班　级　1702019

学　号　17020199055

西安电子科技大学

本科毕业设计论文

题　　目　多视角对比图表示学习

学　　院　电子工程学院

专　　业　电子信息工程

学生姓名　潘思多

导师姓名　邓成教授

毕业设计（论文）诚信声明书

本人声明：本人所提交的毕业论文《多视角对比图表示学习》是本人在指导教师指导下独立研究、写作成果，论文中所引用他人的无论以何种方式发布的文字、研究成果，均在论文中加以说明；有关教师、同学和其他人员对本文本的写作、修订提出过并为我在论文中加以采纳的意见、建议，均已在我的致谢辞中加以说明并深致谢意。

本文和资料若有不实之处，本人承担一切相关责任。

论文作者：_____（签字） 时间：　　年　　月　　日

指导教师已阅：_____（签字） 时间：　　年　　月　　日

摘　要

近几年来，随着深度学习的迅速发展，在大规模的训练数据集的支撑下，计算机视觉中的物体识别算法取得了突破性的进展。但人工收集和标注数据仍然是一项十分耗费人力物力的工作，例如对于濒临灭绝物种的识别，要收集到丰富多样的数据就十分困难。在给定有限或者没有训练图片的情况下，目前的模型很难预测出正确的结果。本课题基于对抗生成网络（Generative Adversarial Networks, GAN）对现有的零样本学习（Zero-shot Learning, ZSL）算法进行改进，利用对抗生成网络将视觉特征与语义特征结合，建立优化算法，使其完善图像空间与语义空间之间的对齐关系，从而提升模型的识别精度。

本文首先基于 GAN 建立了用以生成特征的 f-CLWGAN 图像生成模型，其特点是直接生成图像的特征而不是图像本身，通过生成器和判别器之间的相互对抗训练，使模型生成的图像特征足够真实并能够利用于识别训练。该模型首先通过优化分类损失正则化的 Wasserstein 距离优化语义空间和图像空间的对齐，从而合成未见类别的图像特征，再将该特征输入卷积神经网络进行训练，并最终提升物体识别的效率和精度。

本文在 AWA2、SUN、CUB 这三类具有不同粒度和大小的数据集上进行了实验，同时也对零样本学习和广义零样本学习设置下的两种情形分别进行了测试。在模型的建立过程中依次对模型的框架进行了改进，在实验中也对不同阶段的模型进行了同样数据集的测试，以此对比体现该特征生成模型在零样本学习问题下的适用情况。

最后本文对整个模型的框架结构和测试结果进行了整体的总结和评价，并对该模型在不同情境下的零样本学习问题进行了适用性分析，最终得出模型可以适用于大多数图像识别深层架构的结论。

关键词：零样本学习　对抗生成网络　深度学习　计算机视觉　图像识别

Abstract

In recent years, with the rapid development of Deep Learning and the support of large training datasets, object recognition algorithms in computer vision have made breakthrough progress. However, manual collection and labeling of data is still a labor-intensive task. In this paper, we improves the existing zero-shot learning (ZSL) algorithm based on the generative adversarial networks (GAN). Using the GAN to combine visual and semantic features, and to improve the alignment between image space and semantic space, so as to improve the recognition accuracy of the model.

First, We establishes a f-CLWGAN image generation model based on GAN, which generates features directly instead of the image. By training the generators and discriminators, the image features are real enough to be used in recognition training. The model first optimizes the alignment of semantic space and image space by optimizing Wasserstein distance of classification loss regularization, so as to synthesize image features without labels, then input the features into convolution neural network for training, and ultimately improve the efficiency and accuracy of object recognition.

We conduct experiments on AWA2, SUN, CUB datasets with different granularity and size, and also tests the model in two cases of ZSL and GZSL. In the process of building the model, the framework of the model is improved in turn. We also compare the applicability of the feature generation model under the ZSL problem.

Finally, we evaluated the framework structure and test results of the whole model. And the conclusion is that the model can be applied to most of the deep learning problems of image recognition.

Key words: ZSL GAN Deep Learning Image Recognition

目 录

第一章　绪论

1.1　计算机视觉及深度学习概况

作为人工智能领域备受关注的方向之一，计算机视觉被人们称作是赋予电脑以人类视觉能力的一门学科。计算机视觉能力在一定程度上能体现生物的视觉系统特征，即识别和处理视觉信息的能力[1]。然而，由于许多对普通信息的接收和输出过程都能描述为对信息的识别和处理，因此广义来看，这种视觉的定义并不严格，同时也与近几十年来该领域内的各项研究成果关注的方面有所出入，即缺乏更加具体细致的描述。

从另一个角度看，计算机视觉问题也被人们称为是研究视觉层面的信息感知和交互的问题。感知和交互，是指生物在对周围事物的感受和知晓中，识别、交流、互动和解释视觉信息的过程。由该种定义出发，即可将计算机视觉能力定义为对周围事物的表达和理解能力，该能力的核心便是对输入的视觉信息进行组织，对场景和物体的各类属性进行识别和判定，在此基础上再对图片或视频形式的视觉信息进行解释或互动。计算机视觉作为人工智能的一个分支，也有自己的独特性质。相比于更看重推断、梳理、判决和界定这些逻辑能力的人工智能，计算机视觉更关注的是对不同形式的视觉信息进行识别和交互的问题。总结即为，计算机视觉是以单帧图像或时间序列视频等形式的视觉信息为输入，将对周围事物的识别和交互为目标，研究视觉信息的各类属性、联系。从而对这些事物给出交互或解释的一门学科。

目前看来，该方向的大多数研究都聚焦在对视觉信息的交互阶段，如对图像中某个物体的检测和识别；利用简单的语言描述信息生成对应的图片；或是在不同的事物之中精确识别某特定物体的位置等属性[2]。其解释性地功能目前大多结合自然语言处理，对人类语言信息进行分析和解答，多集中在图片与文字问答交流等领域。在这些应用级的研究中，如何训练计算机让其拥有人类的视觉信息识别交互能力，是一个避不开的问题。常见的方法为搭建神经网络，不断学习给出的

样本并反馈误差来调节每个单独神经元的参数，从而实现输出结果不断靠近期望结果，利用这样一种学习到的能力，再去识别和处理新的或更多已有的视觉信息。在计算机视觉发展的初期，由于硬件的计算能力有限，人们往往找不到足够充沛的计算资源来使神经网络处理大量用以提升效果的数据样本，因此许多看似可行的方法都因为硬件条件不够而最终没有达到预期效果。近几年来随着硬件技术的飞速提高，计算机的体积越来越小而计算能力越来越强，使得基于大量数据样本的分析而进行的神经网络参数改进成为可能，在一个又一个浅层神经网络结构被证明有效后，人们开始设想模型的学习成果是否会和神经网络结构以及层数有更直接关系，从而提出了深层神经网络——即深度学习的概念。

深度学习的概念最早在 2006 年由多伦多大学的 G.E.Hinton 等提出，是一种基于样本数据通过特定的网络模型训练方法进行训练，最终得到包含多层网络结构的机器学习算法[2]。过深度学习得到的深度网络结构符合神经网络的特征，因此深度网络就是深层次的神经网络，即深度神经网络（Deep Neural Networks, DNN）。

深度学习作为自然语言处理和计算机视觉领域备受人们关注的一类算法，相比于浅层的神经网络或其它机器学习框架，其不仅具有结构更复杂，效果更显著等特点，最重要的是解决了许多复杂的模式识别问题。深度学习的模型中大量的神经网络层和节点使得更多更复杂的误差反馈和参数传递成为可能。在之前无法想象的许多应用，如机器翻译、个性化推荐、语音生成、图片生成等，在深度学习出现后都得以实现。近十几年来其在语言和视觉等领域取得的成果，也使人工智能的相关技术实现了质的飞跃，并且诸如人脸识别、语音识别等技术早已投入了人们的日常生活，供以实际使用。

从数学统计角度来看，深度学习在处理视觉信息的识别和交互问题时，关注的不仅仅是视觉信息的输入和输出之间的直接关系，更有对输入的样本信息和特征空间的映射关系。在行人检测的问题中，需要寻找的是人类手部及腿部等多个具有代表性的特征点，将它们分别映射到特征空间并进行后续的信息处理。而在物体的识别问题中，需要寻找的是图像中物体的轮廓、颜色、形状、大小以及物体和周边事物关系等视觉特征，将它们再分别映射到高维特征空间进行网络的学习和优化，并最终实现物体的精确识别。一般来说，此类映射方式可以大体分为三类：第一类为定义核函数将视觉信息特征映射到高维线性空间，如支持向量机（SVM），该种方法需要消耗大量计算资源，同时由于其涉及输入和输出间对应的

复杂函数转换关系，因此对数学专业知识要求也很高；第二类方法为人工编码供计算机进行视觉信息和高维特征之间的转换，此类方法与第一类一样，同样需要较大的计算资源和专业知识；而第三类方法，则是利用多层非线性的神经网络层来搭建深度学习框架，使计算机具有自主学习的能力。直观层面即为，通过样本的训练，来教计算机学会学习，这便是深度学习。由于层数多、框架复杂等特点，因此在同样数量的参数和样本条件下，相比于上文所述的浅层学习框架，深度学习能够更好地拟合非线性函数，从而实现更好更快的学习效果。

