本科毕业设计（论文）

基于生成对抗网络的草图到图像翻译

学院（部、中心）：　电信学部

专　　业：　自动化

班　　级：　自动化 001 班

学生姓名：　王贤超

学　　号：　2203211416

指导教师：　杨旸

2024 年 06 月

摘 要

随着社会发展，当今人们在各个领域内对于图像的需求不断增加，手绘、拍照等传统的图像生成方式逐渐无法满足日益增长的需要。对此，生成对抗网络技术出现所带来的生成式模型，为人们提供了一种使用计算机来合成图片的优质解法。在这些技术中，通过手绘草图来生成真实图像是近年来非常热门的研究方向，前人在这一领域内已经取得了大量的成果，提出了多套高效的草图图像翻译模型。但是针对多类别的草图图像翻译，由于需要大量的数据作为训练的支撑，目前依然存在数据集中手绘草图数据不足的问题。一旦出现草图缺失的现象，生成模型的性能往往就会出现较大幅度的下降。

针对这一问题，本文使用了一种基于开放域的多类别草图图像翻译模型。该模型在开放域下对草图域和图像域的数据进行联合学习，从而使模型在两个域下都能有很高的性能，因而可以利用合成草图来代替真实草图进行学习，解决了草图数据缺失时图像生成质量不佳的问题。模型中还加入了用以判断图像类别的分类器，能够较好地对不同类别图像的颜色和纹理进行学习。此外，该模型通过引入随机混合抽样算法，避免了生成过程中出现的图像模糊问题，进一步提高了生成图像的质量。本文对模型中各组件的具体实现也进行了介绍。

本文对使用的模型开展了相关实验进行验证。首先，使用生成图像的相关评价指标，通过将本模型的生成结果与之前的模型进行对比，体现了本文模型在图像生成质量上的提升。另外，本文通过设置消融实验，减少模型中的关键组件，证明了本模型使用的相关算法和框架对于提升模型具有重要作用。最后，本文通过服务器端的网页设计，以该模型为核心，设计了一套可在线进行草图图像翻译的网页，将模型从理论层面的搭建迁移至实际应用。

关 键 词：生成对抗网络；草图图像翻译；开放域；网站搭建

ABSTRACT

With the development of society, today's people have increasing demand for images in various fields. Hand drawing, photography and other traditional ways of generating images are gradually unable to meet the growing needs. In this regard, the generative model brought by the emergence of Generative Adversarial Network technology provides a high-quality solution for people to synthesize images using computers. Among these techniques, generating real images from hand-drawn sketches is a very popular research direction in recent years, and previous researchers have achieved a lot in this field, proposing several efficient sketch-to-image translation networks. However, for multi-class sketch-to-image translation, there still exists the problem of insufficient sketch data in the dataset due to the requirement for a large amount of data to support training. Once some sketches are missing, the performance of the generative model tends to drop substantially.

To address the problem, this paper uses an open-domain multi-class sketch-image translation model. The model jointly learns data from both sketch and image domains, so that the model can have high performance under the two domains. As a result, synthetic sketches can be utilized to replace real sketches in learning, which solves the problem of low-quality generated images when sketch data is missing. The model also adds a classifier for determining image categories, which can help learn the colors and textures of images of different categories. In addition, the model avoids the problem of blurry images during the generation process by introducing a random sampling algorithm, which further improves the quality of the generated images. The specific implementation of each component in the model is also presented in this paper.

Relevant experiments are carried out to validate the model. First, by comparing the generated results with previous model utilizing relevant metrics, the model in this paper proves to be productive in image generation. In addition, by setting up ablation experiments, namely reducing the key components in the model, the paper demonstrates that the algorithms and frameworks used in this model play an important role in improving the performance. Finally, the paper applies the model to practical use, designing web pages that enable users to generate images from sketches online.

KEY WORDS: Generative adversarial network; Sketch-to-image translation; Open domain; Web development

目 录

1 绪论

1.1 研究背景和意义

当今社会数字化程度不断提高，图像在各个领域的应用越来越广泛，生成需求日益增加。传统的图像生成方式（如手工绘制或拍摄等）往往成本高昂且效率低下，无法满足人们的发展需求。针对这一问题，机器学习技术的发展为我们提供了使用计算机生成图片的解决方案。

在这些技术中，生成对抗网络作为一种强大的深度学习模型，可以高效地完成图像生成的工作。生成对抗网络由一个生成器与一个判别器组成：判别器需要尽可能将生成网络的输出与真实样本进行区分，而生成器的输出结果则要尽可能地骗过判别网络。在两个网络相互对抗的过程中，模型权重等参数不断更新迭代，最终使得模型生成的图片能够以假乱真。

自诞生以来，生成对抗网络的应用边界不断得到扩展。其中，草图到图像的翻译网络，作为生成对抗网络应用的重要分支和图像生成领域的重要问题之一，是近年来备受关注的研究方向。草图图像翻译，是指从手绘风格的、相对抽象的草图，生成带有草图特征且具有真实感的现实图像[1]。在艺术创作、产品设计、医学诊断等多个领域，将手绘草图转化为逼真的实物图像，都有着极为重要的作用。以设计领域为例，草图图像翻译网络可以根据设计师初步构思的草图方案，高效直观地生成与草图风格相近的真实作品，为设计师提供优质的参考。不难看出，该技术有着非常广阔的应用背景，可以为社会发展和人类生活带来诸多便利，也能反过来推动计算机视觉和人工智能等相关技术的进步。

1.2 研究方法与研究现状

1.2.1 常见草图图像翻译模型

近年来，在草图图像翻译这一领域，国内外的研究人员已经提出了多种模型，取得了非常大的进展。针对不同的应用背景和需求，根据对生成图像的控制程度，大体可以将现有的研究工作分为两类：第一类是基于条件生成对抗网络开展的无控制的草图图像翻译，第二类是通过属性和风格来控制输出图像的精细控制的草图翻译[2]。

1） 无控制的草图图像翻译

在此类工作中，目前相当多的方法是基于带有配对或未配对数据的数据集，通过条件生成对抗网络等技术，完成两个图像域之间的转化。在翻译网络中，通过将草图和图像一起输入生成器，使得生成器能够学习草图和图像之间的映射关系。对于现有的模型，还可以根据训练方式进一步分为监督学习和无监督学习：监督学习方法需要带有配对数据的数据集，将配对数据进行一对一的映射；无监督的草图图像翻译方法

在训练时可以使用未配对的数据集，数据经由一组生成对抗网络，从源域映射到目标域再返回源域，完成一次学习。

2） 精细控制的草图图像翻译

此类工作通过研究图像的视觉属性（语义分割图、边缘图等）和风格（颜色、纹理、笔画等），在生成图像时可以带有更加精细的控制信息，从而更好地得到目标图像。一种比较常用的思想，是将图像拆分为局部组件，对每一个局部设计模块进行特征提取和训练，最终再将其进行融合，从而得到全局图像。目前这类工作的研究成果，主要是能够针对人的脸部和发型生成质量较高的图像，但其算法很难推广到其它任务下的草图图像翻译。

1.2.2 研究现状

2014 年，由 Goodfellow I 等人提出的生成对抗网络横空出世[3]。它的出现，使得草图到图像翻译成为了近年来深度学习领域非常热门的研究方向，至今许多学者在这方面已经做出了巨大贡献。

2017 年，Isola P 等人基于条件生成对抗网络，提出了 pix2pix 方法，通过在网络中加入用户提供的图片来确保生成图像和输入图像的匹配性[4]。针对 pix2pix 方法在生成复杂图像时质量下降的问题，Wang TC 等人又提出了改进版本 pix2pixHD。pix2pixHD 引入了多尺度的生成器和判别器以及新的损失函数，来生成更高分辨率和更加逼真的图像。此外，作者还引入了语义控制机制，使得生成的图像在保持逼真的同时能够进行语义上的精确控制[5]。同年，Zhu JY 等人首次提出了 CycleGAN。CycleGAN 引入了循环一致损失，通过同时训练一对生成器和判别器，在两个待转换的域之间建立了一种映射关系[6]。这种设计实现了图像的跨域转换，如马变斑马、照片转换成艺术家风格的作品等。之后，Yi ZL 等人对此模型进行了优化，提出了 DualGAN 方法，通过引入双重生成器和判别器，实现了更加稳定和高质量的图像转换[7]。2018 年，Choi Y 等人提出了 StarGAN 模型，旨在解决多个域之间的图像转换任务。相比之前的方法，StarGAN 采用了一个统一的生成器和判别器，通过引入新的条件约束机制，使模型能在一个统一的框架下实现多个领域之间的图像翻译[8]。为了解决该模型中多个域间进行图像转换时出现的生成质量下降问题，以前者为基础的 StarGAN v2 随后诞生。通过采用单个生成器和判别器、引入跨域对齐和多尺度生成等技术，StarGAN v2 可以实现更加多样化和高质量的图像转换[9]。2019 年，Kim J 等人提出了 UGATIT 模型，以解决无监督图像转换任务中的样式迁移和域适应问题。UGATIT 引入了自适应层、实例归一化和生成器中的注意力机制，以实现更加精准逼真的图像转换[10]。而后，Tang H 等人提出了 AttentionGAN，一种同样用于无监督图像转换的生成对抗网络模型。AttentionGAN 在 UGATIT 的基础上引入了注意力机制，通过关注图像的局部区域，提高了模型对图像细节的把握和表达能力[11]。这一创新使 AttentionGAN 能够更好地实现不同域之间的图像转换，为无监督图像翻译任务带来了新的进展。

针对多类别的草图到图像翻译，Chen WL 等人通过引入屏蔽剩余单元块（MRU）

和多样性损失，训练了以类别标签为条件的编码器-解码器模型，可以生成 50 种不同类别的图像[1]。同年，Lu YY 等人为了解决跨域生成的问题，提出了 ContextualGAN，在图像生成过程中将草图视为一种上下文弱约束，学习草图和图像的联合分布[12]。2019 年，Ghosh A 等人使用一种基于门控的方法来调节类，可以在只有单一生成器网络的情况下，根据用户绘制的草图交互生成对应的图像[13]。Gao CY 等人提出的 SkecthyCOCO 则用生成对抗网络网络把边缘图和对应的图片编码到同一个隐空间中，并通过两个生成器来生成前景和背景[14]。

多类别的图像翻译需要大量的数据作为支撑。尽管上述模型在生成多类别的图像方面取得了一定进展，但是它们对草图数据给出的扩充方法，主要还是利用与实物图贴近、保留大量信息的边缘图。这导致这些模型在面对笔画稀疏、风格抽象的手绘草图时效果不佳。对于多类别图像跨域生成时出现的问题，本文使用了一种开放域下的草图图像多类别翻译模型，并将模型嵌入了架设在服务器上的网页中，制作了一个可视化的在线草图图像翻译工具。

1.3 本文的主要工作

本文针对多类别图像生成，以现有工作的思想方法为基础，设计了一套可在线访问的图像生成框架。本文具体的研究内容主要包括以下几点：

（1）通过文献调研，学习了草图图像翻译的相关原理和技术，实现了基于经典图像生成网络的算法

（2）针对多类别图像生成的常见问题，在现有框架的基础上进行了调整优化，并给出了框架的具体实现

（3）将本文设计的模型和以往模型的性能进行对比，使用主流的评价指标开展实验和测试

（4）利用本文设计的模型，在服务器上搭建了一套网络框架，可供用户在线访问并由草图生成图片

本文设计的模型在生成图像的质量得到提升的同时，通过对模型的打包封装，极大地提高了用户在使用时的易用性。

1.4 本文的组织架构

本文主要对多类别的草图图像翻译进行了相关学习和研究，重点想通过开放域和特定算法提升草图缺失时模型生成图像的性能。为了提高模型的实用程度和易用性，本文对模型的接口进行了修改并搭建了相关的网络框架。论文各章节的内容安排如下：

第 1 章：绪论。本章阐述了本文的研究背景和意义，当前学界的研究方法和研究现状，也分析了多类别草图图像翻译中存在的问题。

第 2 章：草图图像多类别翻译基本原理。通过回顾草图图像多类别翻译中涉及到的基本原理和经典算法，对相关技术进行了学习和总结。

第 3 章：基于开放域的草图图像多类别翻译。本章通过应用第二章所介绍的方法和思想实现了草图到图像的多类别翻译，并介绍了模型在实际使用时内部的具体组件。

第 4 章：草图翻译模型的实验开展和实际应用。通过相关指标对第三章中翻译模型的性能进行了对比和分析，证明了模型的良好性能。通过设置消融实验，验证了模型中各组件的作用。此外，在前文基础上实现了基于模型的可视化服务器网站的搭建。在服务器上搭建的网络框架中内置有本文的模型，以供用户在线访问并可视化地生成结果。

第 5 章：总结与展望。主要对全文的研究进行了总结，并对文中可供后续研究和改进的地方进行了说明和展望。

2　草图图像多类别翻译基本原理

本章重点介绍了基于生成对抗网络的草图图像多类别翻译的基本原理。其中，主要技术为生成对抗网络，涉及到 pix2pix 和 CycleGAN 两个经典模型的原理。另外，对后文模型中使用的分类器也进行了相关介绍。

