

分类号\_\_\_\_\_  
学校代码 10487

学号 D201980609  
密级\_\_\_\_\_

华中科技大学  
博士学位论文

(学术型  专业型 )

基于深度表征的零样本  
图像分类研究

学位申请人： 陈使明

学科专业： 信息与通信工程

指导教师： 尤新革 教授

答辩日期： 2022年11月22日

## 答辩委员会

	姓名	职称	单位
主席	陶大程	教授	悉尼大学/京东探索研究院
委员	赖剑煌	教授	中山大学
	杜博	教授	武汉大学
	白翔	教授	华中科技大学
	徐明亮	教授	郑州大学
	田岩	教授	华中科技大学
	杨欣	教授	华中科技大学

**A Dissertation Submitted in Partial Fulfillment of the Requirements  
for the Degree of Doctor of Engineering**

**Research on Zero-Shot Image Classification  
Based on Deep Feature Representation**

**Ph.D. Candidate : Shiming Chen**

**Major : Information and Communication Engineering**

**Supervisor : Xinge You**

**Huazhong University of Science and Technology**

**Wuhan 430074, P. R. China**

**November, 2022**

## 独创性声明

本人声明所呈交的学位论文是我个人在导师的指导下进行的研究工作及取得的研究成果。尽我所知，除文中已标明引用的内容外，本论文不包含任何其他人或集体已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：

日期： 年 月 日

## 学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，即：学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权华中科技大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

保 密 ，在 \_\_\_\_\_ 年解密后适用本授权书。

本论文属于

不保密 。

(请在以上方框内打“√”)

学位论文作者签名：

指导教师签名：

日期： 年 月 日

日期： 年 月 日

## 摘 要

图像分类是计算机视觉中的一个基本任务。近年来，利用深度学习进行视觉特征表示（即深度表征）的图像分类技术取得了重大进展。然而，这种传统的图像分类算法只能对已知类别的图像进行分类，不能对现实生活中成千上万且持续增长的未知类目标进行识别。在属性等语义信息的指导下，零样本图像分类通过从已知类到未知类的知识迁移实现对未知类图像的分类，已成为当前人工智能领域的研究热点。为挖掘基于深度表征的零样本图像分类的潜力，本文针对跨数据集偏差、视觉-语义特征的表示差异性、视觉-语义特征的异构性等三个问题开展研究。本文的主要研究内容和创新点如下：

(1) 针对基于深度表征的零样本图像分类面临的跨数据集偏差问题，本文面向嵌入式和生成式模型提出基于视觉特征增强的零样本图像分类方法。面向嵌入式零样本图像分类，本文利用图指导双注意力网络融合局部视觉特征和显式全局视觉特征对视觉特征进行增强。面向生成式零样本图像分类，本文利用特征精细化学习机制实现视觉特征增强，并使得生成器生成更真实的未知类伪特征样本。实验结果表明，本文提出的方法能够有效地增强视觉特征的判别性和迁移性，实现有效的视觉-语义交互，提高了零样本图像分类效果。

(2) 针对基于深度表征的零样本图像分类中视觉-语义特征的表示差异性问题，本文提出了基于属性-视觉关键公共语义知识的零样本图像分类方法。该方法基于属性指导的 Transformer 网络，利用单向跨注意力机制学习具有属性定位的视觉特征，对属性-视觉特征之间的潜在公共语义知识进行准确表示。该方法基于互语义蒸馏网络，利用双向注意力模块学习基于属性的视觉特征和基于视觉的属性特征，并在互语义蒸馏学习机制的指导下，充分地挖掘关键语义特征。通过将这两个网络集成为一个统一的模型，提高视觉-语义特征的语义一致性，实现零样本图像分类从已知类到未知类有效的语义知识迁移。

(3) 针对基于深度表征的零样本图像分类中视觉-语义特征的异构性问题，本文提出一种基于层次语义-视觉适应的零样本图像分类方法。不同于现有单步适应的方法只进行分布对准，该方法分别使用结构适应模块和分布适应模块逐步学习一个本