深度学习如今作为机器学习领域的常用方法，已在自然语言处理和计算机视觉等众多应用级领域中经过考验，其灵感来自于建立类似人体脑部中的神经元节点，并在节点之间设置激活单元和输出单元，使其在从外界视觉信息或音频信息等信号时，可以通过多个变换状态来对信息进行分层处理和描述，进而可以给出数据中包含的细节和深层信息，通过网络的逻辑分析和数理推算，最终实现对该信息的交互和解释。在以人类为主的灵长类视觉系统中，彩色静止图像信号的处理流程依次为：首先输入图像的像素信息，通过处理其 RGB 三层通道的信息来检测物体的轮廓，从而得到物体初步整体形状，逐步传递这些特征从而使模型能够检测到图像中对应该物体的颜色、花纹等细节属性，进而实现物体识别。深度学习通过组合低维特征来形成抽象的高维空间表示或其它特征，之所以被人们称为“深度”，是相对于前文提到的定义核函数和人工编码等方式而言的。相比于普通神经网络，在深度学习的神经网络模型中，非线性神经层会更多。前者依靠人工经验来提取样本数据的特征信息然后输入网络，最终获得的是没有层次结构的单一特征；而后者通过对原始的视觉信息进行了层与层之间多次特征映射，将样本在原始特征空间表示的内容映射变换到新的更高维的特征空间，从而获得学会学习能力，即层次化结构的更有利于分类和传递的特征表示[3]。

相比于传统的神经网络，深度学习算法打破了对神经网络的层数限制，网络模型的设计者可以根据实际解决问题的需求来自行选择神经网络的层数，除此之外，相比于传统的神经网络通过随机设定每一个参数的初始取值，并采用后向传播的神经网络算法通过梯度下降的方式进行网络训练。而深度学习算法则将每层分开对待，使该层训练完后，新的一层将前一层的输出特征作为输入值输入并进行训练，并在每层参数训练完毕后，整个网络再进行有监督学习式的微调，这便是深度学习算法的整体思路。

1.2 国内外研究现状

近几年来，随着深度学习的迅速发展，计算机视觉中的物体识别算法在大规模的训练数据集的支撑下取得了突破性的进展。然而在这过程中，人工收集和标注数据仍然是一项非常消耗人力物力的工作，例如若要实现对数量本就不足的濒临灭绝物种精确识别，要收集到丰富多样的数据就十分困难。在给定有限或者没有训练图片的情况下，目前的模型很难预测出正确的结果。目前而言，自 2008 年 Larochelle 等人针对字符分类问题提出了零样本学习（Zero-shot Learning, ZSL）方法以来[4]，这一训练方法已在目标识别任务中得到普及应用。传统的目标识别方法是通过将图像标签分配到训练集中见过的一个类别来预测目标实例的存在，ZSL 则与传统方法不同，其目标是识别之前从未见过的新类别中的目标实例[5]。常见的 ZSL 模型有 DAP、ALE 和 SAE 模型，其基本原理都是设置一个共有的嵌入空间以及相应的映射函数，该函数的目的是对于已见过或未见过的样本类别都进行图像特征与语义特征之间的相容性衡量，从而实现类别的精确分类。在 ZSL 领域，图像空间与语义空间对齐问题是近期研究的侧重点。

另一方面，自 Ian Goodfellow 等人于 2014 年提出对抗生成网络（Generative Adversarial Networks, GAN）以后[6]，这一方向目前也已成为人工智能方向的研究热点。GAN 借鉴二人零和博弈的核心思想，整个模型由生成器和判别器两部分组成。其中，生成器负责将随机信号转化成图像，判别器则负责判别输入的图像是来自真实样本还是生成样本。整个过程中生成器和判别器不断对抗，最终达到彼此共优的效果，即生成器能够生成足够真实的样本数据，判别器也能对逼真的伪造样本进行判别。结合现有成果来看，GAN 可应用于不同领域，主要为计算机图像和视觉领域。除了能生成高分辨率逼真的图像，也能对图像进行修复、风格迁移，以及生成视频并进行预测等[7]。借用其博弈共优的思想，GAN 也还可以生成文本，进行对话生成、机器翻译、语音生成，并在音乐、密码破译等其它领域也都有应用[?]。

对抗生成网络的提出无疑是新颖的，但同时也存在着不少缺点，比如梯度消失问题、模式崩溃等。随着研究的深入，GAN 不断优化扩展，其衍生模型也层出不穷。如何提高 GAN 在其他领域的应用效果也是个值得深入研究的问题，同时也是其接下来的发展趋势所在。

1.3　课题研究目的及意义

本课题目的在于结合对抗生成网络对零样本学习算法进行优化。利用对抗生成网络将语义转换为对应的视觉特征，与原有的视觉特征结合，将零样本学习问题转换成一个普通的分类问题，从而使其解决图像空间与语义空间之间的对齐关系，提升样本的判别性和识别精度，从而在给定有限或没有训练图片的情况下使模型预测出正确结果。

此外，在大规模的训练数据集的支撑下，计算机视觉中的物体识别算法在近几年取得了突破性的进展。但是人工收集和标注数据是一项十分耗费人力物力的工作，对于一些数据本身有限的问题，要保证数据的多样性则更为困难。结合对抗生成网络的零样本学习算法能够有效解决训练数据不足以及识别精度不够的问题，是将深度学习与机器学习巧妙结合在一起的创新性研究。本课题在解决零样本学习算法未能有效解决的图像空间与语义空间对齐的问题上提出了可行的研究方案。

1.4　研究任务及主要工作

本课题核心任务为，在现有的零样本学习算法基础上，结合对抗生成网络的博弈共优思想，对算法中的低维到高维映射关系进行改进。利用对抗生成网络将视觉特征与语义特征结合，建立优化算法，使其完善图像空间与语义空间之间的对齐关系，从而提升模型的识别精度。

本文将利用对抗生成网络将语义转换为对应的视觉特征，与原有视觉特征结合，将零样本学习问题转换成普通分类问题，从而解决零样本学习中图像空间与语义空间的对齐关系问题。项目所需要完成的任务为了解零样本学习以及对抗生成网络并重点学习深度学习算法及框架；借鉴现有方法，利用未见类别的语义生成具有类间差异性的视觉特征，完成精准的零样本识别模型；最后使用基于 TensorFlow 的深度学习框架实现模型，并在多个数据集探究算法的有效性和优越性。

论文共分为四章：

第一章为绪论，介绍课题背景与研究现状等相关内容。

第二章介绍对抗生成网络和在此基础上设计的零样本学习模型。

第三章实验验证、实验结果对比分析。

第四章对提出的模型和所进行的实验进行总结，并对后续工作进行展望。

第二章　基于对抗生成网络的零样本学习模型

本章将对基于 GAN 的 ZSL 模型搭建过程进行详细介绍，首先将分别简单介绍模型涉及的 GAN 和 ZSL 原理和结构，引出将 GAN 应用到 ZSL 情境下生成图像特征的模型主要思想。然后将从特征生成、样本分类两个角度来对模型进行逐步搭建，过程中将介绍模型的定义和模型不同版本的改进优化，最后得到完整的零样本学习模型。

2.1　对抗生成网络

借鉴博弈论中的纳什平衡思想，GAN 的整体框架为设定一个两方参与的游戏，一方为生成器（Generator），负责捕捉真实样本的内在信息, 并将输入的随机噪声信号通过特定的映射方式转化为新的数据样本。另一方为判别器（Discriminator），负责判别输入的样本是真实数据还是生成器生成的数据样本[6]。

图 2.1　对抗生成网络的结构示意图

如图2.1所示，生成器的目标是将随机噪声信号转化成足够逼真的虚假图像，使判别器无法识别出来，而判别器的目标通过比对输入的真实数据来判定出全部的

虚假样本。为了实现各自的目标, 这两方需要不断优化, 各自提高自己的生成能力和判别能力, 整个对抗和学习的优化过程就是寻找二者之间的一个纳什平衡。

在 GAN 的发展前期，其最初是作为一种学习生成模型而提出的，多被应用在图像生成的领域[8]。其中输入生成器的样本为遵循某种特定分布的随机噪声信号，如多元高斯分布等。在 GAN 的发展历程中，DCGAN[9] 通过利用深度卷积神经网络为 GAN 提供了较好的扩展实践训练方案。InfoGAN[10] 则通过将图像的生成过程分解为样式和结构网络来改进 DCGAN，最大化了可解释的潜在变量分布和生成器之间的相互信息。通过将类别标签和句子描述输入生成器和判别器，人们还将 GAN 扩展为了条件 GAN[11]。