2.1 生成对抗网络

2.1.1 生成器和判别器

生成器的输入通常是一个来自正态或均匀分布的随机噪声，在经过一系列变换后，转换为与真实数据分布相似的样本。它的目标是欺骗判别器，使其无法区分生成的假样本和真实样本。为了最大化判别器面对生成结果时判定为真的概率，常常通过极大似然估计或交叉熵来设计生成器的损失函数。

判别器可以被看作一个二分类器，对输入的真实数据和生成的数据，输出一个标量，表示输入数据为真的概率。它的目标是最大化对真实数据判别为真的概率，同时最小化对生成数据判别为假的概率，其损失函数常用二分类交叉熵损失函数来实现。

图 2-1 展示了生成对抗网络的结构。

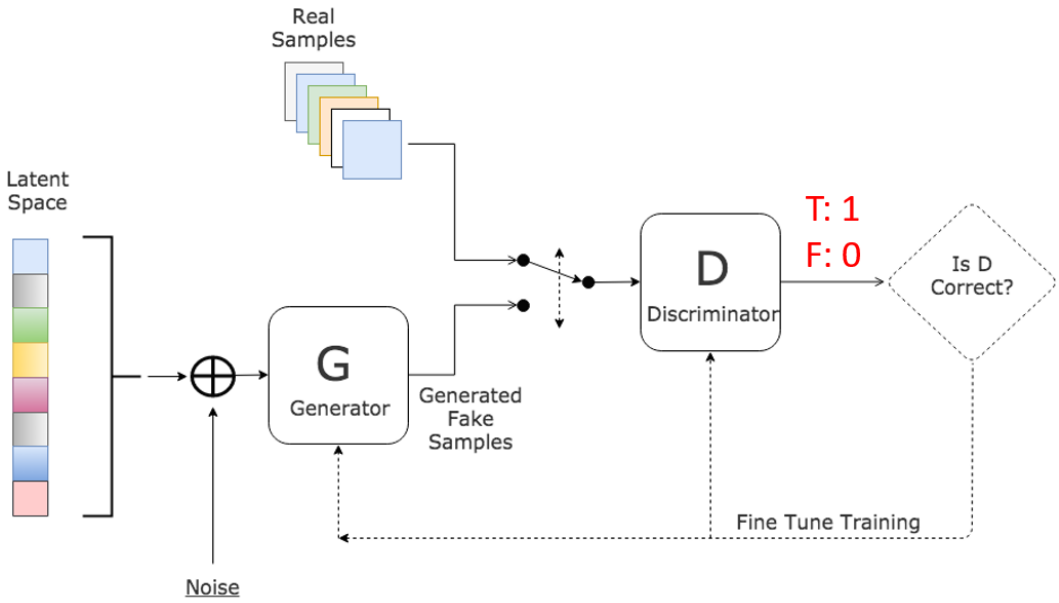

图 2-1　生成对抗网络结构

2.1.2 全连接层和卷积层

全连接层是生成对抗网络中最基本的层之一，它的作用是将输入的特征进行线性变换。在生成器中，全连接层会将生成器接收的低维随机噪声映射到一个高维空间，以供后续的其他层进行处理；而在判别器中，经过多个卷积层提取特征后，结果通常会通过一个或多个全连接层映射为一个标量输出，得到判别概率。

卷积层会通过卷积核在输入数据上的滑动，产生一组特征图，实现局部特征的提取。卷积核的尺寸比输入图像小得多，会通过指定的移动方向遍历整张图像。在卷积层中，卷积核会与大小、维数相同的输入图像块逐元素相乘并求和，将求和结果作为该位置上的输出。卷积核在滑动时每次移动的距离称为步幅，步幅会直接影响后续输出特征图的尺寸。

判别器中的卷积层通常用来提取输入图像的特征，减少数据的空间尺寸，增加其深度特征；而生成器通常使用反卷积层来生成图像，通过逐步增加数据的空间尺寸，实现从低维向量到高维图像的转换。

2.1.3 激活函数和损失函数

激活函数是在网络中引入非线性的关键组件。生成器在中间的网络层常用 ReLU 激活函数来保持梯度流动并引入非线性，而在输出时通过 tanh 函数，将输出值范围限制在[-1, 1]，从而与图像像素值归一化后的范围保持一致；判别器在输出时会使用 Sigmoid 函数，将输出值映射到[0, 1]范围，表示输入数据为真的概率。公式（2-1）、（2-2）和（2-3）中分别给出了 ReLU、tanh 和 Sigmoid 三种激活函数的具体表示：

$$f(x)=\begin{cases}0 & x<0\\ x & x\geq 0\end{cases} \tag{2-1}$$

$$f(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{2-2}$$

$$f(x)=\frac{1}{1+e^{-x}} \tag{2-3}$$

损失函数是网络中指示模型优化方向的重要组件，影响模型的性能与收敛速度。它度量了模型中单个样本的预测值与实际值之间的差异，大多数情况下采用距离公式或直接做差进行计算。在训练过程中，模型的目标是通过不断迭代，调整模型内的参数，使损失函数最小化，从而提高预测精度。公式（2-4）和（2-5）分别给出了常用的均方误差损失函数和交叉熵损失函数在数学上的表达式：

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(Y_i-Y_i)^2 \tag{2-4}$$

$$H(p,q)=-\sum_{x\in X}p(x)\log q(x) \tag{2-5}$$

与损失函数密切相关的还有代价函数和目标函数。代价函数与损失函数形式一致，只是范围定义在整个训练集上，可以用所有样本的误差求和再取平均来计算。而目标函数是最终模型需要优化的函数，一般包括经验风险和结构风险，前者用于衡量模型的损失，与代价函数相关，后者则用于衡量模型的泛化能力，可以通过正则化的处理来实现。引入正则化处理的原因在于，如果只从经验风险的角度考虑，模型可能会出现过拟合的问题，即对特定历史数据的学习过于精准，以致在面对干扰时预测效果反而出现下降。图 2-2 很好地说明了这一问题：图中的绿线代表过拟合模型，黑线代表

正则化模型。虽然对于训练数据，绿线能近乎完美地拟合，但由于拟合得太过贴近，与黑线相比，在新的测试数据上反而会有更高的错误率。为了降低模型对外界输入和内部参数的敏感性，通常使用的正则化方法有 L1 和 L2 两种，其表达式分别由公式（2-6）和（2-7）给出：

$$w^* = \arg\min_w \sum_{i=1}^{N} (w^T x_i - y_i)^2 + \lambda \|w\|_1 \tag{2-6}$$

$$w^* = \arg\min_w \sum_{i=1}^{N} (w^T x_i - y_i)^2 + \lambda \|w\|_2^2 \tag{2-7}$$

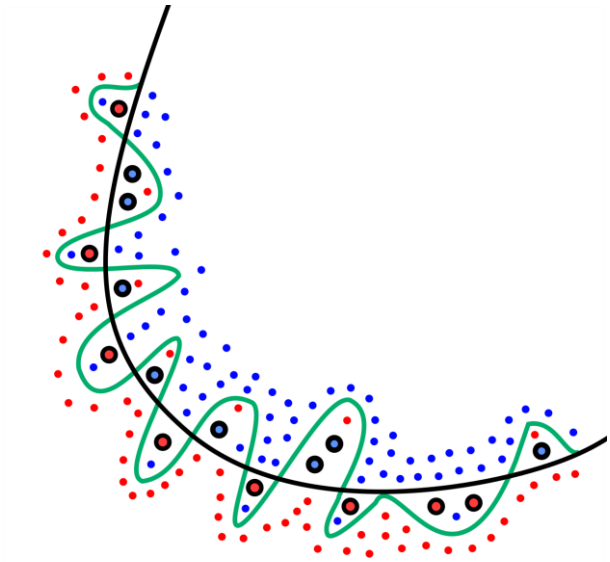

图 2-2　过拟合问题示意图

2.1.4 训练过程

生成对抗网络的最终目标，是找到生成器与判别器之间的纳什均衡点。其实现方法是通过二者在训练过程中的充分对抗，提升系统性能。在训练过程中，训练一者的同时，需要另一者的参数保持固定不变。生成对抗网络涉及的极小化极大算法如公式（2-8）所示：

$$\begin{aligned} \min_G \max_D V(D,G) = & E_{x\sim P_{data}(x)}[\log D(x)] \\ & + E_{z\sim P_z(z)}[\log(1-D(G(z)))] \end{aligned} \tag{2-8}$$

上式中，等式右边的第一项代表输入图像服从原始数据分布时判别器的输出，第二项代表输入图像服从噪声分布时与判别器输出有关的函数，该函数的值随着判别器判别概率的升高而下降。对于此公式，生成对抗网络的做法是：在生成器 G 固定时，上式的值与两项都有关系，要尽可能地使其最大化，对应提高判别器对原始图片的判别概率而降低对生成图片的判别概率；在判别器 D 固定时，上式中第一项的值相应固定，不会影响整个表达式的结果，故要针对第二项进行最小化，对应生成器要尽可能地欺骗判别器。

2.2 经典草图图像翻译算法

2.2.1 pix2pix

pix2pix 是一种基于条件生成对抗网络的生成式模型，它通过输入一对图像进行有条件的图像生成。它可以被用在草图图像翻译的场景，即输入草图和图像组成的数据对，输出拟合的图像。

1） 核心思想

pix2pix 的核心思想是在对抗训练中引入条件变量，用以学习输入和目标间的映射，生成器会根据输入图像生成与目标图像相匹配的输出图像。此方法通过逐像素比较，可以让输出在结构及细节上更加贴近目标图像，使得生成的图像不仅具有真实感，还符合特定的转换要求，比如生成与输入草图线条相符的实物图。该模型的整体框架如图 2-3 所示。

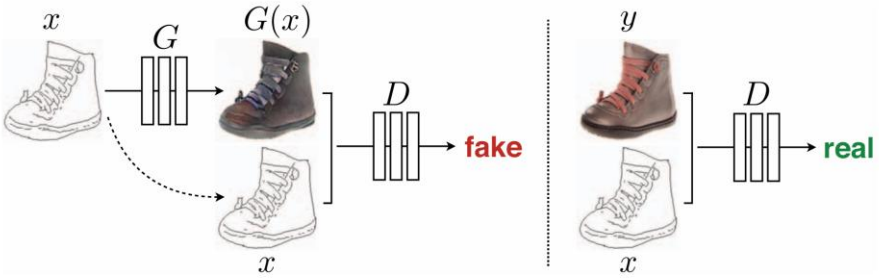

图 2-3 pix2pix 模型框架[4]

2） 模型架构

pix2pix 的模型主要由一组生成器和判别器构成。

为了有效地处理图像的细节和全局特征，生成器采用 U-Net 的编码器-解码器结构，如图 2-4 所示。

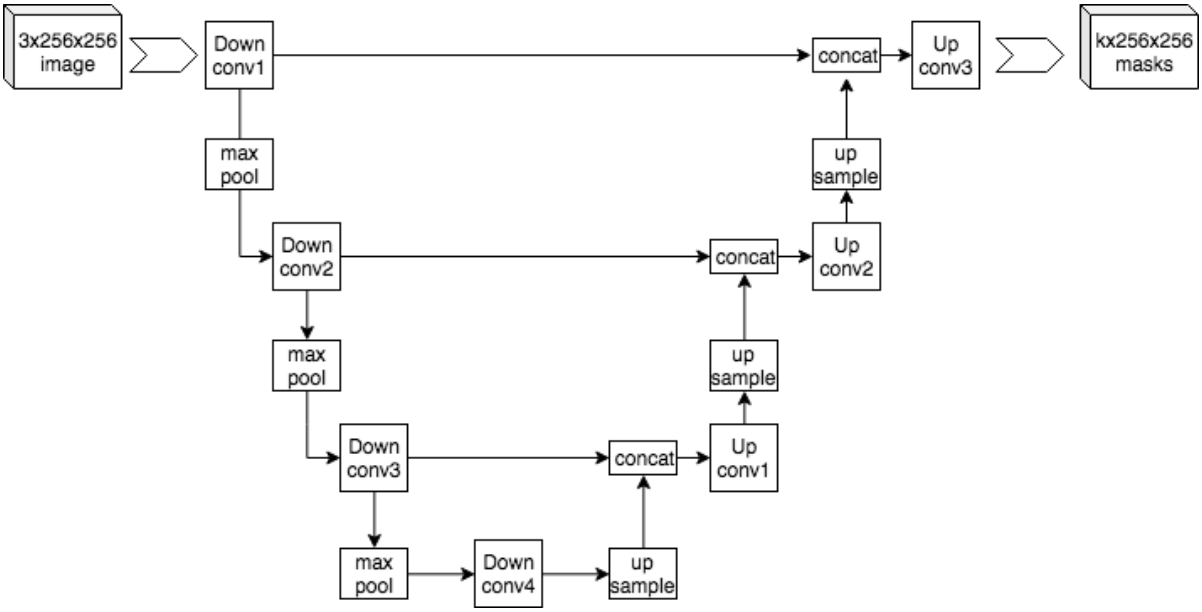

图 2-4 U-Net 的编码器-解码器结构

编码器对输入逐层下采样，以提取图像特征；编码器的每一层特征图通过跳跃连接直接传递到相应的解码器层，以保留更多细节信息；解码器逐层上采样，将编码器提取的特征重建为目标图像。通常来说，编码器由一系列卷积层、批量归一化层和一个 ReLU 激活函数组成，而解码器的结构则是将编码器中的卷积层改为反卷积层。