# 华 中 科 技 大 学 博 士 学 位 论 文

---

真的子空间用于视觉和语义特征的表示，实现视觉特征和语义特征在一个公共子空间中真正对准，提升公共子空间式零样本图像分类方法性能。

本文在多个公开数据集上进行了充分实验，实验结果表明本文方法均取得了显著的性能提升，验证了所提出方法的有效性。

**关键词：**零样本学习；零样本图像分类；深度表征；知识迁移；特征增强

## Abstract

Image classification is a fundamental task in computer vision. In recent years, the image classification technology has made great progress using deep learning for visual feature representation (*i.e.*, deep feature representation). However, such conventional image classification can only classify the images of seen classes, failing to recognize thousands of unknown objects. Under the guidance of semantic information (*e.g.*, attributes), zero-shot image classification (ZSIC) can classify images of the unseen class by transferring knowledges from seen classes, and it has become a popular research in artificial intelligence. To release the potential of the deep feature representations based ZSIC, this thesis focuses on tackling the challenges of: (1) cross-dataset bias; (2) the inconsistency between visual-semantic feature representations; and (3) the heterogeneous feature alignment between vision-semantic. The main contributions and innovations of this thesis are as follows:

(1) To tackle the challenge of cross-dataset bias in deep feature representations based ZSIC, this thesis proposes two visual feature enhancement models for the embedding-based and generative ZSIC methods. As for the embedding-based ZSIC, a graph-guided dual attention network is introduced to fuse the local visual features and explicit global visual features to enhance visual features. As for the generative ZSIC, a feature refinement learning mechanism is proposed to enhance the visual features and encourage the generator to synthesize realistic visual features for unseen classes. The experimental results show that the proposed methods effectively improve the discrimination and transferability of visual features, which enables ZSIC to conduct effective interactions between visual-semantic features and achieve significant performance gains.

(2) To tackle the challenge of the inconsistency between semantic-visual feature representations in deep feature representations based ZSIC, this thesis proposes a key common semantic knowledges between visual-attribute features based ZSIC methods. First, an attribute-guided Transformer network employs the cross-attention to learn the visual features with accurate attribute localization to represent the key common semantic knowledges.

# 华 中 科 技 大 学 博 士 学 位 论 文

---

Then, a mutually semantic distillation network takes bidirectional attention sub-nets to learn attribute-based visual features and visual-based attribute features. Under the guidance of mutually semantic distillation learning, the two sub-nets learn consistent semantic features. Finally, the two networks are integrated into an unified framework to fully and exactly discover the key common semantic knowledges between visual-attribute features, which improves the semantic consistency between visual-semantic features. As such, our method conducts effective semantic knowledge transfer from seen classes to unseen ones for ZSIC.

(3) To tackle the challenge of the heterogeneous feature alignment between vision-semantic in deep feature representations based ZSIC, this thesis proposes a hierarchical semantic-visual adaptation based ZSIC. Different to existing one-step adaptation method that on alignment the feature distributions between visual and semantic domains, this method utilizes a hierarchical adaptation to learn an intrinsic common space for semantic and visual feature representations by adopting sequential structure adaptation and distribution adaptation. To this end, the proposed method realizes the real alignment of visual and semantic features to achieve classification performance gains for common space learning based ZSIC.

This thesis conducts the extensive experiments to demonstrate the effectiveness of the proposed methods, which lead the state-of-the-art performance on several popular benchmark datasets.

**Keywords:** Zero-shot learning, Zero-shot image classification, Deep feature representations, Knowledge transfer, Feature enhancement