为了解决 GAN 的训练过程中出现不稳定的问题，Wasserstein-GAN（WGAN）被提了出来[12]。Wasserstein 距离的起源是最优传输问题，该问题把概率分布想象成一堆石子，过程中移动该堆石子，通过最小的累积移动距离把它堆成另一个目标形状。相比于其它的距离度量方式，Wasserstein 距离能够自然地度量离散分布和连续分布之间的差距，并给出距离度量。还可以在变换为另一分布的同时保持分布自身的几何形态特征。通过将 Wasserstein 距离应用到 GAN 中来构建的 WGAN 比原始 GAN 具有更好的理论性能，但由于过程中对判别器实施了 1-Lipschitz 约束的权重限制，WGAN 仍然遭受消失和爆炸梯度问题的困扰。在该基础上，人们则提出了一种 WGAN 的改进版本，即通过梯度惩罚来进行网络表现的改善[13]。

尽管上述这些方法的 GAN 模型都生成了逼真的图像，但人们均未将 GAN 的思想应用于图像的特征生成。为了进一步改善图像空间和语义空间的对齐关系，本文采用了一种基于 GAN 架构的直接生成卷积神经网络（Convolutional Neural Networks, CNN）特征的模型，它是一种可用于训练零样本学习的判别式分类器，结合了 WGAN 中的损失和约束来将生成的特征损失分类加以区分。

2.2　零样本学习

随着人工智能各个领域的飞速发展，机器变得越来越聪明。但诸如识别、检测等这些视觉领域技能的实现，都依赖于其背后庞大的训练样本数据库。如果没有一定量的标注训练数据，目前的机器无法区分两个很相似的物体，也无法去识别一个从未见过的物体[14]。因此人们便提出了对样本数量要求更少或者无须对应类别样本数据的学习，即零样本学习。

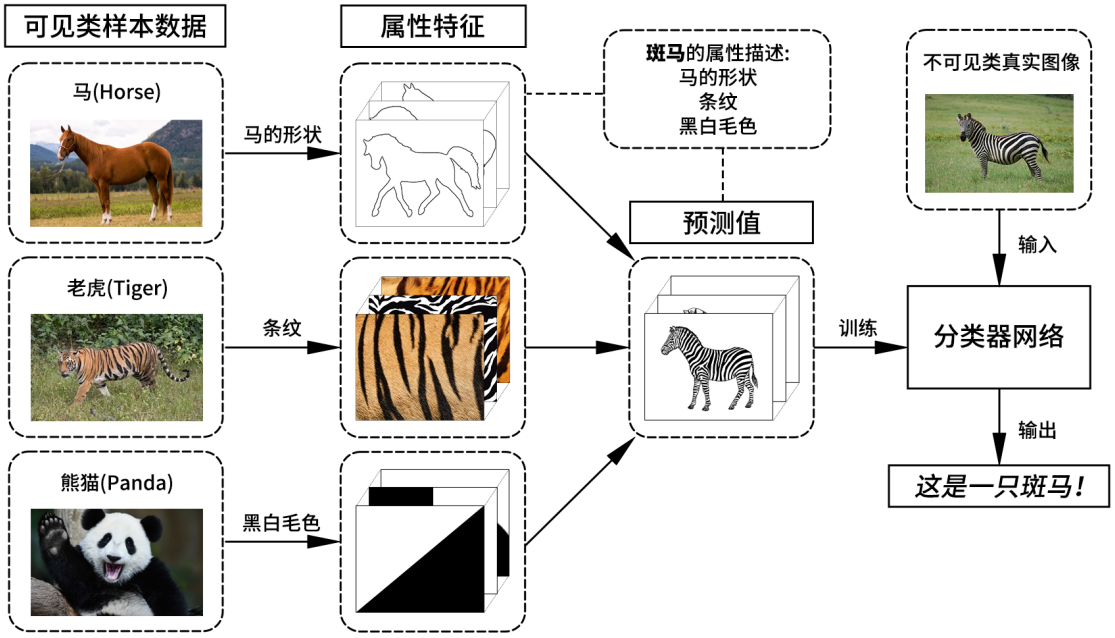

图 2.2 零样本学习示意图

如图2.2所示，结合人类生活中的实际零样本学习例子来看，当我们告诉一位从来没有见过斑马的小孩，有这样一种动物，它有马的形状、老虎的条纹和熊猫的黑白毛色，这种动物叫做斑马。即通过这样的语义级的属性描述，来告诉机器有这样一种类别的物体，它具有可见类中已知物体中的某个或某些属性，机器将自己组合之前学到的属性特征并对新输入的不可见类未知图像进行判定，从而实现识别。

对机器来说，零样本识别依赖于可见类的标注数据集和不可见类中与之相关的语义描述。而人类能够实现零样本学习，很大程度上是因为自己已有对这些事物的自然语言描述基础，如上文举例中对小孩进行语义描述后，建立了他对不可见类的斑马和可见类的条纹、毛色等视觉概念关联，从而使他能够在初次见到该类别物体时识别出来。基于此思想，人们将 ZSL 模型定义为对未标注标签的不可见目标类样本的识别模型。ZSL 利用的是可见类中的语义属性，将它们作为辅助信息，并迁移到不可见类的样本识别中。ZSL 包含两个阶段：（1）训练阶段：这个阶段获取可见类的语义特征描述，并将它提取为属性知识；（2）推理阶段：结合整理综合出的可见类语义属性特征描述，来对不可见类的新样本进行识别[15]。在 ZSL 整个学习和识别的过程中，最关键的一步便是利用语义空间和图像内容的对齐关系来进行视觉特征的表示。为了实现这个目的，人们常使模型先学习并构建一个联合的嵌入空间，其中语义向量（可见类的样本原型特征描述）和视觉特征

向量都被投影到这个空间。接着再使用最近邻搜索在该嵌入空间上匹配图像特征和未见类样本的投影进行反馈和优化，最终实现不可见类的零样本学习。

在 ZSL 的设置条件下，模型训练期间所见到的样本数据集合和测试期间评估的样本集合之间是不相交的。由于不可见类只具有物体样本图像而没有对应的标签，因此不能将监督式学习方法用于此类 ZSL 问题。另一方面，虽然 ZSL 吸引了很多注意力，但在可见类和不可见类都出现在测试期间的广义零样本学习（Generalized Zero-shot Learning, GZSL）情境下，ZSL 似乎没有太多实际作用。本文将通过基于 GAN 的 ZSL 模型为不可见类生成 CNN 特征，进而解决 GZSL 情境下的识别问题。本文的工作将结合 WGAN 使分类损失进一步改善，并在 ZSL 和 GZSL 两种情境任务下对的大多数数据集进行测试验证。

2.3　零样本学习中的特征生成和样本分类

就现有的 ZSL 模型而言，其在训练过程中只能从可见类中看到标记过的数据，从而会将预测结果偏向可见类。本文提出的模型思路为，通过提供不可见类的综合 CNN 特征，经过训练的分类器可基于此继续探索不可见类的词嵌入空间，并最终进行语义特征和图像特征的对齐。本文所采用的方法关键是生成以不可见类中特定语义属性向量为条件的 CNN 特征分布，并利用该特征来使模型进行分类学习。由于该模型可以生成的 CNN 特征在数量上没有限制，理论上可以生成与可见类中数量一样多的数据样本，因此该方法将改善可见类和不可见类之间的样本特征不平衡，从而达到提升训练结果的目的。与此同时，本文建立的模型也将允许直接训练有区别的分类器，即 Softmax 分类器，即使对于不可见类也可以进行同样的训练。

下面将逐步搭建本文基于 GAN 的 ZSL 模型。首先对模型的主体变量和参数进行定义，令

$$\mathcal{S}=\left\{\left((x,y,c(y))\middle|x\in\mathcal{X},y\in\mathcal{Y}^{\mathcal{S}},c(y)\in\mathcal{C}\right)\right\}, \tag{2-1}$$

其中 $\mathcal{S}$ 代表可见类的训练样本数据，$x\in\mathbb{R}^{d_x}$ 是 CNN 特征，y 表示类标签，属于 $\mathcal{Y}^{S}=\{y_1,\ldots,y_K\}$ 中由 K 个离散可见类组成的集合，而 $c(y)\in\mathbb{R}^{d_c}$ 是类嵌入，例如 y 类的属性描述，用于对类之间的语义关系进行建模。另对不可见类的训练数

据标签建立集合，令

$$\mathcal{U}=\left\{\left((u,c(u))\middle|u\in\mathcal{Y}^{\mathcal{U}},c(u)\in\mathcal{C}\right)\right\},\tag{2-2}$$

其中 $\mathcal{U}$ 代表不可见类的样本数据，$\mathcal{Y}^{\mathcal{U}}=\{u_1,\ldots,u_L\}$ 为标签集合，其类别嵌入集是可用的，但缺少了图像和对应的图像特征。对于给定的 $\mathcal{S}$ 和 $\mathcal{U}$，ZSL 的任务是训练分类器 $f_{zsl}:\mathcal{X}\rightarrow\mathcal{Y}^{\mathcal{U}}$，而在 GZSL 中，任务则为训练分类器 $f_{gzsl}:\mathcal{X}\rightarrow\mathcal{Y}^{\mathcal{S}}\cup\mathcal{Y}^{\mathcal{U}}$。