判别器采用 PatchGAN 结构[15]，通过将图像划分成小块，对每个小块单独进行判别，然后综合这些判断结果输出最终的判别概率。这样设计的好处在于，不仅减小了模型的复杂度，还增强了对局部细节的关注。判别器内置了一系列卷积层和一个 ReLU 激活函数来对每个小块进行真伪判别，输出一个二维矩阵，表示图片的真伪判别结果。在输出矩阵的末尾，通过一个全连接层，来综合每个小块的判别结果。

3） 损失函数

pix2pix 的联合损失函数由对抗损失和重建损失组成。对抗损失源自标准的生成对抗网络，用于提升图像的真实感，如公式（2-9）所示：

$$\begin{aligned} L_{GAN}(G,D) &= E_y[\log D(y)] \\ &+E_{x,z}[\log(1-D(G(x,z)))] \end{aligned} \tag{2-9}$$

重建损失是 pix2pix 新引入的机制，可以确保生成图像与目标图像的一致性，如公式（2-10）所示：

$$L_{L1}(G) = E_{x,y,z}[\|y - G(x,z)\|_1] \tag{2-10}$$

最终的目标函数由公式（2-11）给出：

$$G^* = \arg\min_G \max_D L_{cGAN}(G,D) + \lambda L_{L1}(G) \tag{2-11}$$

4） 运行结果

对于 pix2pix 模型，本文使用了包含鞋子和手提包类别的配对草图图像的数据集，训练模型由输入草图生成图像，结果如图 2-5 所示。

图 2-5　pix2pix 对鞋子和手提包类别的草图生成结果

2.2.2 CycleGAN

CycleGAN 是一种无监督学习方法，可以在不需要成对训练样本的情况下，实现图像转换和图像风格迁移，草图图像翻译也是其应用领域之一。

1） 核心思想

CycleGAN 的核心思想是通过一组生成器和判别器，同时学习图像从源域到目标域的映射和从目标域到源域的映射，实现两个域之间的相互转换，如图 2-6 所示。在 CycleGAN 中，两个判别器分别在两个域中，对真实图像和生成图像进行区分，两个生成器则在两个域之间形成两条链路——一条将源域 X 的图像转换为目标域 Y 的图像再转回源域 X，另一条将源域为 Y 的图像转换为目标域 X 的图像再转回源域 Y。为了提升转换后图像重构成原始图像的效果，CycleGAN 引入了循环一致性损失，确保 $F(G(x)) \approx x$ 和 $G(F(y)) \approx y$。这样做的目的在于，模型对两个域的特征都能够进行学习，故能在两个域中都能生成风格相近的图片。

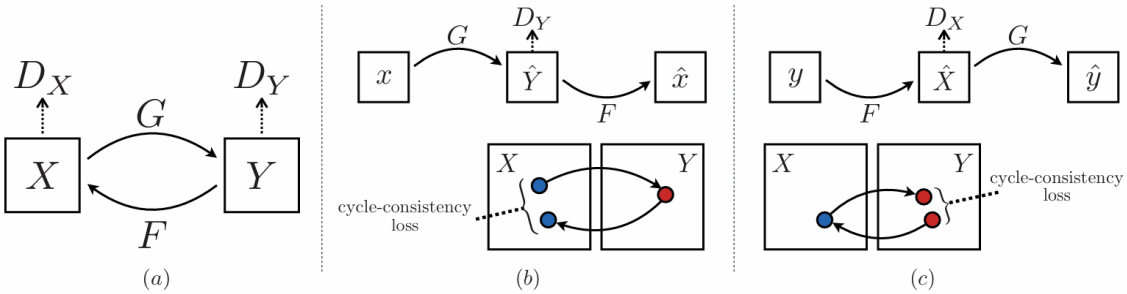

图 2-6 CycleGAN 模型框架[6]

2） 模型架构

CycleGAN 的模型主要由两组生成器和判别器组成。

生成器可以采用前面提到的 U-Net 结构，也可以使用残差神经网络，即 ResNet 结构。ResNet 中引入了图 2-7 所示的残差块。

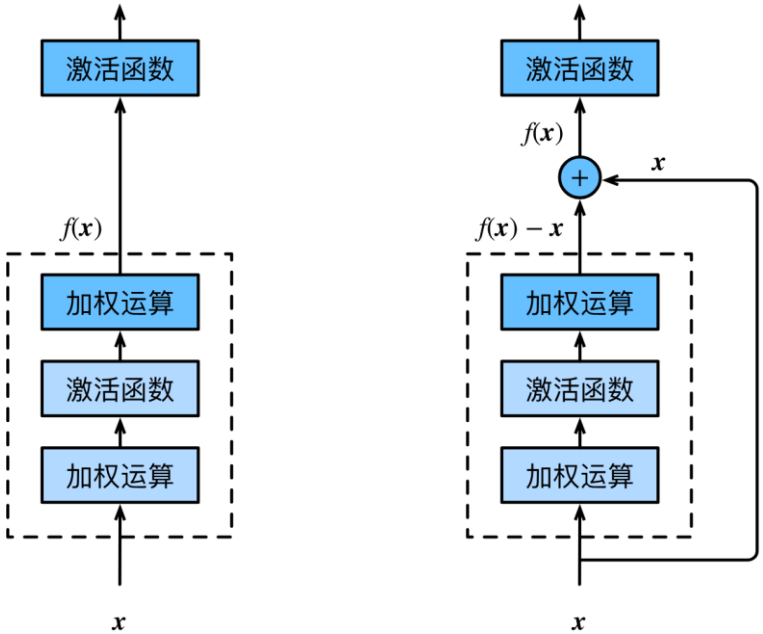

图 2-7 网络结构对比图[16]。左图虚线框的部分需要直接拟合出理想映射 $f(x)$，右图引入残差块后虚线框的部分需要拟合出残差映射 $f(x)-x$

Resnet 结构的核心思想在于网络的每一层不直接去学习预期的输出，而是学习与输入之间的残差关系。这种网络的输出通过添加跳跃连接，即跳过中间的某些网络层来进行恒等映射，再与网络层的输出相加合并。对于一个输入为 x、包含若干层的网络，希望学出的理想映射为 $f(x)$，作为图（2-7）上方激活函数的输入。图中左边的网络由于没有 ResNet 结构，虚线框的部分需要直接对该映射 $f(x)$进行学习拟合；而右边的网络在引入残差块后，网络的参数被重新设定，虚线框拟合的目标变为残差函数 $f(x)-x$，最终与输入相加合并得到输出 $f(x)$。这样做的好处在于，残差映射在实际中更容易优化，并且输入可经由跨层的数据线路，更加快速地向前传播。

判别器采用的结构与之前介绍的 pix2pix 模型中所用的一致，两个判别器均采用了 PatchGAN 模型。

3） 损失函数

CycleGAN 的损失函数包括对抗损失、循环一致性损失和身份损失。生成器 G 和判别器 D_Y 的对抗损失由公式（2-12）给出：

$$\begin{aligned} L_{GAN}(G,D_Y,X,Y) = {} & E_{y\sim p_{data}(y)}[\log D_Y(y)] + \\ & E_{x\sim p_{data}(x)}[\log(1-D_Y(G(x)))] \end{aligned} \tag{2-12}$$

生成器 F 和判别器 D_X 的对抗损失与公式（2-12）形式相同，只需替换相应符号即可。确保输入图像经过两个生成器转换后，能够恢复到原始图像的循环一致性损失如公式（2-13）所示：

$$\begin{aligned} L_{cyc}(G,F) = {} & E_{x\sim p_{data}(x)}[\|F(G(x))-x\|_1] \\ & + E_{y\sim p_{data}(y)}[\|G(F(y))-y\|_1] \end{aligned} \tag{2-13}$$

整体的目标函数由公式（2-14）给出：

$$\begin{aligned} L(G,F,D_X,D_Y) = {} & L_{GAN}(G,D_Y,X,Y) + \\ & L_{GAN}(F,D_X,Y,X) + \lambda L_{cyc}(G,F) \end{aligned} \tag{2-14}$$

其中，λ 的作用是控制对抗损失和循环一致性损失的权重。模型最终需要处理的函数由公式（2-15）给出：

$$G^*,F^* = \arg\min_{G,F}\max_{D_X,D_Y} L(G,F,D_X,D_Y) \tag{2-15}$$

对于风格迁移等场景，有时还需要额外引入身份映射损失，否则生成器很容易自主地修改图像的色调，使得整体的颜色发生改变。身份映射损失的具体计算如公式（2-16）所示：

$$\begin{aligned} L_{identity}(G,F) = {} & E_{y\sim p_{data}(y)}[\|G(y)-y\|_1] \\ & + E_{x\sim p_{data}(x)}[\|F(x)-x\|_1] \end{aligned} \tag{2-16}$$

4） 运行结果

对于 CycleGAN 模型，本文使用了含有菠萝类别的配对草图图像的数据集，训练模型由输入草图生成图像，结果如图 2-8 所示。

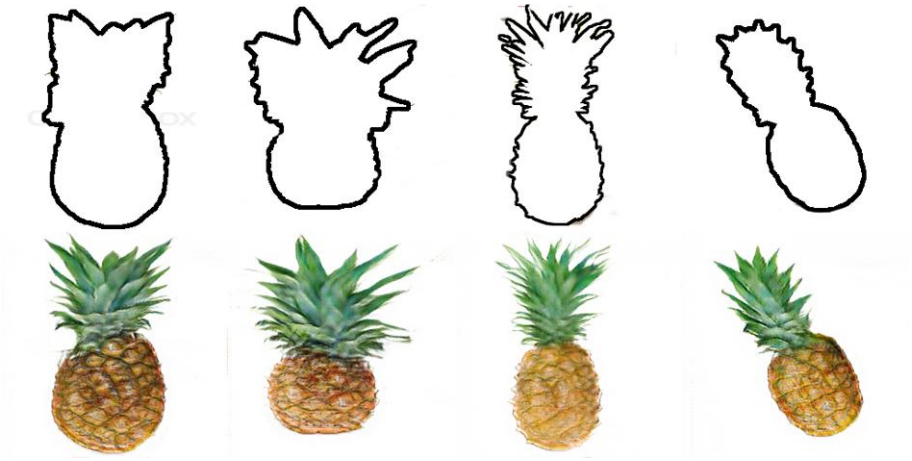

图 2-8　CycleGAN 对菠萝类别的草图生成结果

2.3 分类器

对于多类别的草图到图像翻译模型，对图像进行识别分类也是训练过程中的一环，因此分类器是整个模型的重要组成部分。下面对后续模型将要使用的两种高性能的图像分类算法进行介绍。

2.3.1 HRNet 架构

HRNet 是 Wang JD 等人在 2019 年提出的一种神经网络架构，被广泛应用于人体姿态估计、语义分割和图像分类等任务[5]。

1） 核心思想和模型架构

HRNet 的核心思想是在整个模型中都保持特征是高分辨率的，从而获得更多的特征信息，以提高模型的精准程度。在传统网络架构中，图像的特征往往在卷积层时从高分辨率转为低分辨率，之后再转回高分辨率，这会损失一部分信息。针对这一问题，HRNet 采用了逐步增加不同分辨率的特征并使之融合的方法。

图 2-9 给出了一种通过 HRNet 实现高分辨率网络的例子。

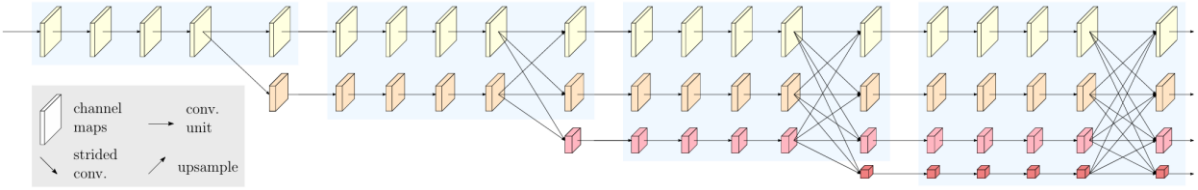

图 2-9　以 HRNet 为基础的高分辨率网络结构[5]

一般来说 HRNet 从一个高分辨率的卷积层开始，使用 3x3 的卷积核来提取初始特征。之后，HRNet 会使用不同大小的卷积核，从高分辨率开始，逐步增加低分辨率的分支。不同的分支之间存在定期的特征交换，以保持不同分辨率之间的信息流动。比如，高分辨率的细节特征能够传递到低分辨率的分支、而低分辨率的全局特征也能传递回高分辨率的分支。在网络的每个阶段结束时，HRNet 会将所有分支的特征进行融合。融合后的特征表示既包含高分辨率的细节信息，也包含低分辨率的全局信息。这

样的网络架构输出的特征表示是多种分辨率特征的综合体，会包含丰富的信息内容。

2） 在图像分类中的应用

设计之初，HRNet 被用于人体姿态估计，但其高效的多分辨率特征表示能力，使其在图像分类问题上同样表现优秀：既包含全局信息又包含细节信息的特征表示，能够更准确地描述图像内容，提高分类的精度；同时，通过优化上下采样操作和特征融合策略，HRNet 的整体计算量和参数个数也相对较低，具有较高的训练和预测效率。研究表明，HRNet 在 ImageNet、CIFAR-10 和 CIFAR-100 等多个标准数据集上，都取得了很高的分类准确性[5]。