## 目 录

<b>摘 要</b> .....	I
<b>Abstract</b> .....	III
<b>中英文缩写对照表</b>	
<b>1 绪论</b>	
1.1 研究背景与意义.....	(1)
1.2 研究内容与主要贡献.....	(3)
1.2.1 基于视觉特征增强的零样本图像分类.....	(4)
1.2.2 基于属性-视觉关键公共语义知识的零样本图像分类 .....	(5)
1.2.3 基于层次语义-视觉适应的零样本图像分类 .....	(6)
1.3 论文章节安排与组织结构 .....	(6)
<b>2 零样本图像分类的研究概述</b>	
2.1 国内外相关研究工作综述 .....	(8)
2.1.1 嵌入式零样本图像分类 .....	(9)
2.1.2 生成式零样本图像分类 .....	(11)
2.1.3 公共子空间式零样本图像分类.....	(12)
2.2 现有工作的不足及存在的主要问题 .....	(13)
2.3 零样本图像分类数据集及评价指标 .....	(14)
2.3.1 零样本图像分类数据集 .....	(14)
2.3.2 零样本图像分类评价指标 .....	(16)
2.4 本章小结 .....	(16)
<b>3 基于视觉特征增强的零样本图像分类</b>	
3.1 引言 .....	(18)
3.2 基于图指导双注意力网络的嵌入式零样本图像分类 .....	(20)
3.2.1 研究动机 .....	(20)
3.2.2 基于图指导的双注意力网络 .....	(21)
3.2.3 实验结果与分析 .....	(26)

# 华 中 科 技 大 学 博 士 学 位 论 文

---

3.3	基于视觉特征精细化的生成式零样本图像分类 .....	(34)
3.3.1	研究动机 .....	(34)
3.3.2	视觉特征精细化学习 .....	(36)
3.3.3	实验结果与分析 .....	(41)
3.4	本章小结 .....	(50)
<b>4</b>	<b>基于属性-视觉关键公共语义知识的零样本图像分类</b>	
4.1	引言 .....	(52)
4.2	基于属性指导 Transformer 的零样本图像分类 .....	(53)
4.2.1	研究动机 .....	(53)
4.2.2	属性指导的 Transformer .....	(54)
4.2.3	实验结果与分析 .....	(58)
4.3	基于互语义蒸馏网络的零样本图像分类 .....	(64)
4.3.1	研究动机 .....	(64)
4.3.2	互语义蒸馏网络 .....	(65)
4.3.3	实验结果与分析 .....	(69)
4.4	联合属性指导 Transformer 和互语义蒸馏网络的零样本图像分类 .....	(75)
4.4.1	研究动机 .....	(75)
4.4.2	基于属性指导 Transformer 的互语义蒸馏网络 .....	(77)
4.4.3	实验结果与分析 .....	(80)
4.5	本章小结 .....	(81)
<b>5</b>	<b>基于层次语义-视觉适应的零样本图像分类</b>	
5.1	引言 .....	(82)
5.2	层次语义-视觉适应网络 .....	(82)
5.3	实验设置 .....	(88)
5.4	超参实验分析 .....	(89)
5.5	消融实验分析 .....	(91)
5.6	定性实验分析 .....	(92)
5.7	HSVA 和其他先进方法的对比 .....	(93)
5.8	本章小结 .....	(95)

# 华 中 科 技 大 学 博 士 学 位 论 文

---

## 6 总结与展望

6.1	论文总结 .....	(96)
6.2	展望 .....	(97)
致 谢 .....	(98)	
参考文献 .....	(99)	
附录 1	答辩委员会决议 .....	(111)
附录 2	攻读博士学位期间取得的研究成果 .....	(112)
附录 3	公开发表的学术论文与博士学位论文的关系 .....	(115)
附录 4	攻读博士学位期间参与的科研项目 .....	(118)