2.3.1 特征生成

GAN 整体是由一个生成器网络 G 和一个判别器网络 D 组成，它们在双方极小与极大的对策中互相竞争。在生成图像像素信息的过程中，D 试图准确地将真实图像与生成的图像区分开来，而 G 则试图通过生成足以假乱真的图像来愚弄判别器。在此基础上，我们通过在 G 和 D 中都添加一个条件变量来将 GAN 扩展为条件 GAN[14]。下面本文将给出逐步搭建条件 GAN 模型的详细内容。本文所采用的 GAN 模型生成的是图像特征而不是图像像素，并且由于确定了图像特征和语义空间的对齐关系，因此若只训练可见类的样本数据，也可以达到生成未可见类图像特征目的。

（1）f-GAN 模型

在给定所见类的训练数据的情况下，模型的目标是训练出一个条件生成器 $G:\mathcal{Z}\times\mathcal{C}\rightarrow\mathcal{X}$，它以随机高斯噪声 $z\in\mathcal{Z}\subset\mathbb{R}^{dz}$ 和类嵌入 $c(y)\in\mathcal{C}$ 为输入，输出一个 y 类别中的 CNN 图像特征 $\widetilde{x}$，其中 $\widetilde{x}\in\mathcal{X}$。一旦生成器 G 通过学习可见类的类嵌入条件 $c(y)\in\mathcal{Y}^{\mathcal{S}}$ 来生成真实图像的 CNN 特征，即 $\mathcal{X}$。那么它将同样也可以通过类嵌入 $c(u)$ 生成任何不可见类中 u 对应的 $\widetilde{x}$。该特征生成模型 f-GAN 的优化函数如下：

$$\min_{G}\max_{D}\mathcal{L}_{GAN}=E\left[\log D(x,c(y))\right]+E\left[\log\left(1-D(\widetilde{x},c(y))\right)\right],\tag{2-3}$$

其中 $\widetilde{x}=G(z,c(y))$。且判别器 $D:\mathcal{X}\times\mathcal{C}\rightarrow[0,1]$ 是以 Sigmoid 函数为最后一层的多层感知器。当 D 试图使损失最大化时，G 则试图使其最小化。尽管 GAN 已经被证明能够捕获复杂的数据分布，生成例如像素图像等视觉信息，但实际来看，它仍然较难获得一个较好的结果[16]，从另一个角度来看，引入 Wasserstein 距离来衡量

高维空间的差异可以起到更好效果，因此本文在此基础上提出改进的 f-WGAN 模型。

（2）f-WGAN 模型

我们将改进的 WGAN 扩展为一个条件 WGAN，将嵌入类 $c(y)$ 集成到生成器和鉴别器中。定义损失函数 $\mathcal{L}_{WGAN}$ 为

$$\begin{aligned}\mathcal{L}_{WGAN} =& E\left[D\left(x, c\left(y\right)\right)\right] - E\left[D\left(\widetilde{x}, c\left(y\right)\right)\right] - \\ & \lambda E\left[\left(\left\|\nabla_{\hat{x}} D\left(\hat{x}, c(y)\right)\right\|_2 - 1\right)^2\right],\end{aligned} \tag{2-4}$$

其中，$\widetilde{x} = G(z, c(y))$，$\hat{x} = \alpha x + (1+\alpha)\widetilde{x}$，其中 $\alpha \sim U(0,1)$，而 λ 则为惩罚系数。与 GAN 相比，此处的判别器网络定义为 $D : \mathcal{X} \times \mathcal{C} \to \mathbb{R}$，它取代了 Sigmoid 层并输出了实数值。同时定义中也删除了公式2-3中的对数项，取而代之的是方程式2-4中的前两个项近似取值的 Wasserstein 距离。第三项则是梯度惩罚项，它强制 D 的梯度沿直线并在成对的实点和生成点之间具有单位范数。同样定义优化问题为

$$\min_{G}\max_{D} \mathcal{L}_{WGAN}. \tag{2-5}$$

（3）f-CLWGAN 模型

由于模型的目标是使得生成的 CNN 特征能用来训练判别器网络，而 f-WGAN 不能保证生成的 CNN 特征恰好符合需求。于是不妨通过鼓励生成器构造可以由判别器网络正确分类的要素（在输入数据上进行训练）来缓解此问题。为此，本文在 f-WGAN 基础上进行改进，建立 f-CLWGAN 模型以使生成特征的分类损失降至最低，使用负对数来定义损失函数

$$\mathcal{L}_{CLS} = -E_{\widetilde{x}\sim p_{\widetilde{x}}}\left[\log P\left(y|\widetilde{x};\theta\right)\right], \tag{2-6}$$

其中 $\widetilde{x} = G(z, c(y))$，$y$ 是 $\widetilde{x}$ 的类别标签，$P\left(y|\widetilde{x};\theta\right)$ 表示 $\widetilde{x}$ 被预测为其真实类别标签 y 的概率。该条件概率是通过由 θ 参数化的线性 Softmax 分类器计算得出的，该分类器已在可见类的真实样本特征上进行过预训练，然后再对测试样本进行计算。分类损失可以看作是强制生成器构造判别式特征的正则化网络，通过该网络对优化的参数进行进一步的学习和判定，并使其能够对输入的图像进行真实图像和虚

假图像的判定，则最终模型的整体优化目标变为

$$\min_{G}\max_{D}\mathcal{L}_{WGAN}+\beta\mathcal{L}_{CL}, \tag{2-7}$$

其中，β 是对分类器进行加权的超参数。

2.3.2 样本分类

给定任何未知类 $u\in\mathcal{Y}^{\mathcal{U}}$ 中的 $c(u)$，通过对噪声 z 重采样可以生成多个 CNN 特征 $\widetilde{x}$，其中 $\widetilde{x}=G(z,c(u))$。对每个看不见的类重复此特征生成过程后，即可获得综合训练集 $\widetilde{\mathcal{U}}=\{(\widetilde{x},u,c(u))\}$。然后通过训练多模式嵌入模型或 Softmax 分类器来训练学习分类网络。该模型生成的特征允许对那些实际看到的类别数据 $\mathcal{S}$ 和生成的不可见的类别数据 $\widetilde{\mathcal{U}}$ 都进行训练。

（1）多模式嵌入模型

在 ALE[17]、DEVISE[18]、SJE[19] 以及 LATEM[20] 等许多零样本学习方法中，人们都将可见类数据 $\mathcal{S}$ 对应的图像特征空间 $\mathcal{X}$ 和类嵌入空间 $\mathcal{C}$ 之间的进行多峰嵌入。而该模型的功能为将这些方法结合可见类数据 $\mathcal{S}$ 和不可见类数据 $\mathcal{U}$ 一起训练，从而学习得到更强大的分类器。由 W 参数化的嵌入模型 $F(x,c(y);W)$ 进行测量任一图像特征 x 与类嵌入 $c(y)$ 之间的兼容性得分。给定图像特征 x，分类器将通过以下方式搜索具有最高兼容性的嵌入类：

$$f(x)=\operatorname*{argmax}_{\mathrm{y}}\mathrm{F}(\mathrm{x},\mathrm{c}(\mathrm{y});\mathrm{W}), \tag{2-8}$$

在 ZSL 中，$y\in\mathcal{Y}^{\mathcal{U}}$，而在 GZSL 中，$y\in\mathcal{Y}^{\mathcal{S}}\cup\mathcal{Y}^{\mathcal{U}}$。

（2）Softmax 分类器

利用标准的 Softmax 分类器将对数似然损失减到最小，即

$$\min_{\theta}-\frac{1}{|T|}\sum_{(x,y)\in T}\log P(y|x;\theta), \tag{2-9}$$

其中 $\theta\in\mathbb{R}^{d_x\times N}$，为将图像特征 x 映射到 N 个不正规概率全连接层的权重矩阵，其

中 N 是类别的数目，且

$$P\left(y|x;\theta\right)=\frac{\sum_{i}^{N}\exp(\theta_{y}^{\mathcal{T}}x)}{\exp(\theta_{i}^{\mathcal{T}}x)}, \tag{2-10}$$

$\mathcal{T}$ 的范围取决于任务，当任务为 ZSL 时，$\mathcal{T}=\widetilde{\mathcal{U}}$，当任务为 GZSL 时，$\mathcal{T}=\mathcal{S}\cup\widetilde{\mathcal{U}}$，则预测函数定义为：

$$f\left(x\right)=\underset{\mathrm{y}}{\mathrm{argmax}}\mathrm{P}\left(\mathrm{y}|\mathrm{x};\theta\right), \tag{2-11}$$

在 ZSL 中，$y\in\mathcal{Y}^{\mathcal{U}}$，而在 GZSL 中，$y\in\mathcal{Y}^{\mathcal{S}}\cup\mathcal{Y}^{\mathcal{U}}$。