2.3.2 EfficientNetV2 架构

EfficientNetV2 是 Tan MX 等人在 2021 年提出的一种高效且灵活的卷积神经网络架构，适用于图像分类等任务[17]。

1） 核心方法和模型架构

EfficientNetV2 改进了 EfficientNet 中提出的复合缩放方法[18]，使其在深度、宽度和分辨率上都能进行更加精细的调整，适用于不同尺寸的输入。此外，针对低分辨率的图像输入，该架构在标准的 MBConv 层的基础上引入了 Fused-MBConv 层，如图 2-10 所示。通过将标准卷积和逐点卷积分离进行，在低分辨率输入下减少了模型的计算量，改善了其特征提取能力。

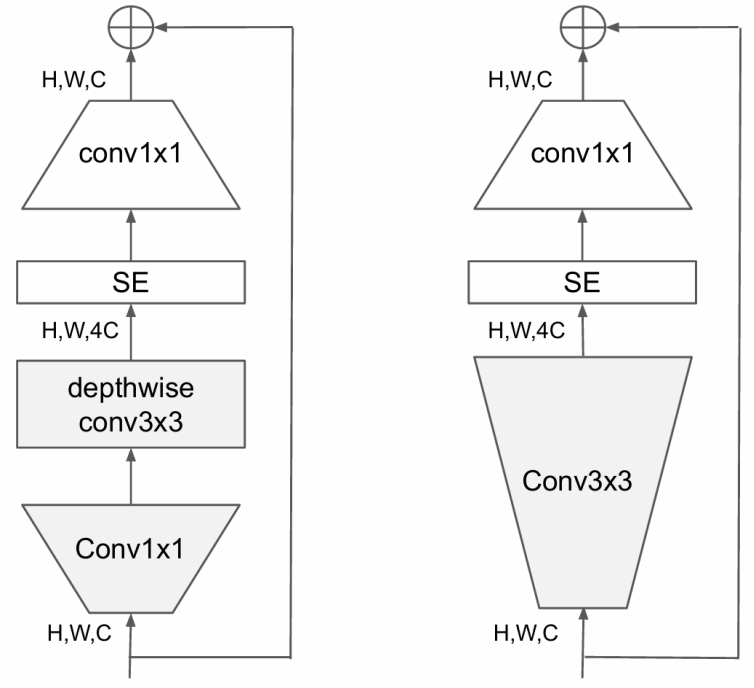

图 2-10　MBConv 层（左）和 Fused-MBConv 层（右）的结构[17]

值得一提的是，EfficientNetV2 引入了一种新的渐进式学习率调度策略，这对于模型性能的提升有着极为重要的作用。该策略开始时使用较低的分辨率和较高的学习率，然后逐步增加图像的分辨率并降低学习率，从而加速训练过程。同时，为了解决速度提升过程中识别精度下降的问题，该架构还使用了适应性正则化机制：通过随机裁剪、

水平翻转、色彩抖动等多种数据增强技术，来增加数据多样性，提高模型的泛化能力；通过 L2 正则化、Dropout 和数据增强结合的方式，有效地防止过拟合，提高模型的稳定性。

2） 性能表现

通过渐进式训练，EfficientNetV2 在 ImageNet 和 CIFAR/Cars/Flowers 等多个数据集上都具有很好的性能。在分类精度上，以 ImageNet 数据集为例，EfficientNetV2 的 top-1 准确率为 82.1%，比 EfficientNet-B0 提高了 2.3%。在训练速度上，EfficientNetV2 相比之前的 EfficientNetV1 提高了约 2 倍[17]。

2.4 本章小结

本章的内容主要分析讨论了生成对抗网络的相关技术和实现，介绍和总结了草图图像翻译中涉及的常用技术和原理。本章对 pix2pix 和 CycleGAN 两种经典的草图图像翻译模型进行了重点介绍，说明了其在实现过程中的重要思想方法，并给出了这两种模型的测试运行结果。最后介绍了 HRNet 和 EfficientNetV2 两种高性能的分类器。这些技术都会在后文的模型中得到应用。

3　基于开放域的多类别草图图像翻译

本章主要介绍了基于开放域的多类别草图图像翻译，运用第二章介绍的相关理论和思想，在开放域下训练模型，并加入了随机混合抽样算法，解决了多类别草图到图像翻译过程中存在的问题。对于模型实现过程中用到的具体技术，本章给出了详细的说明。

3.1 多类别草图图像翻译中存在的问题

多类别的草图图像翻译需要大量的训练数据作为支撑，但现有的数据集大部分只有图片域的数据，而缺少草图域的数据。因此根据现有条件，如何在草图数据缺失的情况下依然完成图像的学习和生成，成为了翻译模型需要解决的问题。

如果直接使用边缘图来代替草图进行训练，得到的模型[19]在输入为草图时表现不佳，其结果由图 3-1 给出。

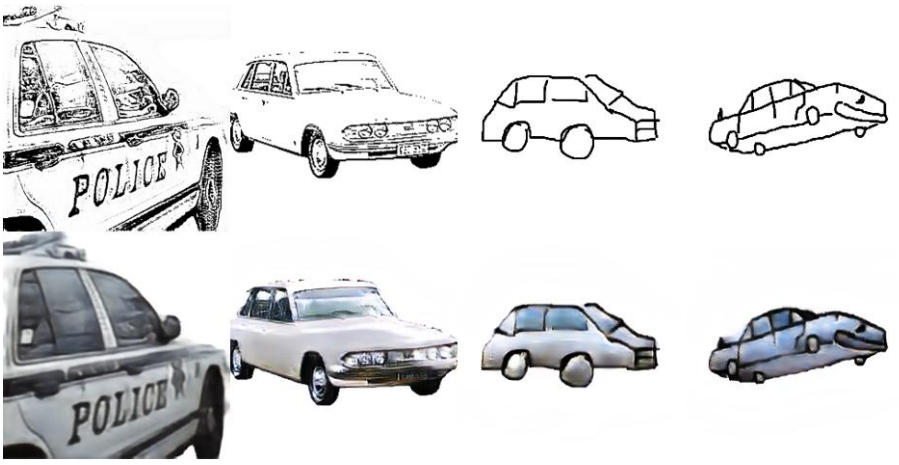

图 3-1　利用边缘图代替真实草图的模型生成结果

可以看到，输入为左半边的边缘图时模型很好地对其进行了上色和纹理填充，而输入为右半边的草图时生成的图片线条很不自然。其中的原因在于，具有详细实物信息的边缘图与简单抽象的草图之间存在很大的域差。从图 3-2 中可以明显看出，与草图相比，边缘图包含更多的线条和背景信息，而手绘草图与实物存在不对齐的情况，不能反映实物的真实边界。因此，模型的学习无法从对边缘图的生成迁移到对手绘草图的生成。

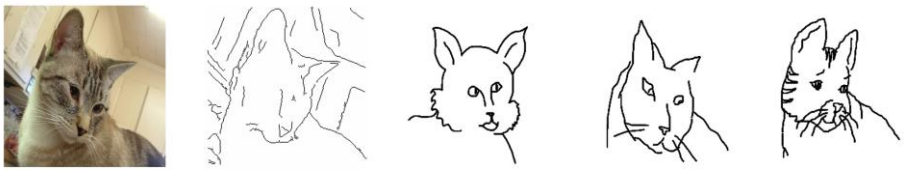

图 3-2　真实图片、边缘图和手绘草图的对比[1]

另一种可能的解法，是预训练图像转草图的模型[20]，用合成草图来填补真实草图的缺失。对于使用合成草图训练出的模型，图 3-3 给出了其输入分别为合成草图和真实草图时的生成结果。可以看出，输入为左半边的合成草图时生成效果良好。但面对右半边的真实草图输入时，模型虽然能对草图进行基本正确的上色和纹理填充，但不能很好地将背景和实物区分开来。造成此结果的原因在于，尽管肉眼下合成和真实草图极为相近，但在模型内部二者无法被区分开来。因此，单纯使用合成草图进行训练也不能很好地满足多类别草图图像生成的需求。

图 3-3　利用合成草图代替真实草图的模型生成结果

对于第二种方法，是否存在一种改善机制，能够提高合成草图的生成效果，减小与真实草图间的域差，使利用合成草图进行学习成为可能呢？结合 pix2pix 和 CycleGAN 两种经典模型中蕴含的思想方法，本文在 3.2 节中提出了一种可能的针对多类别草图翻译的解法。

3.2 基于开放域的多类别草图图像翻译

上面提到的问题中，其实涉及到了草图域和图像域两个域的学习和生成，这和 CycleGAN 模型是极为相似的。因此通过 CycleGAN 的模型思想，对草图和图像两个域进行联合学习，来同时提高模型在两个域中的生成精度，理论上来说就可以解决上述的问题。基于这样的想法，本节介绍了一种可行的草图图像多类别翻译方法，用未配对的草图图像数据进行训练，通过在开放域下联合学习草图转图像和图像转草图的模型并使用特定算法，提升模型在草图缺失时生成图像的质量。

3.2.1 模型目标

本模型会在开放域下设计模型，将数据集中草图缺失的类别和正常的类别放在一起进行训练和测试。以此为前提，模型很容易出现草图缺失的类别因学习数据不足而导致图像生成质量下降，或者转变为草图未缺失类别的域内生成等问题。因此本模型的目标是，通过一定的策略和算法，在保证生成图片质量的情况下，使模型对草图缺失的类别也能够顺利地进行学习和生成。图 3-4 直观地说明了这一目标。