## 插图清单

1.1	传统的图像分类和零样本图像分类的差异性对比。 . . . . .	1
1.2	零样本图像分类应用场景。 . . . . .	2
1.3	本文研究内容示意图。 . . . . .	3
2.1	基于不同视觉-语义交互方式的零样本图像分类。 . . . . .	9
2.2	直接属性预测和间接属性预测模型结构示意图。 . . . . .	10
2.3	CUB、SUN、AWA2 等三个主流数据集部分样本展示图。 . . . . .	15
3.1	跨数据集偏差问题阐述与偏差量定量度量。 . . . . .	19
3.2	不同嵌入式零样本图像分类方法的对比。 . . . . .	20
3.3	本章提出的 GNDAN 模型结构示意图。 . . . . .	22
3.4	RAN 分支上的特征区域块数目 $K$ 对 GNDAN 的影响。 . . . . .	28
3.5	模型的组合系数 $(\alpha_1, \alpha_2)$ 对 GNDAN 的影响。 . . . . .	28
3.6	自校准损失权重 $(\lambda_{SC})$ 对 GNDAN 的影响。 . . . . .	29
3.7	GNDAN 中 RAN 和 RGAT 子网络学习的特征图可视化。 . . . . .	31
3.8	由 GNDAN 的子网络 RAN、RGAT 以及其完整模型在 CUB 数据集上学习的视觉特征 t-SNE <sup>[1]</sup> 可视化。 . . . . .	32
3.9	现有的生成式零样本图像分类方法和本文基于视觉特征精细化的生成式零样本图像分类方法对比。 . . . . .	35
3.10	本章提出的 ViFR 模型结构示意图。 . . . . .	37
3.11	平衡因子 $\gamma$ 对 ViFR 模型的影响。 . . . . .	43
3.12	损失权重 $\lambda_{SAMC}$ 对 ViFR 模型的影响。 . . . . .	44
3.13	损失权重 $\lambda_{Ra}$ 对 ViFR 模型的影响。 . . . . .	44
3.14	未知类伪视觉特征样本的数量 $N_{syn}$ 对 ViFR 模型的影响。 . . . . .	45
3.15	不同特征成分对 ViFR 的影响。 . . . . .	46
3.16	ViFR 和 CNN Backbone 在 CUB 数据集上学习的视觉特征 t-SNE <sup>[1]</sup> 可视化对比。 . . . . .	47

# 华 中 科 技 大 学 博 士 学 位 论 文

---

4.1 零样本图像分类的关键任务阐述：通过挖掘视觉特征和语义特征关键的公共语义知识实现有效地已知类到未知类的语义知识迁移。 . . . . .	53
4.2 现有基于注意力机制的零样本图像分类方法和本章节提出的基于属性指导 Transformer 的零样本图像分类方法的模型对比。 . . . . .	54
4.3 本章提出的 TransZero 模型结构示意图。 . . . . .	55
4.4 标准 Transformer 的自注意力机制和属性指导 Transformer 的跨注意力机制对比。 . . . . .	57
4.5 损失权重 $\lambda_{AR}$ 对 TransZero 模型的影响。 . . . . .	60
4.6 损失权重 $\lambda_{SC}$ 对 TransZero 模型的影响。 . . . . .	60
4.7 现有基于注意力机制方法（例如，AREN <sup>[2]</sup> ）和 TransZero 学习的特征图可视化。 . . . . .	61
4.8 TransZero 和 CNN Backbone 在 CUB 数据集上学习的视觉特征 t-SNE <sup>[1]</sup> 可视化对比。 . . . . .	62
4.9 本章提出的 MSDN 模型结构示意图。 . . . . .	65
4.10 两个注意力子网络的组合系数 $(\alpha_1, \alpha_2)$ 对 MSDN 的影响。 . . . . .	70
4.11 自校准权重 $(\lambda_{SC})$ 对 MSDN 的影响。 . . . . .	71
4.12 语义蒸馏损失权重 $(\lambda_{distill})$ 对 MSDN 的影响。 . . . . .	71
4.13 本章节使用的基准方法（baseline）模型结构示意图。 . . . . .	72
4.14 MSDN 的两个注意力子网络学习的特征图可视化。 . . . . .	73
4.15 MSDN 的不同模型和基准方法在 CUB 数据集上学习的视觉特征 t-SNE <sup>[1]</sup> 可视化。 . . . . .	73
4.16 不同方法学习的特征图可视化对比。 . . . . .	76
4.17 不同细粒度类别的样例。 . . . . .	76
4.18 本章提出的 TransZero++ 模型结构示意图。 . . . . .	78
5.1 基于单步适应的和基于层次语义-视觉适应的公共子空间学习方法。 . . . . .	83
5.2 本章提出的 HSVA 模型结构示意图。 . . . . .	84
5.3 有监督的对抗差异性学习机制。 . . . . .	85
5.4 CZSL 设置下每个未知类生成的伪特征样本个数 $N_u$ 对 HSVA 的影响。 . . . . .	90
5.5 GZSL 设置下每个已知类和未知类生成的伪特征样本个数 $(N_s$ 和 $N_u)$ 对 HSVA 的影响。 . . . . .	91

