯定不同，同时由于运动监督的作用，控制点指向的图像中的主体部分会向目标点偏移，而偏移的距离由于神经网络学习的特殊性，是无法得知的，此时控制点仍在原始图像对应的位置（例如嘴角已经向上偏移，此时未更新的控制点将在嘴角下方，指向错误的主体位置），如果在下一次迭代过程中，依然使用该未更新的控制点，便会导致运动监督损失函数计算产生较大误差，进而生成的图像与预期不符，更为严重的是，这个误差会在不断迭代过程中累积，最终导致生成结果不稳定甚至不可控的情况。

在生成新图像之后，为了及时准确地更新控制点位置，原作者采用了点跟踪的方法，即通过在未更新的控制点周围进行最近邻搜索，找出与原始位置最相似的图块，得到新控制点的位置，形式化表述为：

$$p_i = \underset{q_i \in \Omega_2(p_i, r_2)}{\operatorname{argmin}} \| f'(q_i) - f_i \|_1$$

即找到新图像中与原始图像控制点附近区域最相似的区域，并将其中心点（实际是按点遍历的）作为新的控制点。

4.6 模型整合及分析

将上述四个板块整合在一起，就得到 DragGAN 的整个模型框架，即输入一幅图像，控制者指定一个或者多个成对的控制点和目标点，经过多层 StyleGAN2 生成器，并在其中加入运动监督调整隐变量，使用点跟踪实时更新控制点，最终到达精准编辑图像的目的并得到期望生成的图像。

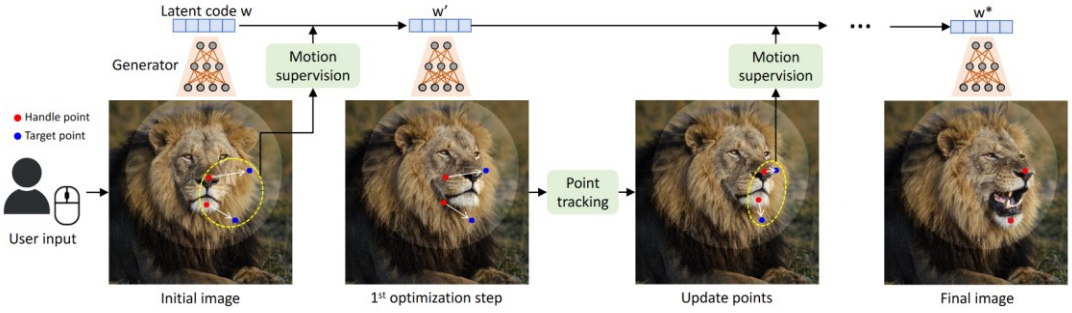


图 18 DragGAN 模型示意图（图自原论文）

4.6.1 巧妙的遮罩实现

在我以往接触的图像处理及自然语言处理的过程中，希望模型不对某一个部分进行操作或者期望模型对某一部分是未知的，就会使用遮罩（mask）掩盖这一部分，而这样的遮罩通常是全零矩阵或者全零向量，这使得模型从中无法获得有效信息。但是在 GAN 这样的

图像生成模型中，图像作为输入，全连接层的存在使得每一个部分都会影响最终的输出，如果对某一部分施加全零矩阵遮罩，即图像区域变黑，必然会对最终生成的图像造成较大的影响。如何加入遮罩使得模型在生成图像的过程中保持某一个区域不变呢。

该论文中利用了在损失函数中加入重构损失项$\|(f-f_0)\cdot(1-M)\|_1$，模型在追求损失最小化的情况下必然将保持遮罩之外的部分尽量不变，从而达到了设计目的。这样的“软”遮罩相比以往的“硬”遮罩给生成模型留下了更多自由发挥的空间，使得模型能够调节图像真实性与维持图像局部区域不变之间的平衡，类似于线性判别分析 LDA 中软间隔的实现与效果。

4.6.2 向图像生成模型中融入先验知识

DragGAN 将 StyleGAN2 作为基础模型或者说图像生成器，一方面，StyleGAN2 自身的结构特性保证了图像生成过程中能够维持主体特征，另一方面，StyleGAN2 作为一个预训练模型，在训练过程中通常针对同一类物体进行学习，这保证了模型学习到了该类物体的大部分显著特征，而这使得模型在处理输入图像时能够引入这些特征，类似于先验知识，例如输入图像中主体是一只狮子，那么由于 StyleGAN2 之前预训练时见过其他情景下的狮子，那么它可能就会知道狮子闭着的嘴里有犬牙（当然这是通过像素数据分布得出的，而不意味着模型真的能够思考），从而在生成的图像中，出现原本图像中不存在的犬牙。

当然，这样的特性可能会为模型带来一定的麻烦，比如当修改幅度较大时，模型可能会将图像主体完全替换掉，来使得当前图像中的像素值的组合在分布中为最大概率。例如下图中试图修改汽车的高度、车轮的大小、车斗的大小，最终生成的图像中的汽车的品牌与原始图像中的汽车品牌完全不同，当然这也是模型生成多样性的体现之一。