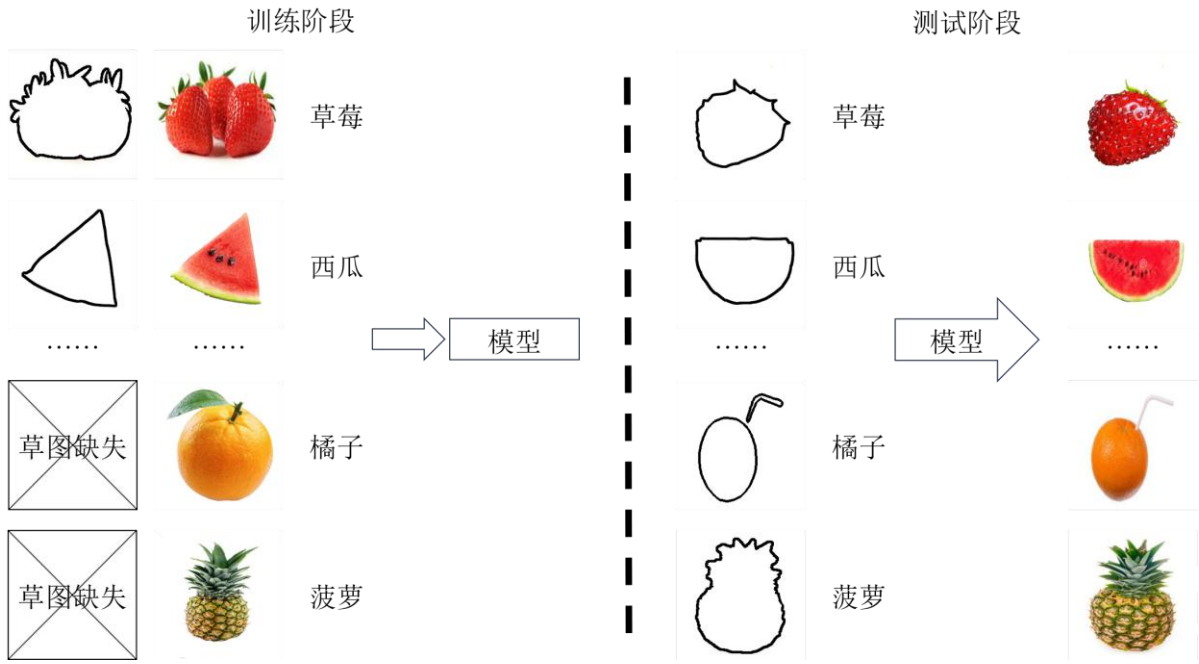

图 3-4　开放域下多类别草图图像翻译模型的目标

3.2.2 模型框架

本模型的算法流程图如图 3-5 所示，主要包括两个生成器，两个判别器和一个分类器。

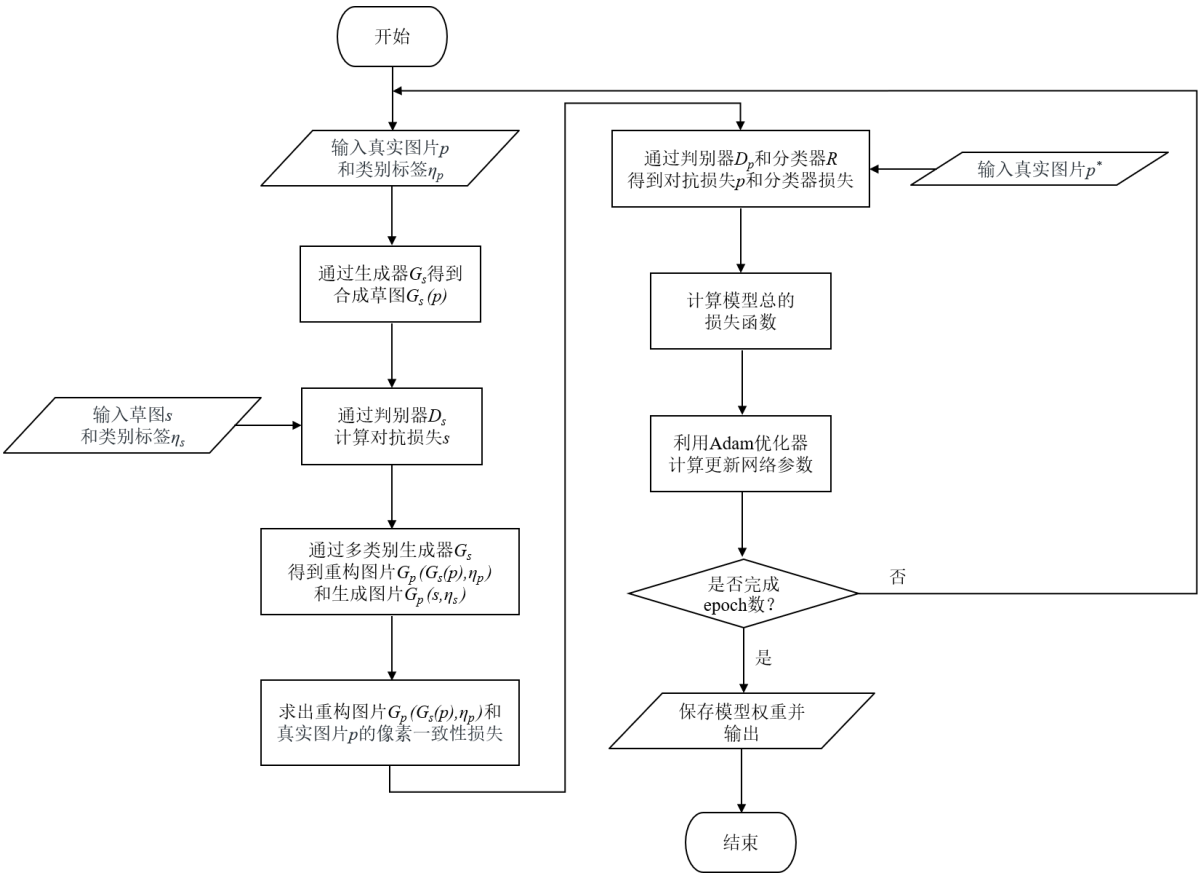

图 3-5　基于开放域的草图-图像多类别翻译模型

生成器 G_s 和判别器 D_s 构成一组从真实图像转换成合成草图的生成对抗网络，生成器 G_p 和判别器 D_p 构成另一组多类别的由草图生成图像的对抗网络。这种设置是受到前文 2.2.2 节中所介绍的 CycleGAN 模型的启发，通过对草图和图像两个域进行联合学习，使模型不管面对训练数据完整还是缺失的类别，都能够捕捉到其中的特征。分类器 R 会判断真实图片和生成图片的类别标签，以确保最终的输出能与输入标签 η_s 保持一致。

在训练过程中，一张真实照片首先输入生成器 G_s 得到合成草图 $G_s(p)$。在判别器 D_s 中，合成草图 $G_s(p)$和真实草图 s 会被进行比较，以不断提升合成草图的生成质量，减小合成草图与真实草图之间的域差。然后，合成草图 $G_s(p)$和真实草图 s 连同它们对应的标签 η_p 和 η_s 将被送入生成器 G_p，得到重构图像 $G_p(G_s(p),\eta_p)$ 和合成图像 $G_p(s,\eta_s)$。为了提高重建图像的生成质量，前文 2.2.1 节介绍的 pix2pix 重构损失的思想被引入到了本模型中，用来对输入图像和重构图像进行逐像素的比较。合成图像 $G_p(s,\eta_s)$ 最后会被送入判别器 D_p 和分类器 R：前者用来保证图像的真实性，后者用来确保模型针对不同类别的图像都能学习到相应的特征。

整个模型的损失函数主要包含四个部分：图像转草图网络的对抗损失 L_{G_s}，草图转图像网络的对抗损失 L_{G_p}，重构图像的像素一致性损失 L_{pix} 和合成图像的分类器损失 L_η。整体的目标函数如公式（3-1）所示：

$$\begin{aligned} L_{GAN} &= \lambda_s L_{G_s}(G_s, D_s, p) + \lambda_p L_{G_p}(G_p, D_p, s, \eta_s) \\ &+ \lambda_{pix} L_{pix}(G_s, D_s, p, \eta_p) \\ &+ \lambda_\eta L_\eta(R, G_p, s, \eta_s) \end{aligned} \tag{3-1}$$

实际在开放域中进行训练时，如果直接用公式（3-1）中的损失函数来训练多类别生成器，那么对于草图缺失的类别，其损失函数将不再与涉及真实草图的项相关，而转变为公式（3-2）所给出的形式：

$$L^*_{GAN} = \lambda_s L_{G_s}(G_s, D_s, p) + \lambda_{pix} L_{pix}(G_s, D_s, p, \eta_p) \tag{3-2}$$

这会导致草图转图像的生成器 G_p 将完全受到像素一致性损失的影响。对于像素一致性损失，常用的 L_1 和 L_2 正则化都会因为计算中均值的引入，导致生成的图像变模糊。针对这一问题，即将在 3.2.2 中介绍的随机混合采样算法可以在输入时最小化真实和合成草图之间的域差，提升草图缺失类别的生成质量。

3.2.3 开放域下的随机混合采样算法

开放域下的随机混合采样算法，旨在让生成器尽可能同等对待真实草图和合成草图。在草图缺失的类别中增加合成草图这一影响因素后，其损失函数不再只受像素一致性损失的控制，从而解决上述图像生成模糊的问题。该算法的实现主要分以下几步：

（1）准备一个缓冲区，存放通过生成器 Gs 生成的历史草图数据（包含合成草图和类别标签）。

（2）将真实图像 p 输入生成器 G_s，输出合成草图 $G_s(p)$，并将数据对 $(G_s(p),\eta_p)$ 存入上述的缓冲区中。

（3）按一定概率选取真实草图数据对或者缓冲区中的合成草图数据对，作为生成器 G_p 的输入。此概率的大小与草图缺失类别的数量有关，缺失的越多，采用生成的假草图作为输出的概率越大。

（4）生成器 G_p 对输入相应生成两张图像

（5）对上述过程计算损失函数

（6）重复第 2 步到第 6 步，直至模型收敛

整个算法的直观表达如图 3-6 所示。

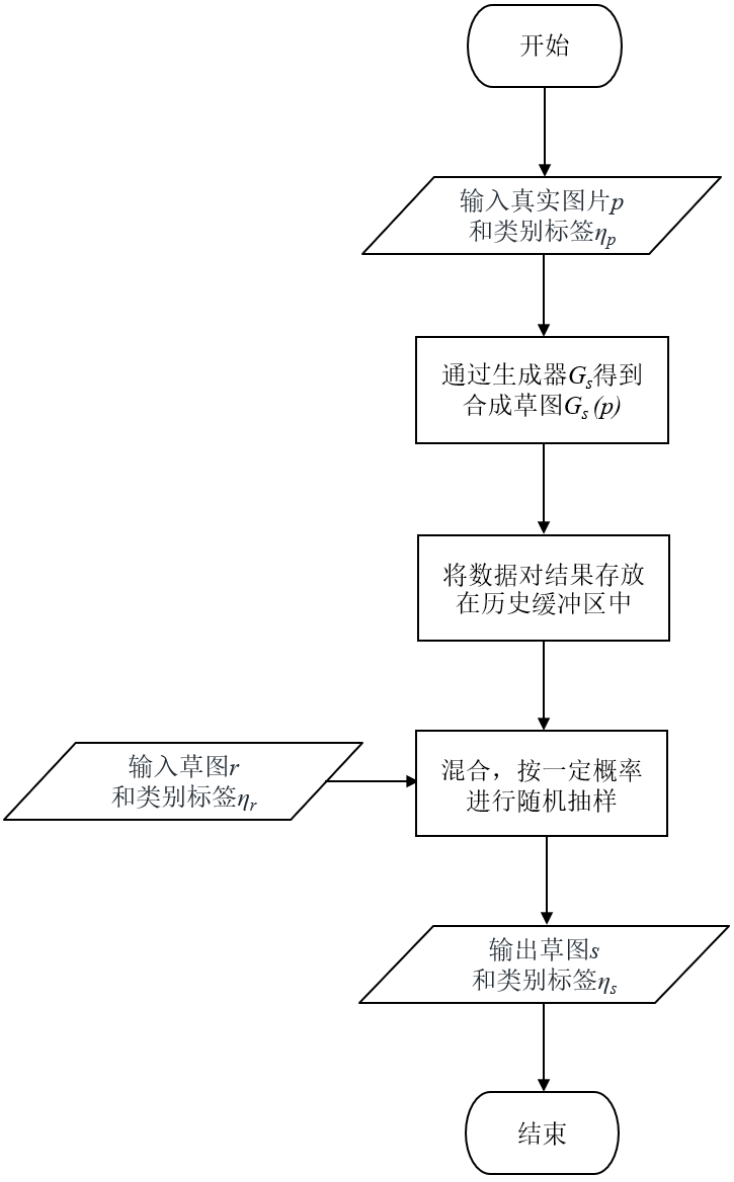

图 3-6　随机混合采样算法示意图

通过将真实草图和生成草图随机作为输入，可以模糊域内类别和开放域类别之间的界限，减小真实草图和生成草图之间的域差。相比于使用边缘图和单纯使用合成草

图进行数据增广，生成器 G_p 能够学习真假草图之间存在的变形，提高整个模型生成图片的质量。生成器 G_p 的损失函数也随之改变，由公式（3-3）给出：

$$\begin{aligned} L^{*}_{GAN} &= \lambda_s L_{G_s}(G_s, D_s, p) + \lambda_p L_{G_p}(G_p, D_p, s_{fake}, \eta_s) \\ &+ \lambda_{pix} L_{pix}(G_s, D_s, p, \eta_p) + \lambda_\eta L_\eta(R, G_p, s_{fake}, \eta_s) \end{aligned} \tag{3-3}$$

式中：s_{fake} —— 该策略下随机采样的图像。

此外，由于该采样算法只针对生成器 G_p，不会影响到判别器和分类器对真假数据的正常迭代。因此使用此算法，可以在不影响模型其它部件的情况下，让利用生成草图代替缺失的真实草图进行训练变得可行。

3.3 模型的具体实现

本小节将对实际训练过程中，模型框架内部生成器、判别器和分类器的具体实现进行说明，并对模型整体的目标函数进行分析。

3.3.1 生成器

1） 图像转草图的生成器 G_s

本模型中生成器 G_s 的结构如图 3-7 所示。此生成器采用了 Johoson J 等人提出的图像转草图模型[21]。该模型包含一个将 RGB 图像映射到特征空间的卷积层，两个下采样层，九个残差块，两个上采样层，和一个将特征空间重新映射回 RGB 图像的卷积层。此外，为了更好地对输入数据的特征进行学习，生成器 G_s 的网络中还使用了实例归一化处理。

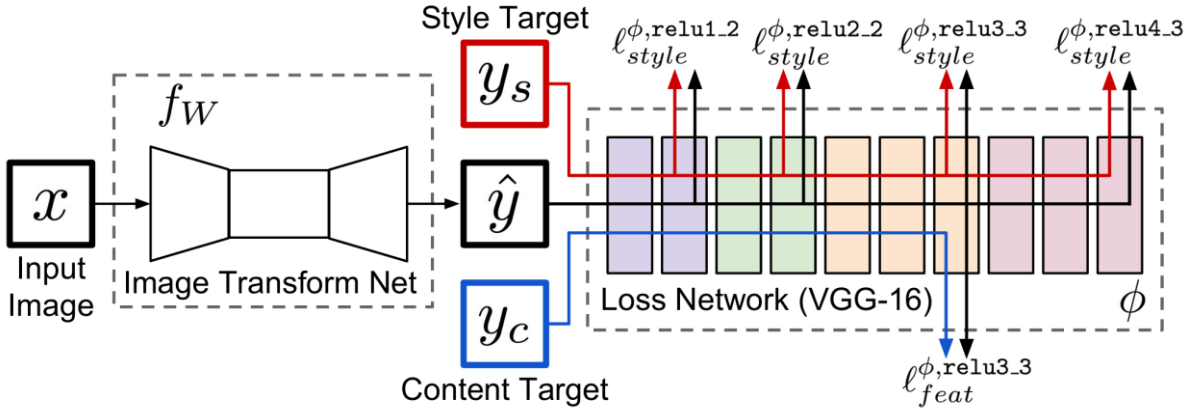

图 3-7 图像转草图生成器 G_s 的模型结构[21]

2） 草图转图像的生成器 G_p

本模型中生成器 G_p 的结构由图 3-8 给出。与生成器 G_s 相近，生成器 G_p 的网络结构中包括一个特征映射的卷积层，两个下采样层，九个残差块，两个上采样层，和一个映射 RGB 图像的卷积层。相比前者，为了进行多类别的图像生成，生成器 G_p 修改了网络内部的上采样层，并在残差块的处理上进行了相应的调整。