# 华 中 科 技 大 学 博 士 学 位 论 文

---

5.6	结构对准子空间的特征维度 Dim_S 对 HSVA 的影响。 . . . . .	91
5.7	分布对准子空间的特征维度 Dim_D 对 HSVA 的影响。 . . . . .	92
5.8	HSVA 和 CADA-VAE <sup>[3]</sup> 学习的子空间中视觉嵌入特征（图中的“○”） 和语义嵌入特征（图中的“×”）的 t-SNE 可视化 <sup>[1]</sup> 。 . . . . .	93

## 表格清单

2.1 四个最常用的零样本图像分类属的数据信息统计。 . . . . .	15
3.1 不同模型成分设置下，GNDAN 在 CUB 数据集 <sup>[4]</sup> 和 SUN 数据集 <sup>[5]</sup> 上的性能表现。 . . . . .	29
3.2 GNDAN 使用不同图模型的情况下，其在 CUB 数据集 <sup>[4]</sup> 和 SUN 数据集 <sup>[5]</sup> 上的分类性能表现。 . . . . .	30
3.3 GNDAN 和和其他先进的零样本图像分类方法在 CUB <sup>[4]</sup> 、SUN <sup>[5]</sup> 、AWA2 <sup>[6]</sup> 数据集上的 CZSL/GZSL 实验结果对比。 . . . . .	33
3.4 不同的输入图像分辨率对 ViFR 的影响。 . . . . .	42
3.5 基于不同模型成分的 ViFR 在 CUB 数据集 <sup>[4]</sup> 和 AWA2 数据集 <sup>[6]</sup> 上的实验结果。 . . . . .	46
3.6 Pre-FR 模块集成到 TF-VAEGAN <sup>[7]</sup> 、HSVA <sup>[8]</sup> 等两个生成式零样本图像分类模型在 CUB <sup>[4]</sup> 数据集和 AWA2 <sup>[6]</sup> 数据集上的实验结果。 . . . . .	49
3.7 ViFR 和其他先进的零样本图像分类方法在 CUB <sup>[4]</sup> 、SUN <sup>[5]</sup> 、AWA2 <sup>[6]</sup> 数据集上的 CZSL/GZSL 实验结果对比。 . . . . .	50
4.1 不同模型成分设置下，TransZero 在 CUB <sup>[4]</sup> 和 SUN <sup>[5]</sup> 数据集上的零样本图像分类性能表现。 . . . . .	61
4.2 TransZero 和其他先进的零样本图像分类方法在 CUB <sup>[4]</sup> 、SUN <sup>[5]</sup> 、AWA2 <sup>[6]</sup> 数据集上的 CZSL/GZSL 实验结果对比。 . . . . .	63
4.3 在不同模型成分设置下，MSDN 在 CUB 数据集 <sup>[4]</sup> 和 SUN 数据集 <sup>[5]</sup> 上的实验结果。 . . . . .	72
4.4 MSDN 和其他先进的零样本图像分类方法在 CUB <sup>[4]</sup> 、SUN <sup>[5]</sup> 、AWA2 <sup>[6]</sup> 数据集上的 CZSL/GZSL 实验结果对比。 . . . . .	74
4.5 TransZero++ 和 TransZero（章节 4.2）、MSDN（章节 4.3）以及其他先进的零样本图像分类方法在 CUB <sup>[4]</sup> 、SUN <sup>[5]</sup> 、AWA2 <sup>[6]</sup> 数据集上的 CZSL/GZSL 实验结果对比。 . . . . .	80