2.4　本章小结

本章结合已有的相关研究，对 GAN 和 ZSL 模型分别进行了简单介绍。本章介绍了 GAN 的网络结构，也介绍了 ZSL 中通过语义描述将图像特征映射到高维空间的思想，最终引出将 GAN 应用到 ZSL 情境下生成图像特征的主旨。

在模型搭建的过程介绍中，本章首先搭建初始的 f-GAN 模型，逐步分析模型的缺点并进行提升改进，后续引入 Wasserstein 距离来构建 f-WGAN 模型，并进一步优化得到最终的 f-CLWGAN。在该模型中，条件概率是通过由 θ 参数化的线性 Softmax 分类器计算得出的，该分类器已在可见类的真实样本特征上进行过预训练，在此基础上再对测试样本进行计算。其分类损失则可以看作是强制生成器进行特征生成的正则化网络，通过该网络对优化的参数来进行进一步的学习，并使其能够分辨真实图片和虚假图片。

在样本分类的模块介绍中，本章主要介绍了多模式嵌入模型和 Softmax 分类器，该两者均在 ZSL 和 GZSL 两种情境下进行了测试，前者的功能为，将这些方法结合可见类数据 $\mathcal{S}$ 和不可见类数据 $\mathcal{U}$ 一起训练，从而学习得到更强大的分类器。并使由 W 参数化的嵌入模型 $F(x,c\left(y\right);W)$ 进行测量，并最终输出的为任一图像特征 x 与类嵌入 $c\left(y\right)$ 之间的兼容性得分。Softmax 的功能则相对简单，即通过降低对数似然损失来优化网络。

综上，本章介绍了一种基于 GAN 的 ZSL 模型，其特点为能够直接生成图像特征进行训练，目的是解决 ZSL 和 GZSL 两种情境下的问题。在下一章中本文将介绍模型的具体实验的方案与结果。

第三章 基于对抗生成网络的零样本学习实现与测试

本章将首先对 f-CLWGAN 模型进行框架结构的搭建，描述框架中各模块的功能含义及实现方式。接着从数据集选取、特征生成、测试评估协议和实验细节四个方面来对模型的实现过程进行介绍。然后将该模型应用在 ZSL 和 GZSL 两种情境下进行实际数据样本的测试，观察并分析其在不同数据集上的测试结果。为方便下文描述，本文将 f-GAN、f-WGAN 以及 f-CLWGAN 统称为 f-xGAN 系列模型，本章最后将从稳定性、泛化性、CNN 体系结构以及类嵌入等方面在不同数据集上的影响方面来分析 f-xGAN 系列模型。

3.1 模型框架结构搭建

本节将结合上一章中对 f-CLWGAN 模型的描述介绍，对模型框架结构进行搭建，整个框架目的是通过最小化生成特征的分类损失和具有梯度惩罚的 Wasserstein 距离来对网络进行优化。如图3.1所示，对于生成器 G 而言，模型输入的是遵循高

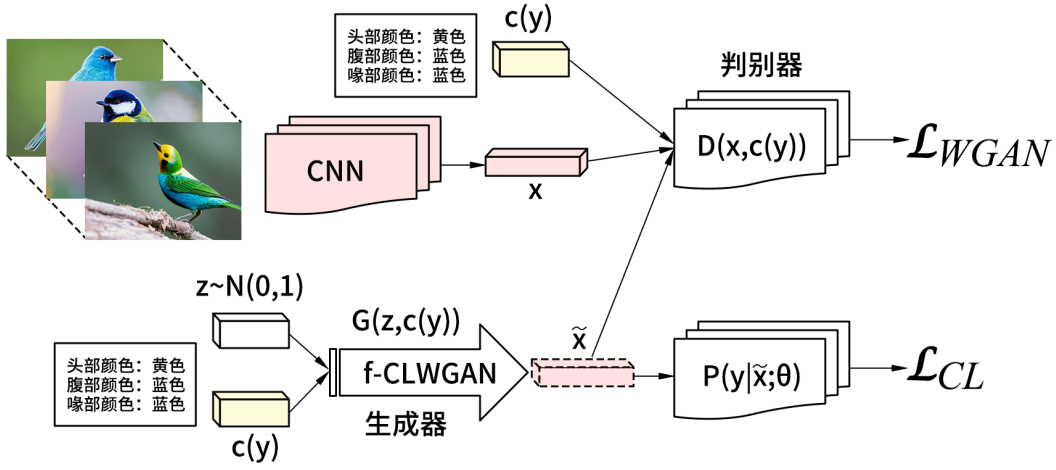

图 3.1 f-CLWGAN 模型框架结构示意图

斯分布的随机噪声 z 和整合的不可见类样本语义描述 $c(y)$，通过其对应映射关系生成 CNN 下的图像特征 $\tilde{x}$，该 $\tilde{x}$ 将通过损失函数 $\mathcal{L}_{CLS}$ 的反馈和鼓励来不断优化，进而生成可以由判别器网络判定为正确分类的特征。而对于判别器 D 而言，其需

要结合输入进的语义描述 $c(y)$ 和真实数据样本的 CNN 图像特征 x 来与生成的虚假图像特征 $\widetilde{x}$ 进行不断分析比对，进而使损失函数 $\mathcal{L}_{WGAN}$ 的值最小, 由此实现基于 GAN 的 ZSL 模型框架结构搭建。

3.2　实验方案的设计与实现

本节将从数据集选取、特征生成、测试评估协议和具体实验细节四个方面来详细介绍实验方案。

（1）数据集选取

Animal-with-Attributes-2（AWA2）[21]、SUN Attribute（SUN）[22] 以及 Caltech-UCSD-Birds 200-2011（CUB）[23] 都是细粒度的数据集。其中 AWA2 包含 50 类动物共 30475 个图像，40 类作为训练集，10 类作为测试集，类别语义为 85 维。CUB 包含来自 200 种不同类型鸟类的 11788 个图像，类别语义为 312 维。SUN 包含来自 717 个场景的 14340 张图像，类别语义为 102 维。

表 3.1　各数据集的统计信息

数据集	属性维数	句子描述	总类别数	训练集 + 验证集数	测试集
AWA2	85	无	50	27+13	10
SUN	102	无	717	580+65	72
CUB	312	有	200	100+50	50

如表3.1所示，各数据集标注出了属性描述的维数、是否有句子描述以及训练集 + 验证集 $\mathcal{Y}^{\mathcal{S}}$ 和测试集 $\mathcal{Y}^{\mathcal{U}}$ 的样本类数。本文使用 AWA2，CUB 和 SUN 三类数据集来分别对搭建好的 f-xGAN 系列模型进行 ZSL 测试。

（2）特征生成

对于真实 CNN 图像特征，本文采用将整个图像从 101 层 ResNet[24] 中提取出 2048 维顶层池化单元的方式，将图像的 CNN 特征提取出来。过程中不进行任何图像预处理（例如裁切）或其他任何数据增强技术。其中 ResNet 已在 ImageNet 1K 上进行了预训练且没有进行微调。而对于生成的虚拟 CNN 图像特征，本文将直接使用 f-xGAN 模型生成 2048 维的 CNN 特征作为输入来训练模型。对于类嵌入，则直接按 AWA2（85 维）、SUN（102 维）和 CUB（312 维）的类属性进行设置。此外，对于有句子描述的 CUB 数据集，本文从细粒度的视觉描述（每个图像 10 个

句子）中提取了基于 CNN-RNN 的 1024 维特征。$\mathcal{Y}^{\mathcal{U}}$ 的句子描述在 CNN-RNN 训练期间不可见，因此模型将通过平均同一类的 CNN-RNN 特征来构建每类句子。

（3）测试评估协议

在 ZSL 情境下的测试中，测试目标是为图像分配一个不可见类标签，即 $\mathcal{Y}^{\mathcal{U}}$。而在 GZSL 情境下，所测试的图像空间将包括可见类和不可见类两者，即 $\mathcal{Y}^{\mathcal{S}} \cup \mathcal{Y}^{\mathcal{U}}$。本文将使用统一评估协议来对两种情境下的测试结果进行评估，在 ZSL 情境下平均精度是针对每个类别独立计算的，然后再将其累计总和除以类别数，即衡量的是排名第一的类别与实际结果相符的准确率（T1）。而在 GZSL 情境中，本文计算表示为 s 的可见类 $\mathcal{Y}^{\mathcal{S}}$ 的平均每类 T1 准确率，表示为 u 的不可见类 $\mathcal{Y}^{\mathcal{U}}$ 的平均每类 T1 准确率及其调和平均数 H，即 $H = 2 * (s * u)/(s + u)$。