图 19 图像修改后图像主体不匹配的示意

4.6.3 模型局限性

尽管 StyleGAN2 为模型带来了相应的物体先验知识，能够生成超出当前图像中的物体，更加接近真实图像，但是同时也为模型带来了局限性，即由于模型并没有真正理解图像的信息，模型的生成能力取决于 StyleGAN2 预训练数据集的分布及训练效果，对于超出数据集分布的情景，模型依然难以应对，例如下图展现的对于不常见的人的姿势，模型生成的姿势就显得较为奇怪。

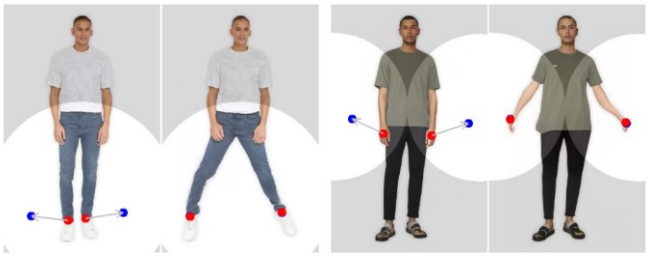


图 20 超出原始数据分布时生成图像的示意

另外，点追踪使用的最近邻匹配算法在某些特殊情况下可能失效，例如缺失纹理、表面光滑的情景，这时，对于控制点的一个特定邻域，像素点的值差别较小，导致当控制点所指代的位置的像素迁移时，处于控制点原有位置的区域像素点变化较小，模型便无法仅依赖匹配算法更新控制点的位置，导致模型失效。

图 21 物体表面无明显纹理时匹配算法失效的示意，左图为纹理缺失，右图纹理正常

5 学习图像处理与分析课程的收获

这门课程中，朱旗老师采用深入浅出的方式介绍复杂的处理方法，并提供大量图像处理的范例和效果展示，加深了我对处理方法的理解，结合课程实验，培养了我对该领域的学习兴趣和动手实践能力，为今后从事该领域相关研究或应用开发打下了坚实的基础。

5.1 知识学习

图像，作为现实生活中最为常见的数据形式之一，承载着丰富而复杂的信息。它不仅仅是视觉感知的载体，更是一种具有潜在知识和巨大价值的数据形态。图像处理与分析课程作为人工智能专业的专业基础课程，为我提供了处理与分析图像的关键知识和理论基础。

在这门课程中，我主要学习了以下重要知识：

- 图像处理的基本概念，包括连通性、距离测度、灰度级的定义、RGB 与 HIS 等颜色模型、灰度与彩色图像的数据形式；
- 图像处理的基本运算，例如灰度直方图统计、点运算、阈值变换、对数变换、傅里叶变换以及多种滤波算子与常用卷积核等；
- 图像处理的经典方法，例如直方图均衡化、直方图匹配、空间和频域滤波、形态学处理（腐蚀和膨胀，开运算和闭运算，边界和骨架等），也了解并使用多种滤波器；
- 图像分割，了解了阈值与图像分割，边界提取与轮廓跟踪，Hough 变换，区域增长和分裂，基于聚类的图像分割等
- 图像的表示和描述，图像压缩等等

这些知识帮助我构建了关于图像处理与分析的知识体系，这将帮助我在之后处理与分析图像的过程中更加得心应手，更好地理解图像中的信息和模式，根据具体任务的需求选

择合适的图像预处理方法，从而提升后续任务的处理效果。

5.2 科研帮助

在课程中，我还了解到图像智能分析应用方向，包括视频目标跟踪与分析，车牌识别，人脸、掌纹识别，以及图像处理与分析在脑科学、医学影像处理等前沿领域的应用。这拓展了我的学术视野，加深了我对计算机视觉领域的喜爱，为我以后的学习与科研方向提供了有效的指导。

这样系统的基础知识学习，为我日常科研提供了有力支撑。例如，传统卷积核的使用为了理解卷积神经网络提供了极大的帮助，多样的传统图像增强方法为我进行图像对比学习提供了巨大支持，多种距离测度的定义也让我理解对比损失函数的具体含义及变种的实际意义。

6 参考文献

[1] Li C, Wand M. Combining markov random fields and convolutional neural networks for image synthesis[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2479-2486.

[2] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25.

[3] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[4] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[5] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015: 234-241.

[6] Kirillov A, Mintun E, Ravi N, et al. Segment anything[J]. arXiv preprint arXiv:2304.02643, 2023.

[7] Ledig C, Theis L, Huszár F, et al. Photo-realistic single image super-resolution using a generative adversarial network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4681-4690.

[8] Van Den Oord A, Kalchbrenner N, Kavukcuoglu K. Pixel recurrent neural networks[C]//International conference on machine learning. PMLR, 2016: 1747-1756.

[9] Shi W, Caballero J, Huszár F, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 1874-1883.