# 华 中 科 技 大 学 博 士 学 位 论 文

---

5.1	HSVA 中的网络结构设置细节。 . . . . .	89
5.2	在不同模型成分设置下, HSVA 在 AWA1 <sup>[9]</sup> 、AWA2 <sup>[6]</sup> 、CUB <sup>[4]</sup> 和 SUN <sup>[5]</sup> 数据集上的实验结果。 . . . . .	92
5.3	在 CZSL 设置下, HSVA 和其他先进的公共子空间式零样本图像分类方法在 AWA1 <sup>[9]</sup> 、CUB <sup>[4]</sup> 、SUN <sup>[5]</sup> 数据集上的实验结果对比。 . . . . .	93
5.4	在 GZSL 设置下, HSVA 和其他先进的零样本图像分类方法在 AWA1 <sup>[9]</sup> 、AWA2 <sup>[6]</sup> 、CUB <sup>[4]</sup> 、SUN <sup>[5]</sup> 数据集上的实验结果对比。 . . . . .	94

## 0 中英文缩写对照表

ZSIC	Zero-Shot Image Classification (零样本图像分类)
ZSL	Zero-Shot Learning (零样本学习)
CNN	Convolutional Neural Network (卷积神经网络)
SIFT	Scale Invariant Feature Transform (尺度不变特征变换特征)
SURF	Speeded Up Robust Features (加速鲁棒特征特征)
HOG	Histogram of Oriented Gradients (梯度方向直方图特征)
LSS	Local Self-Similarity (局部自相似性特征)
CZSL	Conventional Zero-Shot Learning (传统的零样本学习)
GZSL	Generalized Zero-Shot Learning (广义的零样本学习)
DAP	Direct Attribute Prediction (直接属性预测模型)
IAP	Indirect Attribute Prediction (和间接属性预测模型)
VAE	Variational Auto-Encoder (变分自编码器)
GAN	Generative Adversarial Network (生成对抗网络)
H	Harmonic Mean (调和均值)
MMD	Maximum Mean Discrepancy (最大平均差异性)
RAN	Region-guided Attention Network (区域指导的注意力网络)
RGAT	Region-guided Graph Attention Network (区域指导的图注意力网络)
MLP	Multi-Layer Perceptron (多层感知机)
Pre-FR	Pre-Feature-Refinement (前置特征细化)
Post-FR	Post-Feature-Refinement (后置特征细化)
AGT	Attribute-Guided Transformer (属性指导 Transformer 网络)
VSEN	Visual-Semantic Embedding Network, (视觉-语义嵌入网络)
AVT	Attribute→Visual Transformer (属性 → 视觉 Tranformer 子网络)
AVT	Visual→Attribute Transformer (视觉 → 属性 Tranformer 子网络)
SA	Structure Adaptation (结构适应)

# 华 中 科 技 大 学 博 士 学 位 论 文

---

DA	Distribution Adaptation (分布适应)
SAD	Supervised Aversarial Dscrepancy Learning (有监督的对抗差异性学习机制)
SWD	Sliced Wasserstein Discrepancy (切片 Wasserstein 差异性)
CORAL	Inverse CORrelation ALignment (反相互关系对准)

# 1 绪论

## 1.1 研究背景与意义

近年来，随着大数据、云计算、人工智能等领域的快速发展及交互融合，智慧电商、智慧医疗、智慧城市等概念越发受到关注。随着人们对更智能、更便捷、更高质量生活的实际需求，同时伴随着重大的研究价值和广阔的商业前景，众多高校、科研机构、政府部门均对相关产业投入了大量的人力、物力和财力。人工智能被全球当做新时代工业革命的引擎，正悄然渗入到各行各业并改变着人们的生活方式。计算机视觉是人工智能领域的重要分支，旨在研究如何让计算机类似于人类视觉系统智能地感知、分析、处理现实世界。以图像和视频为信息载体的各项计算机视觉算法，早已渗透到人们的日常生活中，如人脸识别、辅助驾驶、商品检索、智能监控、视觉导航等。图像分类技术，作为计算机视觉领域中基础的、重要的研究方向之一，一直是研究人员的关注热点。