（4）实验细节

在所搭建的 f-xGAN 模型中，生成器和判别器都是具有 LeakyReLU 激活的多层感知网络。其中生成器 G 由具有 4096 个隐藏单元的单个隐藏层组成，因为模型旨在学习 ResNet-101 的最大合并单元，因此该生成器网络的输出层为单层 ReLU 层。为了使 GAN 训练更加稳定，f-GAN 的判别器 D 为一个具有 1024 个隐藏单元的隐藏层，而 f-WGAN 和 f-CLWGAN 的判别器 D 则均为一个具有 4096 个隐藏单元的隐藏层。输入生成器网络的随机噪声 z 均服从高斯分布，其维数与类嵌入的维数相同。实验最终设置 $\lambda = 10$，$\beta = 0.01$ 进行测试和分析。

3.3 训练及预测结果

本节将利用搭建好的网络 ZSL 和 GZSL 两种情境下，对常见的数据集分别进行测试和分析。第一组实验设置为在 AWA2、SUN 和 CUB 三类数据集的 ZSL 和 GZSL 情境下评估 f-CLWGAN 的性能，并采用不同的分类器来进行对比分析。第二组实验则将搭建的 f-xGAN 系列模型在 6 种不同分类器下进行了 ZSL 与 GZSL 的对比分析。

（1）ZSL 情境下的 f-CLWGAN 模型测试

表3.2展示了 ZSL 情境下的 f-CLWGAN 在不同分类器设置中的 T1 准确率，在此测试期间所进行判断的样本均来自于不可见类 $\mathcal{Y}^{\mathcal{U}}$。在所有情况下，f-CLWGAN 都可以改善无特征生成情况下的识别准确度。在各数据集上的提升都较为显著，其中 AWA2 的整体准确性由 65.5% 提升至 68.7%，SUN 的整体准确性由 56.6% 提升

至 61.9%，CUB 的整体准确性由 54.4% 提升至 61.3%。另一方面，观察到结果显示该特征生成网络适用于所选的全部多模式嵌入模型和 Softmax 分类器。这在一定程度上表明，本文用以生成不可见类视觉特征的 f-CLWGAN 模型具有较好的通用性。

表 3.2　ZSL 情境下不同分类器的 f-CLWGAN 测试结果

分类器	特征生成	AWA2	SUN	CUB
ALE	无	59.9	54.9	54.9
	f-CLWGAN	68.2	61.9	61.3
SJE	无	65.5	53.7	53.9
	f-CLWGAN	66.9	56.5	58.4
DEVISE	无	54.2	56.5	52.0
	f-CLWGAN	66.8	60.9	60.3
LATEM	无	55.1	55.3	49.3
	f-CLWGAN	68.7	61.2	60.8
Softmax	无	—	—	—
	f-CLWGAN	68.2	60.8	57.3

（2）GZSL 情境下的 f-CLWGAN 模型测试

在 GZSL 情境下模型测试所关联的样本空间包含可见类和不可见类两者，即 $\mathcal{Y}^{\mathcal{S}} \cup \mathcal{Y}^{\mathcal{U}}$。因此，本文计算表示为 s 的可见类 $\mathcal{Y}^{\mathcal{S}}$ 的平均每类 T1 准确性，表示为 u 的不可见类 $\mathcal{Y}^{\mathcal{U}}$ 的平均每类 T1 准确性及其调和平均数 H，即 $H = 2 * (s * u)/(s + u)$。

表 3.3　GZSL 情境下不同分类器的 f-CLWGAN 测试结果

分类器	特征生成	AWA2			SUN			CUB		
		u	s	H	u	s	H	u	s	H
ALE	无	16.8	76.1	26.8	21.8	33.1	25.4	23.7	62.8	31.4
	f-CLWGAN	47.6	57.2	52.0	41.3	31.1	35.5	40.2	59.3	47.9
SJE	无	11.3	74.6	19.6	14.7	30.5	19.8	11.3	74.6	19.6
	f-CLWGAN	37.9	70.1	49.2	36.7	25.0	29.7	37.9	70.1	49.2
DEVISE	无	13.4	68.7	22.4	16.9	27.4	20.9	23.8	53.0	32.5
	f-CLWGAN	35.0	62.8	45.0	38.4	25.4	30.6	52.2	42.4	46.7
LATEM	无	7.9	71.7	13.3	14.7	28.8	19.5	15.2	57.3	24.0
	f-CLWGAN	33.0	61.5	43.0	42.4	23.1	29.9	53.6	39.2	45.3
Softmax	无	—	—	—	—	—	—	—	—	—
	f-CLWGAN	57.9	61.4	57.5	42.6	36.6	38.8	43.7	57.7	48.9

如表3.3所示，对于所选取的全部数据集，本文采用的 f-CLWGAN 均显着改善了 H 度量。其中 AWA2 的整体准确性由 26.8% 提升至 57.5%，SUN 的整体准确性

由 25.4% 提升至 38.8%，CUB 的整体准确性由 32.5% 提升至 48.9%。准确性的提高可以归因于 f-CLWGAN 模型学习到了不可见类中样本的 CNN 特征，尽管模型没有见过这些类的任何实际 CNN 特征，但仍然可以生成足以逼真的特征来帮助模型训练。

此外，从表3.3中还可观察到，在所有模型上都没有生成特征时，可见类精度明显高于不可见类精度，这表明许多样本在测试时被错误地分配给了可见类。通过 f-CLWGAN 生成的特征则可以提高不可见类的准确性，同时也保持可见类的准确性，从而在可见类和不可见类的准确性之间找到平衡。需要注意的是，在直接生成 CNN 特征后，模型仅适用简单的 Softmax 分类器便取得了相较于其他分类器更优的结果，这在一定程度上表明了特征生成对各项识别任务的可推广性。

（3）ZSL 与 GZSL 情境下 f-xGAN 系列模型测试

为了评估模型框架的重要组成部分的影响，本文对 f-xGAN 系列生成模型也进行了测试评估。本文选择 SUN 与 CUB 两个细粒度数据集来对不同版本的生成模型在 ZSL 和 GZSL 两种情境下进行了测试。

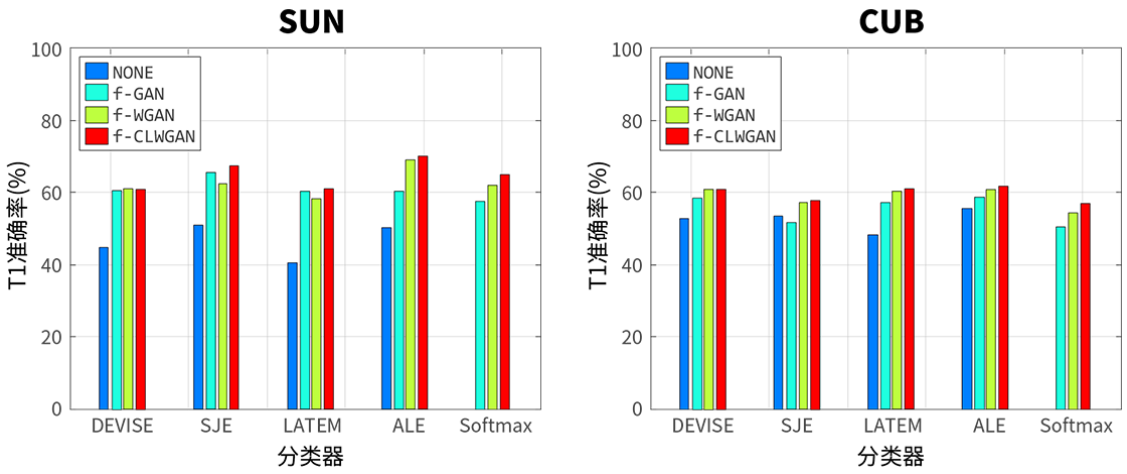

图 3.2　f-xGAN 系列模型在 ZSL 情境下的测试对比结果

如图3.2与图3.3所示，本文选择了 SUN 和 CUB 这两类具有代表性的细粒度数据集进行测试，并以条形统计图的形式展示了 ZSL 情境下的 T1 准确率和 GZSL 情境下的调和平均数 H。首先可以观察到，对于 ZSL 和 GZSL 两种情境，所有生成模型在所有情况下相对无生成 CNN 特征的模型都有了改进。尤其在 GZSL 情境中，无生成模型与 f-xGAN 系列模型之间的测试结果差异非常显着。第二个观察结果则为不同生成模型之间的横向对比分析，对于该两种数据集而言，几乎在所有