[10] Kingma D P, Welling M. Auto-encoding variational bayes[J]. arXiv preprint

arXiv:1312.6114, 2013.
[11] Huang H, He R, Sun Z, et al. Introvae: Introspective variational autoencoders for photographic image synthesis[J]. Advances in neural information processing systems, 2018, 31.
[12] Vahdat A, Kautz J. NVAE: A deep hierarchical variational autoencoder[J]. Advances in neural information processing systems, 2020, 33: 19667-19679.
[13] Mirza M, Osindero S. Conditional generative adversarial nets[J]. arXiv preprint arXiv:1411.1784, 2014.
[14] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks[C]//International conference on machine learning. PMLR, 2017: 214-223.
[15] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv preprint arXiv:1511.06434, 2015.
[16] Liu M Y, Tuzel O. Coupled generative adversarial networks[J]. Advances in neural information processing systems, 2016, 29.
[17] Isola P, Zhu J Y, Zhou T, et al. Image-to-image translation with conditional adversarial networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1125-1134.
[18] Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2223-2232.
[19] Pan X, Tewari A, Leimkühler T, et al. Drag your gan: Interactive point-based manipulation on the generative image manifold[C]//ACM SIGGRAPH 2023 Conference Proceedings. 2023: 1-11.
[20] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 4401-4410.
[21] Karras T, Laine S, Aittala M, et al. Analyzing and improving the image quality of stylegan[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 8110-8119.

附录

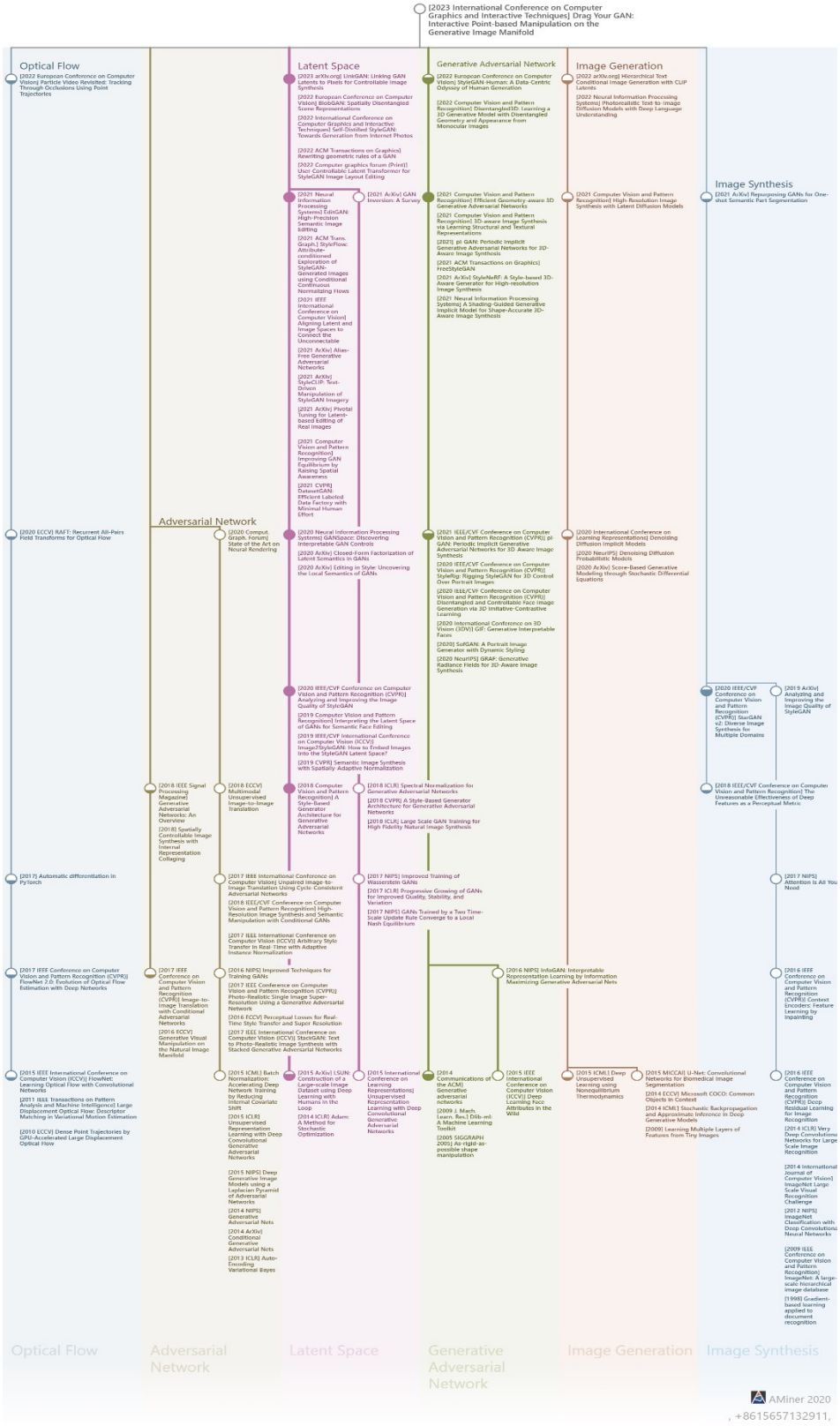


图 22 AMiner 中生成的 DragGAN 论文溯源树