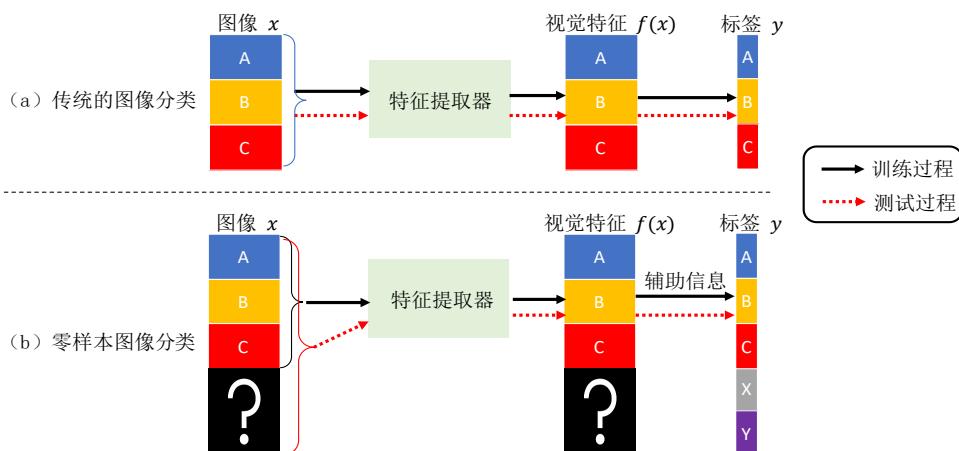


图 1.1 传统的图像分类和零样本图像分类的差异性对比。

得益于收集的大量数据和计算资源，传统的图像分类技术取得了非常大的进步。依托于大量的有标注或无标注的数据，它们利用有监督学习<sup>[10-12]</sup>（Supervised Learning）、半监督学习<sup>[13,14]</sup>（Semi-supervised Learning）、无监督学习<sup>[15-17]</sup>（Unsupervised Learning），通常把深度神经网络模型作为特征提取器对图像的视觉特征进行表示（即深度表征）并用于分类，如图1.1(a)所示。事实上，我们不能对实际场景



图 1.2 零样本图像分类应用场景。

中所有事物都采集相应的数据去训练这样一个图像分类系统，特别是对于稀缺或者新增事物目标数据，例如军事领域、智慧医疗领域的数据。我们当前人工构造的最大图像数据集 ImageNet<sup>[18]</sup> 也包含大约 22,000 个类别，远未达到生物学<sup>[19]</sup> 研究表明的“人类能够识别大约 30,000 个基本的种类和更多的亚类”。此外，现实世界中远不止 22,000 个类别的事物，且不断会有新类别目标产生。为此，传统的图像分类只能对已经在训练集里面出现过的类别（即已知类）进行分类，无法对现实生活中成千上万且持续增长的、在训练集中未出现的类（即未知类）目标进行识别。然而人类的认知系统却大不相同，人类不仅可以从大量的示例样本中学会辨识不同的目标，也能在只提供少量示例或者目标描述的情况下对新目标进行辨别。

针对传统图像分类技术的不足，零样本图像分类（Zero-Shot Image Classification, ZSIC）于 2008 年由 Larochelle 等<sup>[20]</sup> 首次提出并称为“Zero-Data Learning”，并于 2009 年被 Palatucci 等人<sup>[21]</sup> 定义为零样本学习（Zero-Shot Learning, ZSL）且沿用至今。零样本图像分类是指：在属性描述等语义辅助信息的指导下，模型通过视觉-语义交互（即视觉-语义的准确映射关系）从已知类到未知类进行知识迁移，实现对未知类图像的分类<sup>[20-23]</sup>，如图1.1(b) 所示。零样本图像分类是通过模仿人对新事物认知过程而设计的学习模型。例如，同学们对未知类“斑马”的认知过程：假设同学们并不认识斑马，但可以通过老师对斑马的描述“它像一只体型较小的马，同时有老虎和大熊猫一样的黑白条纹”去构想斑马的模样，当同学们见到斑马时可以对这一未知事物快速地识别/认知。我们在现实生活中面临大量零样本场景需求。如图