由于 G_p 的输入为草图和类别标签组成的图像对，因此需要对残差块的归一化层进

行修改。对此，本文采取的方法是自适应实例归一化[22]。在自适应实例归一化中，草图充当内容输入，而类别标签充当风格输入，确保网络对于每个类别都能学习到正确的背景、纹理和颜色等信息。另外，网络中还加入了卷积层和 PixelShuffle 层[23]，来对特征图进行上采样。子像素卷积的引入，能在减轻生成图像棋盘效应的同时，降低模型的参数量和计算量。

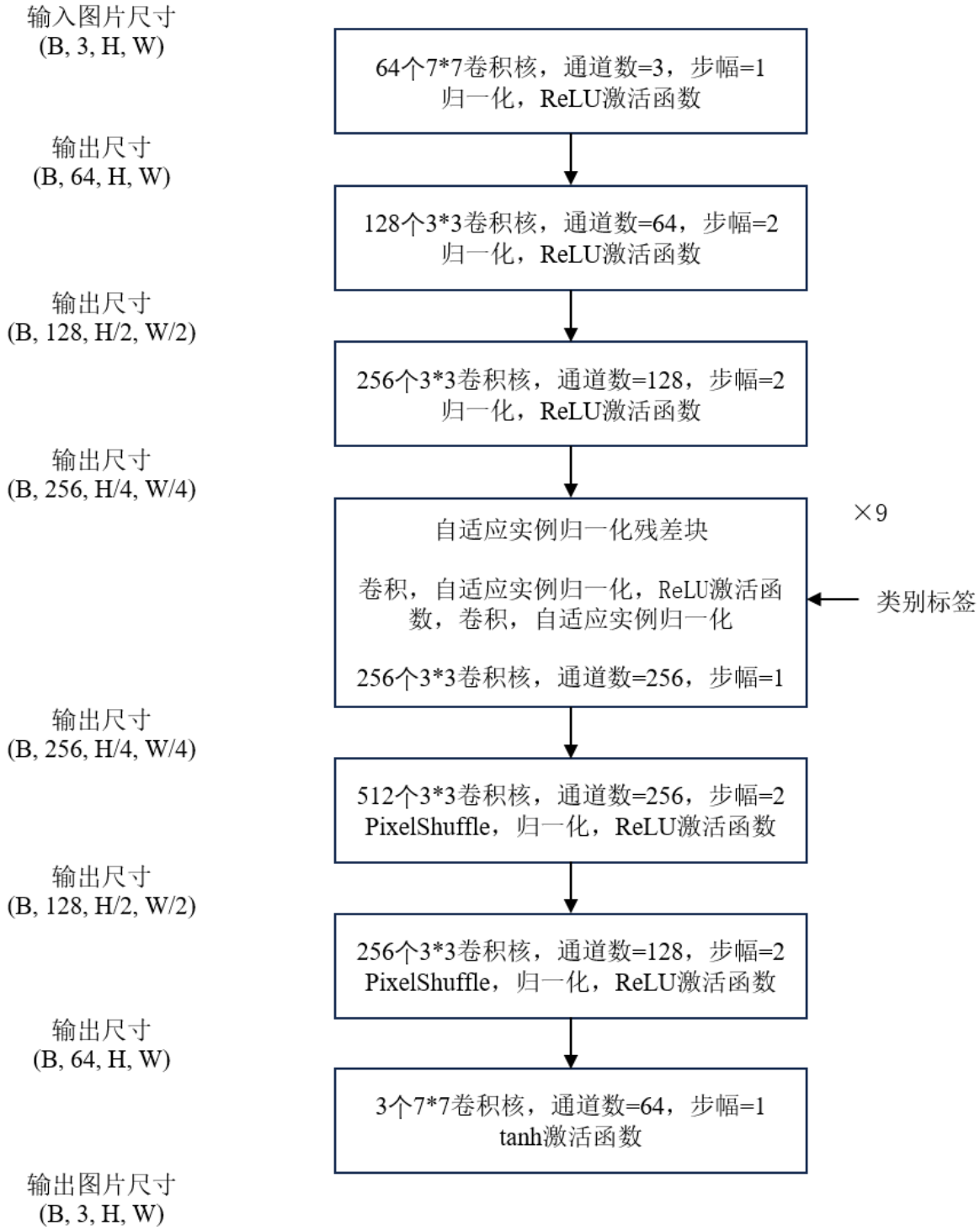

图 3-8　草图转图像生成器 G_p 的模型结构

3.3.2 判别器

判别器 D_s 和 D_p 的实现均使用了上文 2.2.1 小节介绍的 PatchGAN 模型。该模型包括了五个卷积层，能对分辨率为 256×256 的输入图像输出一个 30×30 的张量，张量中的每一个元素保存对输入图像各小块的判别结果。最终的判别结果由每一小块判别结果的均值给出。

3.3.3 分类器

图像分类器的实现了采用了上文 2.3.1 小节介绍的 HRNet 模型，模型结构如图 3-9 所示。数据先通过图中灰色部分(a)的低分辨率子网络，由 VGGNet[24]实现高分辨率到低分辨率的卷积；再通过黄色部分(b)的高分辨率恢复子网络，重新得到高分辨率的数据。

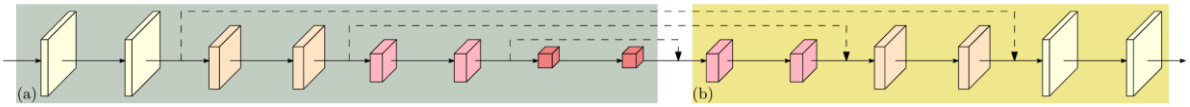

图 3-9　本文使用的 HRNet 模型结构[25]

HRNet 模型接收一张分辨率为 256×256 的图像作为输入，并输出一个 n 维向量作为分类结果。分类结果中向量的维数，会在输出前的全连接层中被修改为当前训练图像的类别个数。正如前文所说，HRNet 模型在分辨率较高时依然具有良好的性能，适用于本文的应用场景。

3.3.4 目标函数

首先给出公式（3-1）中每一项对应的损失函数表达式，如公式（3-4）到公式（3-7）所示：

$$L_{G_s}(G_s, D_s, p) = -\mathrm{E}_{p\sim P_{data}(p)}[\log_{D_s}(G_s(p))] \tag{3-4}$$

$$L_{G_p}(G_p, D_p, s, \eta_s) = -\mathrm{E}_{s\sim P_{data}(s)}[\log_{D_p}(G_p(s, \eta_s))] \tag{3-5}$$

$$L_{pix}(G_s, G_p, p, \eta_p) = \mathrm{E}_{p\sim P_{data}(p)}[\left\| G_p(G_s(p), \eta_p) - p \right\|_1] \tag{3-6}$$

$$L_{\eta}(R, G_p, s, \eta_s) = \mathrm{E}[\log P(R(G_p(s, \eta_s)) = \eta_s \mid G_p(s, \eta_s))] \tag{3-7}$$

对于模型中的两个判别器 D_s 和 D_p，给出其相应的损失函数，如公式（3-8）和公式（3-9）所示：

$$\begin{aligned} L_{D_s}(G_s, D_s, p, s) = &-\mathrm{E}_{s\sim P_{data}(s)}[\log D_s(s)] \\ &+\mathrm{E}_{s\sim P_{data}(s)}[\log D_s(G_s(p))] \end{aligned} \tag{3-8}$$

$$\begin{aligned} L_{D_p}(G_p, D_p, s, p, \eta_s) = &-\mathrm{E}_{p\sim P_{data}(p)}[\log D_p(p)] \\ &+\mathrm{E}_{s\sim P_{data}(s)}[\log D_p(G_p(s, \eta_s))] \end{aligned} \tag{3-9}$$

分类器的损失函数，由公式（3-10）给出：

$$\begin{aligned} L_R(R, G_p, s, p, \eta_s, \eta_p) = &\mathrm{E}[\log P(R(p) = \eta_p \mid p)] \\ &+\mathrm{E}[\log P(R(G_p(s, \eta_s)) = \eta_p \mid G_p(s, \eta_s))] \end{aligned} \tag{3-10}$$

对于分类器而言，真实图片和其标签能让其对每个类别学习正确的判别，而合成图片能够迫使其忽视域差，像对待真实图片一样对待生成图片。这也是模型中分类器需要和其它部分一起进行训练的原因。

模型对于判别器，采用了二值交叉熵损失；对于分类器，采用 Focal Loss[26]，用于在难分类的样本上分配更大的权重；对于图片重构中的像素一致性损失，采用了 L1 范数。

3.4 本章小结

本章利用第二章介绍的思想方法实现了开放域下的多类别草图图像翻译模型。首先，利用联合学习草图和图像的思想，使通过合成草图代替真实草图进行训练成为可能。其次，在开放域下加入分类器，控制每个草图类别的生成方向，避免其转变为正常类别的域内生成。最后，通过引入随机混合采样算法，解决了生成图像模糊的问题。本章对模型中生成器、判别器和分类器的具体实现进行了详细的说明，对模型的目标函数也进行了相关分析。

4 草图翻译模型的实验开展和实际应用

针对第 3 章提出的基于开放域的草图图像翻译模型，本章将开展相关实验，根据常见的评价指标，测试模型的性能，并与其它模型进行比较。然后，本章将对以该模型为基础搭建的在线图像翻译网站进行说明。

4.1 草图翻译模型的实验设置

4.1.1 数据集

对于后续的实验，本文使用的数据集均为 Scribble 数据集[13]。

Scribble 数据集是一个包含配对的草图图像数据的数据集，其中图片为白色背景下的实物图，草图是勾勒出实物图大致线条走势的手绘图。整个数据集包括橘子、菠萝、蛋糕等 10 个类别的数据，图片格式为 png 格式。

4.1.2 模型参数

在对模型进行训练和测试时，本文使用的平台为 GPU 云服务器，所用显卡为 1 张 Nvidia GeForce RTX 4090，运行系统为 Linux 系统。在训练过程中，重要的系统参数在表（4-1）列出。

表 4-1 模型训练中的重要参数

参数名	参数值
batch 大小	1
训练方向	BtoA
epoch 个数	200
GAN 模型	Vanilla
no_dropout	True
pool_size	50
生成器学习率	0.0001
判别器学习率	0.0004
数据集模式	unaligned
学习率策略	Ramp

针对上表给出的训练参数，作以下几点补充说明：

（1）训练方向为从草图域向图像域生成

（2）pool_size 代表随机混合抽样算法中，存储历史生成数据对的缓冲区的大小

（3）学习率策略 Ramp 使用的是 Pytorch 中通过 LambdaLR 自定义的函数：前 100 个 epoch 内学习率为 0.0002，后 100 个 epoch 内均匀递减至 0

4.1.3 评价指标

1） Fréchet Inception Distance

Fréchet Inception Distance (FID)是一种用于评价生成式模型图像生成质量的指标，它的基本思想是通过比较真实图像和生成图像的分布差异来量化生成图像的质量。FID 值越小，意味着生成图像与真实图像之间越接近，生成质量越高。相较于另一种常见的评价指标 Inception Score，FID 更能胜任评判图像真实度的工作。

本文使用预训练的 InceptionV3 模型[27]对生成图像和真实图像进行特征提取，并分别计算其图像特征的均值和协方差矩阵，最终得到 FID 的计算结果。FID 的具体计算由公式（4-1）给出：

$$FID(x,g)=\left\|\mu_x-\mu_g\right\|+Tr(\Sigma_x+\Sigma_g-2\sqrt{\Sigma_x\Sigma_g}) \tag{4-1}$$

其中，μ_x 和 Σ_x 分别是真实图像通过 InceptionV3 输出的特征向量的均值和协方差矩阵，而 μ_g 和 Σ_g 分别是生成图像对应的同类型的矩阵。实际代码实现时，矩阵根号的结果会出现复根，对此直接取其实部即可。

2） 分类器准确率

对于生成图像，使用训练好的分类器进行类别判断。判断准确率通过测试图像中类别判断正确的个数除以测试图像的总数来计算。判断准确率越高，显然说明模型对不同类别图像的学习效果越好，图像的真实度越高。

4.2 草图翻译模型的实验开展

整个翻译模型与 CycleGAN 的思想关系密切，因此是在 CycleGAN 项目的基础上进行上修改搭建的，并与 pix2pix，CycleGAN 以及 EdgeGAN 三个模型在相同的条件下的生成结果，进行了指标比较。

4.2.1 运行结果

在 Scribble 数据集下开展训练和测试时，本文选取了 6 个类别的数据进行实验，对每个类别的数据选取一定数目分别用作训练和测试，其中 2 个类别的草图数据是完全缺失的。数据集的具体划分情况如表（4-2）所示，模型最终的草图到图像生成结果如图 4-1 所示。

表 4-2 数据集的划分情况（带*的类别草图数据完全缺失）

类别名	数据对总数	训练草图数	训练图像数	测试草图数
饼干	151	146	146	5
菠萝*	156	0	151	5
草莓	150	146	146	4
蛋糕	151	146	146	5
橘子*	151	0	146	5
西瓜	151	146	146	5