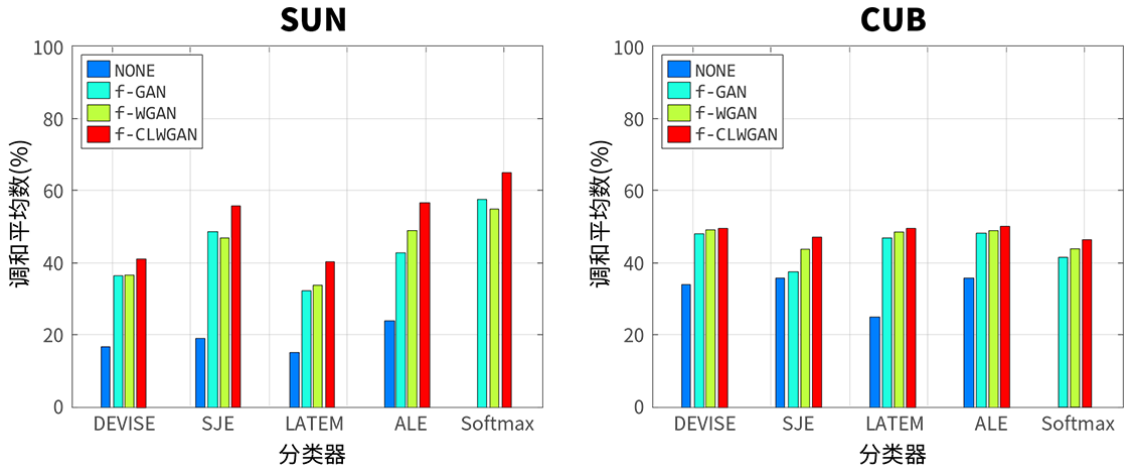

图 3.3　f-xGAN 系列模型在 GZSL 情境下的测试对比结果

情况下，f-CLWGAN 模型都是性能最佳的生成模型。由此可以得出结论，当缺少训练样本数据时，生成 CNN 特征以支持分类器进行学习训练是一种实用的解决方案。

3.4　算法的对比分析与评价

在本节中，我们将综合考虑稳定性、泛化性、用于提取实际 CNN 特征的 CNN 体系结构以及类嵌入这几大方面，选择 SUN 和 CUB 两个细粒度数据集来分析其对 f-xGAN 系列模型的影响。

（1）稳定性和泛化性

我们首先分析不同的生成模型对于可见类样本数据的拟合情况，使用 Softmax 来对可见类的样本特征进行学习，并在测试集上测试预测分类的准确性。如图3.4所示，显示了分类准确率随训练迭代次数增加的变化情况。模型在选定的 SUN 和 CUB 两个细粒度数据集上都展示了逐渐稳定的训练趋势。

由图3.4可见，与通过真实图像（RealData）获得的监督式分类精度（即虚线所标记的上限值）相比，即使在收敛之后，f-GAN 模型的预测结果仍然很弱，这表明 f-GAN 存在欠匹配问题。而 f-WGAN 和 f-CLWGAN 则相对有所改善，几乎达到了监督式分类精度的上限值。在确定了 f-xGAN 系列模型可以带来稳定的训练效果并生成高度描述性的特征后，本文对 f-xGAN 系列模型针对不可见类样本的推广预测能力进行了评估。首先用模型生成不可见类的 CNN 特征。然后使用不可见类的这些 CNN 合成特征与可见类的真实 CNN 特征来共同训练 Softmax 分类器。

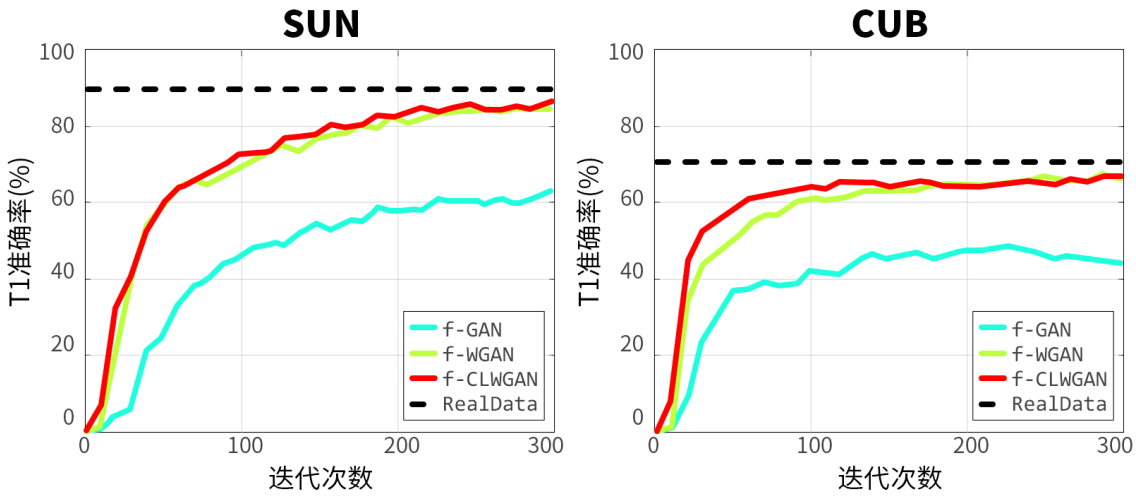

图 3.4 ZSL 情境下模型测试准确率结果

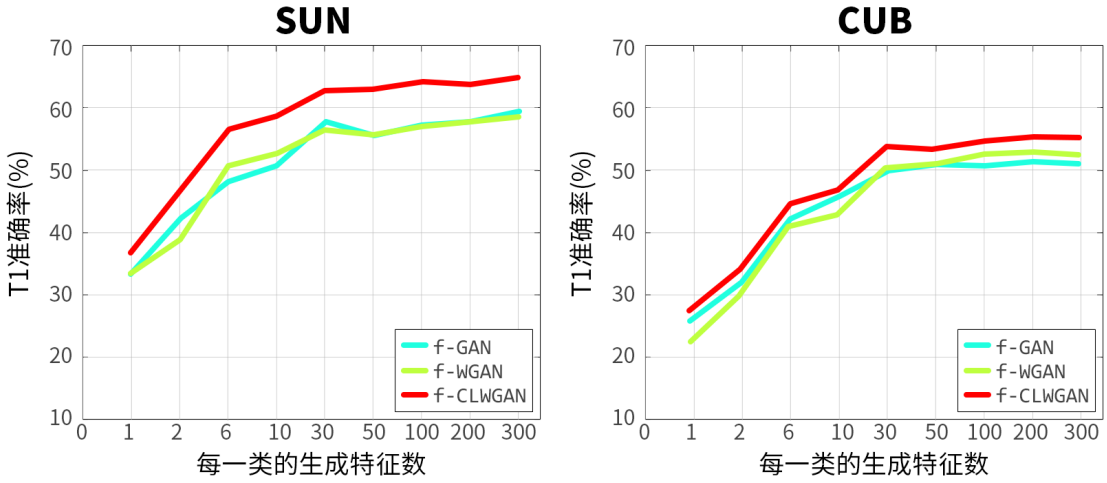

图 3.5 ZSL 情境下模型准确率随生成特征数量的变化情况

如图3.5所示，在 ZSL 情境下，将不可见类的生成特征数量从 1 增加到 100 会使模型准确率的显着提高，例如 SUN 从 35.5% 提升至了 65.4%，CUB 从 29.5% 提升至了 55.8%。与生成可见类特征的情况一样，此处的模型在 SUN 和 CUB 两个数据集上的准确率排序均为 f-GAN < f-WGAN < f-CLWGAN。根据这些结果可以得出，f-xGAN 系列生成模型可以很好地推广到生成不可见类的数据分布问题中，例如可以被广泛使用在 GZSL 情境下的识别任务中。

（2）CNN 架构的影响

该项测试的目的是为了确定 f-xGAN 系列生成模型中所采用的 CNN 编码器对识别精度的影响。如表3.4所示，可以观察到使用 GoogLeNet 特征进行训练时，结果要比使用 ResNet 特征获得的结果要低。这表明 ResNet 特征比预期的 GoogLeNet 更适合 f-xGAN 系列模型训练。而在两种 CNN 架构中，都可以看出添加 f-CLWGAN

生成模型的网络性能均优于无特征生成模型的网络。

表 3.4　GZSL 情境下 f-xGAN 系列模型采用不同 CNN 编码器的测试结果

CNN 架构	特征生成	u	s	H
GoogLeNet	无	20.2	34.8	25.8
	f-CLWGAN	35.3	38.6	36.9
ResNet-101	无	23.5	61.8	34.4
	f-CLWGAN	43.5	57.7	49.7

具体来看，GoogLeNet 特征下的模型准确率从 25.8% 提高到了 36.9%，ResNet 特征下的模型准确率从 34.4% 提高到了 49.7%。这证明了本文的 f-xGAN 系列模型不仅限于学习 ResNet 特征的分布，同时也可以学习其他的 CNN 特征分布。

（3）类嵌入的效果

条件变量，即类嵌入，也是本文 f-xGAN 系列模型的重要组成部分。由于 CUB 数据集包含两类嵌入数据，因此本文评估了 CUB 上每一类属性描述和每一类句子描述这两种不同的类嵌入效果。

表 3.5　GZSL 情境下 f-xGAN 系列模型采用不同 CNN 编码器的测试结果

类嵌入类型	特征生成	u	s	H
属性描述	无	23.7	62.8	34.2
	f-CLWGAN	42.3	56.6	49.5
句子描述	无	38.8	53.8	45.1
	f-CLWGAN	50.3	58.3	54.0