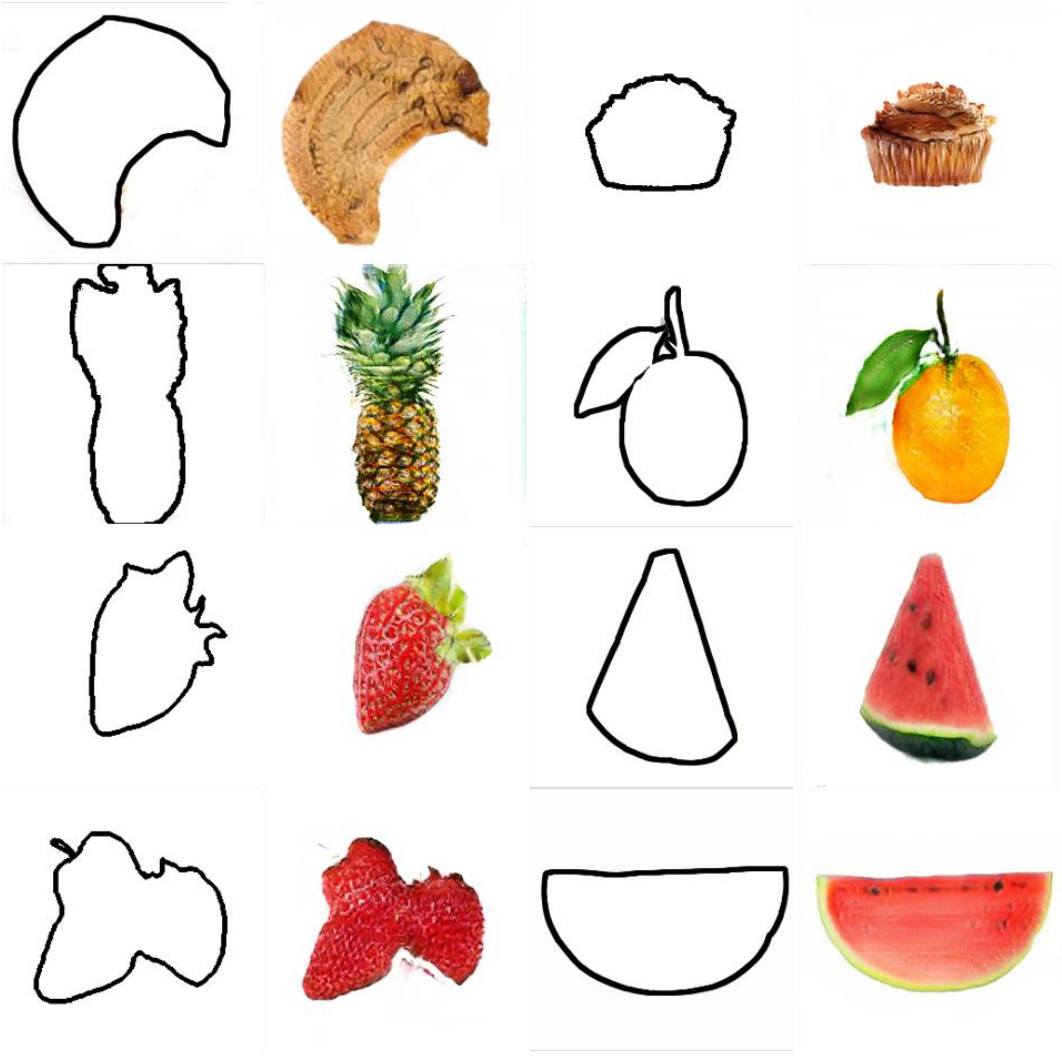

图 4-1　本文模型对 Scribble 数据集的生成结果

上图中，菠萝和橘子为草图缺失的类别，其余为正常类别。可以看出，模型对于两者都具有较好的生成效果。

4.2.2 性能比较

在同等条件下对前文提到的四个模型进行训练和测试，可以得到生成图像的结果。对于 pix2pix，CycleGAN 以及 EdgeGAN，由于它们并不能对草图缺失的类别很好地进行学习和生成，故在比较时只选用了草图数据完整的类别。通过 4.1.2 中介绍的评价指标，将每个模型的生成结果连同对应的真实图像输入进计算框架，输出的结果如表（4-3）所示。

表 4-3 不同模型生成结果的量化指标对比

性能指标	pix2pix	CycleGAN	EdgeGAN	本文模型
FID	385.9	283.3	267.4	**242.8**
分类准确率（%）	51.7	82.8	**100.0**	**100.0**

从上面的表中可以明显看出，本文使用的模型在生成图像的质量和准确度上都具有非常好的效果。进一步分析来说，本文模型在 CycleGAN 模型的基础上利用了其优点，且没有采用 EdgeGAN 中与真实草图存在较大域差的边缘图来进行学习，相比于 pix2pix 也引入了更加复杂有效的学习机制，因此具有更加优越的性能。

4.2.3 消融实验

为验证本文模型中的提出的算法和框架的必要性，在控制变量的条件下，针对原模型删减组件，开展了消融实验，结果如图 4-2 所示。图中的三种草图类别中，菠萝和草莓为正常的数据类别，而橘子为草图数据缺失的类别。相同输入下，列(a)为原模型，列(b)的模型去除了随机混合抽样算法，列(c)的模型使用训练好的图像转草图网络代替联合学习，列(d)的模型删掉了分类器。

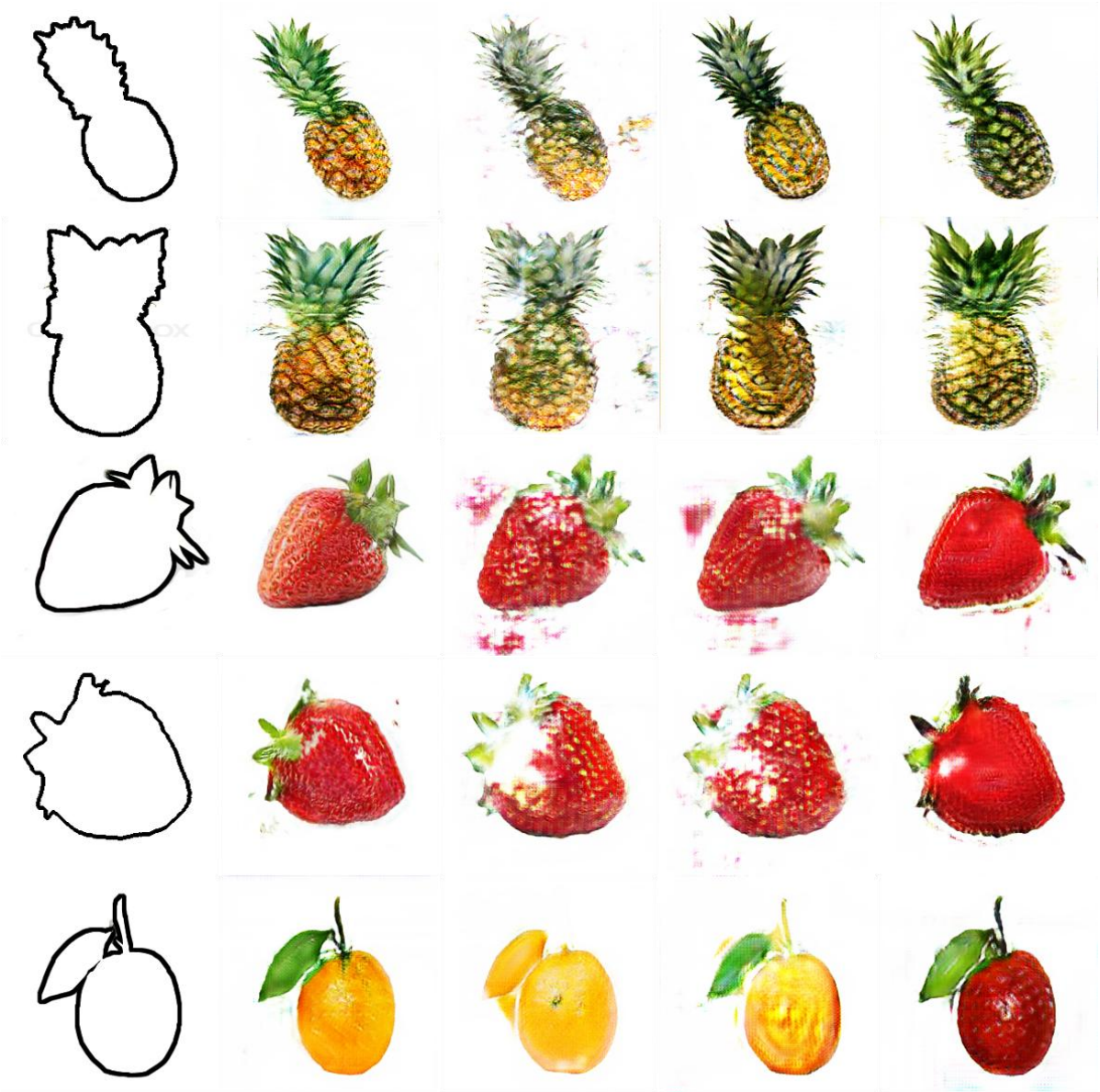

图 4-2　针对本文模型开展的消融实验的结果

从图中可以明显地看出，本文模型中的相关组件对提升性能具有重要作用。在不使用随机混合抽样算法的情况下，模型对于图片背景和内容的学习出现了问题；在不使用联合学习时，草图未缺失类别的生成效果相比于前一种情况有所提升，但是对于草图缺失的开放域类别，模型并不能学习到其内部的纹理信息；在将分类器删除后，可以很直观地通过上图中最右下角橘子的生成图像发现，模型失去了对图像类别的控制，只能用域内类别图像的纹理去填充开放域类别的草图。

此外，本文对缺失类别的个数对模型的影响也进行了研究。在总类别数为 6 的情况下，开放域类别从 0 增长到 3，依次增长的类别为橘子，菠萝和蛋糕，最终的结果如表（4-4）所示。

表 4-4 开放域类别个数对于模型的影响

缺失类别个数	0	1	2	3
FID	209.5	223.1	242.8	255.2
分类准确率（%）	100.0	100.0	93.1	96.6

从上表中可见，当开放域类别的数目增多时，FID 的值也随之增大。这是显然的：合成草图和真实草图之间的域差只是尽可能得被缩小了，而不可能被抹平，因此真实草图缺失得越多，模型在两个域上的域差越大，生成图片的效果自然也就下降。但是在分类准确率上，尽管草图有所缺失，总体的判断效果却并没有明显的下降，这是模型中分类器监督机制起到作用的体现。

4.3 草图翻译模型的实际应用

本小节介绍了为将本文中的多类别草图图像翻译模型应用到现实场景中，在服务器上搭建网站实现在线草图翻译功能的过程。

4.3.1 环境配置

搭建网站时使用的服务器为阿里云的轻量级云服务器，所用的系统为 Linux20.04 ，使用的服务器软件为 Apache2.4.59，用到的网页技术主要包括 HTML、JavaScript、CSS 和 php，其中 php 的版本为 8.3。

此外，原始模型是在高性能的 GPU 处理器下进行训练和测试的，而网站服务器使用的是 CPU 处理器，因此需要对模型的相关代码进行修改适配。由于 CPU 性能上的限制，草图翻译的类别太多时会出现处理时间过长甚至无法处理的情况，故此处的草图翻译模型仅针对菠萝和橘子两个类别。

4.3.2 前置分类器

本文的原始模型和过去的模型一样，都需要通过将存放图片的文件夹名设置成图片的类别名，来指定草图的类别标签，进行训练和生成。在实际应用中，传统做法是通过在网页中添加下拉菜单或单选框的方式，供用户在上传图片的同时，指定图片的

类别标签进行生成。但这种做法在图片类别较多时显然是不够智能的，增加了用户的操作。

为了提高模型的易用性和用户的使用体验，本文在前述模型的基础上加入了一个前置分类器。分类器使用的是前文 2.3.2 节中介绍的 EfficientNetV2 架构，用于对输入草图的类别进行学习和检测，代替用户完成指定类别标签的工作。实测效果发现该分类器的效果良好，对草图类别的识别准确率为 100%。在需要翻译的图像类别较多时，这种由后台完成图像的分类和生成工作、用户只需要输入图片就能得到输出的模式更加符合人的使用习惯，可以为未来的大模型生成提供一种思路。