如表3.5所示，由属性描述生成的 f-CLWGAN 特征不仅使最终测试结果显着提高（从 34.2% 提高到了 49.5%），同时也使得 s 和 u 的结果相比不加生成模型而言更加平衡（从 23.7% 比 62.8% 变为了 42.3% 比 56.6%）。这是因为生成特征的 CNN 结构可帮助网络探索看不可见类的样本空间，而其它方法则仅仅是分析可见类的样本特征，因此会更可能将图像判定到可见类中去。最后，由句子描述生成的 f-CLWGAN 特征在训练和测试准确率相比属性描述的情况而言变高了。这是由于句子描述所提供的高度具体描述语义内容使模型生成了更高质量的特征，从而提升了最终的预测准确率。并且该结果表明，f-CLWGAN 模型可在更高质量的描述下可以学习生成更高质量的 CNN 特征。

3.5 本章小结

本章详细介绍了基于 GAN 的 ZSL 模型搭建过程，首先对 f-GAN、f-WGAN 以及 f-CLWGAN 统称为的 f-xGAN 系列模型进行逐步的框架结构搭建与优化，并最终展示了 f-CLWGAN 的结构示意，描述了框架中各模块的功能含义及实现方式。接着本文对模型进行了测试，分别从数据集选取、特征生成、测试评估协议和实验细节四个方面来进行介绍，并在 ZSL 和 GZSL 两种情境下利用 AWA2、SUN 和 CUB 三个细粒度数据集上进行对比测试，观察并分析其在不同数据集上的测试结果。最后，本文从稳定性、泛化性、CNN 体系结构以及类嵌入等方面在不同数据集上的影响来分析 f-xGAN 系列模型，得出了添加特征生成框架的 ZSL 方法能够显著提升预测结果准确率，并具有较好的泛化性和实用性，同时其生成的 CNN 特征质量也与类嵌入的描述精度有关，精确的语义描述能够提升模型的准确率。

第四章　总结与展望

4.1　课题总结

为了改善语义空间和图像空间的对齐关系，提升零样本学习的识别效率，本文基于 GAN 建立了用以生成特征的 f-xGAN 系列图像生成模型，其中 f-CLWGAN 模型通过生成器和判别器之间的相互对抗训练来使生成的图像特征用于训练网络，并通过优化分类损失正则化的 Wasserstein 距离进行语义空间和图像空间的对齐，从而合成未见类别的图像特征，最终提升物体识别的效率和精度。

本文在 ZSL 和 GZSL 两种情境下使用 AWA2、SUN、CUB 这三类细粒度数据集分别进行了实验和测试，并对比分析该特征生成模型在零样本学习问题下的稳定性、泛化性、CNN 体系结构以及类嵌入等方面的影响。并最终得出了添加特征生成框架的 ZSL 方法能够显著提升预测结果准确率的结论，并证明其具有较好的泛化性和实用性。同时其生成的 CNN 特征质量也与类嵌入的描述精度有关，且在不同数据集上采用不同分类器方法后的测试结果也整体一致。

4.2　研究展望

本课题研究的工作主要集中在特征生成模型的搭建和改进上，在整个算法的框架结构上对生成 CNN 特征的 f-xGAN 系列模型进行了较深入的思考，并最终得到了在不同数据集上均较好的实验结果。然而所用的数据集是常见的识别算法训练数据集，数据量庞大且语义种类丰富，尽管其中包含较多不可见类数据，但该方法相比于实际人类的学习情况依然有所差距。即实际生活中人类并不需要通过成千上百的数据样本来学习一个新事物，且回到文章最初涉及的问题，少数已濒危的物种确实存在其自身性状就是独有的情况，对于该类问题，或类似已灭绝动物的识别问题，零样本学习真正要做到不需要样本实现类别判定仍有一段距离。尤其在对样本进行特征提取的效率和低维到高维空间的映射关系上有可以继续改进的空间。希望在今后的相关研究中能对此做更多的探索。

致 谢

随着毕业设计论文最终章节的落笔,四年的本科生活也即将划上一个圆满的句号。回顾过去，大学让我收获了许多许多。初入大学时的懵懂无知，到此刻临近毕业时的成熟冷静，我回忆起四年的时光，仿佛一切都还历历在目。一路走来，曾遇到过的每一个人都浮现在我的眼帘，团结友爱的同学们，孜孜不倦的老师们，还有那些爱我的和我爱的朋友家人们，他们都让我倍感珍惜，我真心地感谢他们！

首先我最想感谢的是我的毕业设计指导老师，邓成教授。邓教授在我的整个毕业设计的过程中都给了我许多帮助。从最初的课题选择及确定，到任务计划和工作安排的调整，再到中期答辩的讨论、初稿的确定以及最终的成稿，我在其中遇到了很多困难，邓教授一直都在耐心地指导我，教我如何用更好地方法解决问题，教我考虑整体、注意细节，并指导我最终完成毕设，非常感谢邓教授的帮助！

其次我想感谢的是实验室里的师兄师姐，他们在我遇到困难和感到疑惑的时候，及时地为我指明了方向，他们非常乐意向我传授经验，分享知识，帮助我解决了很多疑惑。感谢师兄师姐们的帮助！

同时非常感谢大学里各位同学和朋友的鼓励与帮助，我们相处的时间说长不长，说短不短，我们见证了彼此的青春，也共同经历了苦辣酸甜。你们在我需要的时候给了我毫不犹豫的帮助，如果没有你们，我无法变成今天这个成熟自信的我。非常感谢你们对我的理解和包容，你们在我失意时鼓励着我，在我伤心的时候安慰我，遇见你们，是我一生的幸运。

我还要诚挚感谢的是我的父母，是他们一直在背后给我鼓励，并且无条件支持我做的决定，这让我也更加有动力来面对自己的学习和生活。在以后的研究生学习生涯里，我会更加努力地学习，争取创造出属于自己的成果，让父母为我骄傲。

最后，我也要感谢本论文所引用文章的各位作者，如果没有这些学者的研究成果的启发和帮助，我将无法完成本篇论文的最终写作。金无足赤，人无完人。由于我的学术水平有限，所写论文难免有不足之处，恳请各位老师批评和指正！

参考文献

[1] 刘海玲. 基于计算机视觉算法的图像处理技术[J]. 计算机与数字工程, 2019, 47(3):672-677.

[2] 尹宝才, 王文通, 王立春. 深度学习研究综述[J]. 北京工业大学学报, 2015, 41(1):48-59.

[3] 余凯, 贾磊, 陈雨强, 等. 深度学习的昨天, 今天和明天[J]. 计算机研究与发展, 2013, 50(9): 1799-1804.

[4] 冀中, 汪浩然, 于云龙, 等. 零样本图像分类综述: 十年进展[J]. 中国科学: 信息科学, 2019 (10):5.

[5] 周翔. 基于深度生成模型的零样本学习[D]. 电子科技大学, 2019.

[6] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//Advances in neural information processing systems. 2014: 2672-2680.

[7] 王坤峰, 苟超, 段艳杰, 等. 生成式对抗网络 GAN 的研究进展与展望[J]. 自动化学报, 2017, 43(3):321-332.

[8] 刘恋秋. 基于深度卷积生成对抗网络的图像识别算法[J]. 液晶与显示, 2020, 35(4):383-388.

[9] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv preprint arXiv:1511.06434, 2015.

[10] Chen X, Duan Y, Houthooft R, et al. Infogan: Interpretable representation learning by information maximizing generative adversarial nets[C]//Advances in neural information processing systems. 2016: 2172-2180.

[11] Mirza M, Osindero S. Conditional generative adversarial nets[J]. arXiv preprint arXiv:1411.1784, 2014.

[12] Arjovsky M, Chintala S, Bottou L. Wasserstein gan[J]. arXiv preprint arXiv:1701.07875, 2017.

[13] Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of wasserstein gans[C]//Advances in neural information processing systems. 2017: 5767-5777.

[14] 魏杰. 零样本学习中的细粒度图像分类研究[D]. 广东工业大学, 2019.

[15] 王阳. 零样本学习方法及其应用研究[D]. 南京理工大学, 2019.

[16] Arjovsky M, Bottou L. Towards principled methods for training generative adversarial networks. arxiv e-prints, art[J]. arXiv preprint arXiv:1701.04862, 2017.

[17] Akata Z, Perronnin F, Harchaoui Z, et al. Label-embedding for image classification[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 38(7):1425-1438.

[18] Frome A, Corrado G S, Shlens J, et al. Devise: A deep visual-semantic embedding model[C]// Advances in neural information processing systems. 2013: 2121-2129.

[19] Akata Z, Reed S, Walter D, et al. Evaluation of output embeddings for fine-grained image classification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 2927-2936.

[20] Xian Y, Akata Z, Sharma G, et al. Latent embeddings for zero-shot classification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 69-77.

[21] Lampert C H, Nickisch H, Harmeling S. Attribute-based classification for zero-shot visual object categorization[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 36(3): 453-465.

[22] Patterson G, Hays J. Sun attribute database: Discovering, annotating, and recognizing scene attributes[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012: 2751-2758.

[23] Welinder P, Branson S, Mita T, et al. Caltech-ucsd birds 200[J]. 2010.

[24] 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.