4.3.3 网页编写

整个网站的处理流程如图 4-3 所示，图中的虚线框代表被封装起来的网页前后端交互部分，可以看出整个框架的封装性较好。

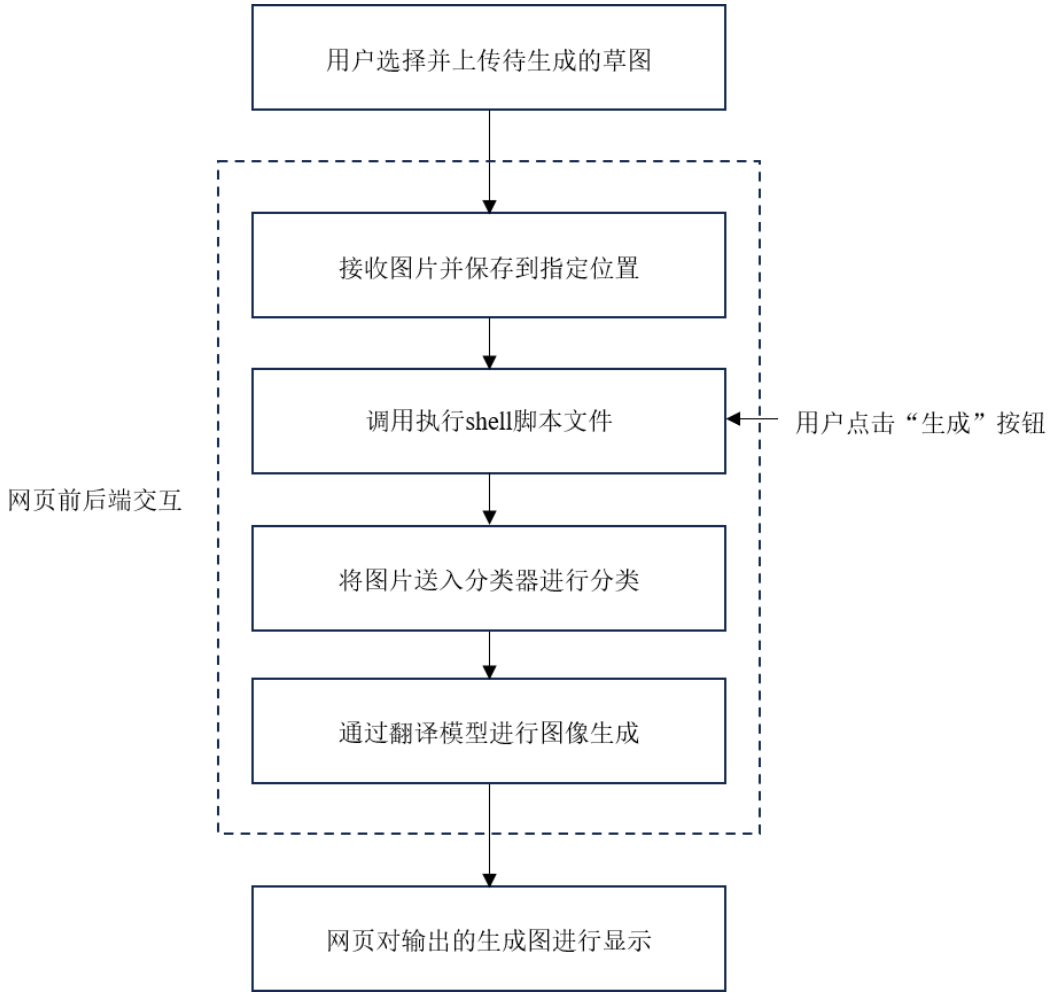

图 4-3　在线网站的处理流程图

为了实现上面的流程图，网站在代码编写上主要分为前端和后端两个部分。

首先在网页前端通过 HTML 表单，处理用户上传的图片。表单中的 action 属性可以调用相关的 php 文件，检测图片上传是否成功、后缀名与大小是否符合要求，并对满足条件的图片进行保存和显示。在用户点击页面上的“生成”按钮后，通过另一个

php 文件链接执行 linux 下的 shell 脚本，切换至后端的处理。

后端的处理主要由编写 shell 脚本来实现，涉及到切换目录、调用 miniconda 下的 python 环境以及执行 python 文件等指令。为了让 shell 能够顺利执行 python 文件，需要对原模型中 python 代码的相关接口进行适当更改，从而使输入的图片先后经过分类器和翻译模型，得到最终的输出，并保存在指定的位置。

最后在完成功能的基础上，可以在前端通过 CSS 层叠样式表对网站页面进行美化。

4.3.4 网站搭建结果

网站最终的搭建效果如图 4-4 所示。用户上传图片后，点击“生成”按钮，经过数秒的等待，即可在网站上看到输入的草图和生成的图片。

图 4-4　在线网站的运行结果截图

4.4 本章小结

本章对第三章中的模型进行了实验验证和实际应用。对于模型性能，通过 FID 和分类准确率这两个指标，与之前的模型进行了对比，证明了本模型具有良好的效果。对于模型的内部结构，通过减少相关组件开展消融实验，说明模型内的各组件对提升性能来说不可或缺。在整个模型的基础上，本章还给出了一种实际应用的框架，通过将模型适配进网站服务器的后端，实现了用户在线访问并执行草图图像翻译的效果。

5 结论与展望

5.1 本文工作总结

本文为了获得了质量较高的图像生成模型，在开放域下对草图和图像域开展联合学习，并使用随机混合抽样算法进一步提升生成效果。在该模型相对完善后，将其实际运用在了服务器网站上，以供在线生成图像。本文具体的研究工作如下：

（1）在具体研究工作之前，首先对多类别草图图像翻译中主要使用的模型和思想进行了介绍和总结。本文详细说明了生成对抗网络的原理，对 pix2pix 和 CycleGAN 两种经典模型进行了学习并给出了运行结果，对 HRNet 和 EfficientNetV2 两种高效分类器的结构和性能进行了研究。其中，两种经典草图翻译模型的思想均被运用在了后文的多类别草图图像翻译模型中，相关原理的说明也为后文模型里各组件的实现打下良好的基础。

（2）利用前文介绍的思想方法，本文实现了基于开放域的多类别草图图像翻译。与之前的模型相比，本文使用的模型较为有效地解决了多类别草图翻译中存在的草图缺失的问题。通过联合学习草图转图像和图像转草图的翻译网络，以及引入随机混合抽样算法，该模型对于草图缺失类别和正常类别都能生成质量较高的图像。为测试模型性能而开展的实验，通过对比本文模型和之前模型的评价指标，证明了本文模型具有更好的效果。此外，针对模型内部组件开展的消融实验也证明，该模型内部的框架和算法在提高模型性能上都发挥了积极作用。

（3）在综合考虑模型和网站服务器的性能后，本文给出了一种模型的实际应用方案。通过将模型的接口由 GPU 调整为 CPU，使得模型能够适配当前市面上绝大多数的网站服务器；通过在模型前引入前置分类器，使得用户不需要额外指定图像类别，提升了模型的易用性；通过网页的前后端处理，使得网站能够较为迅速地响应用户的需求，实现了在线的草图图像翻译网络。

5.2 研究展望

本文针对当前多类别草图图像翻译模型存在的问题，在开放域下实现了一种多类别的草图图像翻译模型。与之前的翻译模型相比，本文的模型具有生成图像质量好、分类精准度高、实用性强等优点。但该模型也存在一些不足，未来的研究可从以下几个方面开展改进工作：

（1）本文的模型所针对的输入输出图像分辨率均为256×256，在现实应用中，这样的分辨率往往是不够的。后续可以继续展开研究，让模型能够处理和生成更高分辨的图像。

（2）本文在实验过程中使用的 Scribble 数据集，其草图数据虽然均为手绘，但是

草图线条依然比较规则，较好地包含了实物图的轮廓信息。对于其它数据集以及更加不规则且抽象的手绘草图等情况，本文并没有进行实验验证，需要在后续研究中进行补充完善。

（3）模型的评价标准相对单一。本文只针对可以量化的评价指标进行了比较和测试。然而在现实应用中，草图到图像翻译模型的最终目标，是使得生成的图像在人眼的判别下也能够以假乱真。因此，后续的研究可以挑选实验对象，用人的肉眼来识别评判图像的真实度，以判断结果的准确率作为评判指标。这样更能评估模型在实际运用中的价值。

致　谢

对于本篇论文的产出，我首先要感谢指导我的杨旸老师和任嘉欣学姐。在完成毕业设计的过程中，她们给予了我非常大的帮助。在毕设的几个关键节点上，杨旸老师对我的研究给出了很有价值的建议，指出了我毕业设计中存在的许多问题，并给我以耐心的指导。任嘉欣学姐在整个过程中也给我提供了很多参考意见。从我在这个研究方向上刚刚入门，到有了一定的基础，学姐对于我提出的各种问题都进行了认真的回复，解答了我的诸多疑问。

其次，我要感谢参与论文评审和最终答辩的各位老师。谢谢你们能在百忙之中抽空阅读这篇论文，并给到我宝贵的指导意见。感谢论文评审阶段老师的耐心审阅，你们的批评与建议让我对这一领域有了更加充分的了解，不仅拓宽了我的视野，也能激发我对该领域更深层次的思考。同时，也感谢最终参加我论文答辩的各位老师，你们认真听取了我的毕设答辩，并且提出了许多有深度的问题和建议，让我能够认识到自己的不足。感谢你们的无私奉献和悉心教导，我会将这些教导铭记于心，并在未来的道路上不断努力，争取取得更大的进步。

最后我要感谢本科学习生活中遇到的三位舍友，感谢你们四年来的陪伴，从你们身上我学到了许多。最重要的是，我要感谢父母从小到大在学业上给我的支持，这是我能够顺利完成学业的根基。

作为一名即将毕业的学生，我深知，我取得的一切成果，都离不开周围人的支持和鼓励。在此，再次向本科四年里从各方面给到我帮助的老师、同学和家人，表达我最衷心的感谢！

参考文献

[1] Chen W, Hays J. SketchyGAN: Towards Diverse and Realistic Sketch to Image Synthesis. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition[C]. Salt Lake City, UT. 2018-06. IEEE, 2018: 9416-9425.

[2] 王建欣, 史英杰, 刘昊, 等. 基于 GAN 的手绘草图图像翻译研究综述[J]. 计算机应用研究, 2022, 39(8): 2249-2256.

[3] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Nets[J]. Advances in neural information processing systems, 2014, 27.

[4] Isola P, Zhu J Y, Zhou T, et al. Image-to-Image Translation with Conditional Adversarial Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)[C]. Honolulu, HI. 2017-07. IEEE, 2017: 5967-5976.

[5] Wang T C, Liu M Y, Zhu J Y, et al. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8798-8807.

[6] Zhu J Y, Park T, Isola P, et al. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. 2017 IEEE International Conference on Computer Vision (ICCV)[C]. Venice. 2017-10. IEEE, 2017: 2242-2251.

[7] Yi Z, Zhang H, Tan P, et al. DualGAN: Unsupervised Dual Learning for Image-to-Image Translation[C]. Proceedings of the IEEE international conference on computer vision. 2017: 2849-2857.

[8] Choi Y, Choi M, Kim M, et al. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8789-8797.

[9] Choi Y, Uh Y, Yoo J, et al. StarGAN v2: Diverse Image Synthesis for Multiple Domains[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 8188-8197.

[10] Kim J, Kim M, Kang H, et al. U-GAT-IT: Unsupervised Generative Attentional Networks with Adaptive Layer-Instance Normalization for Image-to-Image Translation[J]. arXiv preprint arXiv:1907.10830, 2019.

[11] Tang H, Liu H, Xu D, et al. AttentionGAN: Unpaired Image-to-Image Translation using Attention-Guided Generative Adversarial Networks[J]. IEEE transactions on neural networks and learning systems, 2021, 34(4): 1972-1987.

[12] Lu Y, Wu S, Tai Y W, et al. Image Generation from Sketch Constraint Using Contextual GAN[C]. Proceedings of the European conference on computer vision (ECCV). 2018: 205-220.

[13] Ghosh A, Zhang R, Dokania P K, et al. Interactive Sketch & Fill: Multiclass Sketch-to-Image Translation[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 1171-1180.

[14] Gao C, Liu Q, Xu Q, et al. SketchyCOCO: Image Generation From Freehand Scene Sketches. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)[C]. Seattle, WA, USA. 2020-06. IEEE, 2020: 5173-5182.

[15] Demir U, Unal G. Patch-Based Image Inpainting with Generative Adversarial Networks[J]. arXiv preprint arXiv:1803.07422, 2018.

[16] Zhang A, Lipton Z C, Li M, et al. Dive into Deep Learning[J]. arXiv preprint arXiv:2106.11342, 2021.

[17] Tan M, Le Q V. EfficientNetV2: Smaller Models and Faster Training[C]. International conference on machine learning. PMLR, 2021: 10096-10106.

[18] Tan M, Le Q V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks[C]. International conference on machine learning. PMLR, 2019: 6105-6114.

[19] Winnemöller H, Kyprianidis J E, Olsen S C. XDoG: An eXtended difference-of-Gaussians

compendium including advanced image stylization[J]. Computers & Graphics, 2012, 36(6): 740-753.
[20] Liu R, Yu Q, Yu S. Unsupervised Sketch-to-Photo Synthesis[C]. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16. Springer International Publishing, 2020: 36-52.
[21] Johnson J, Alahi A, Fei-Fei L. Perceptual Losses for Real-Time Style Transfer and Super-Resolution[C]. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. Springer International Publishing, 2016: 694-711.
[22]Huang X, Belongie S. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization[C]. Proceedings of the IEEE international conference on computer vision. 2017: 1501-1510.
[23] 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.
[24] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
[25] Wang J, Sun K, Cheng T, et al. Deep High-Resolution Representation Learning for Visual Recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2020, 43(10): 3349-3364.
[26] Lin TY, Goyal P, Girshick R, et al. Focal Loss for Dense Object Detection[C]. Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.
[27] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the Inception Architecture for Computer Vision[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2818-2826.