# 华中科技大学博士学位论文

1.2(a) 所示，我们受限于现实条件而无法获取新物种、濒危物种的视觉数据，传统的图像分类方法不能对这些类别进行识别。如图 1.2(b) 所示，无人超市中会增加新的商品，部署的系统无法对新商品识别并进行销售；如图 1.2(c) 所示，在自动驾驶场景中会遇到很多未知的驾驶场景，自动驾驶系统无法准确地推断新场景并作出正确的决策。零样本图像分类作为前沿技术，其极具挑战性并且是实现人工智能更加智能的关键研究方向，近十年来相关研究呈现逐年上升<sup>[24-30]</sup>。由于深度神经网络对图像提取的视觉特征比人工设计的特征表示能力更强（例如 AlexNet<sup>[11]</sup>, ResNet<sup>[31]</sup>, Transformer<sup>[32]</sup> 等），基于深度表征的零样本图像分类已成为当前主流。因此，基于深度表征的零样本图像分类是计算机视觉领域具有较高科学和实用价值的重要研究课题，本文对此展开研究。

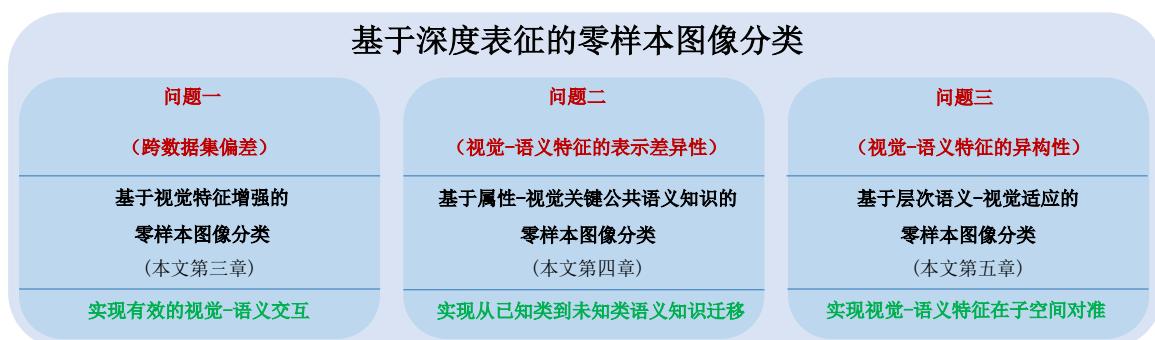


图 1.3 本文研究内容示意图。

## 1.2 研究内容与主要贡献

近年来，深度学习在传统的图像分类中发挥出了强大潜力。零样本图像分类也受益于强大的深度学习框架，但其也给零样本图像分类带来了新的挑战，如跨数据集偏差问题、视觉-语义特征的表示差异性问题、视觉-语义特征的异构性问题。如图1.3所示，针对以上几个关键问题，本论文将从三个方向展开研究：(1) 探究基于视觉特征增强的零样本图像分类算法，拟解决跨数据集偏差对零样本图像分类方法在视觉-语义交互上的限制；(2) 研究基于属性-视觉关键公共语义知识的思路，以移除视觉-语义特征的表示差异性问题对零样本图像分类从已知类到未知类知识迁移的阻碍；(3) 采用基于层次语义-视觉适应的公共子空间学习机制，应对视觉-语义特征的异构性造成视觉和语义特征在子空间中难以对准的挑战。视觉-语义交互是零样本

# 华中科技大学博士学位论文

---

图像分类的具体任务，从已知类到未知类的知识迁移是零样本图像分类的具体目标，视觉-语义异构特征在公共子空间对准是实现零样本图像分类的一种重要方法。

综上所述，本文聚焦于基于深度表征的零样本图像分类，分别从零样本图像分类的任务、目标和方法等三个层面对相应的问题进行剖析与并提出相应方法进行解决，从而全面推进零样本图像分类的进一步发展。

## 1.2.1 基于视觉特征增强的零样本图像分类

基于深度表征的零样本图像分类模型通常使用在 ImageNet<sup>[18]</sup> 上预先训练的深度卷积神经网络（Convolutional Neural Network, CNN）对零样本图像分类标准数据集（例如，粗粒度数据集 AWA2<sup>[6]</sup> 和细粒度数据集 CUB<sup>[4]</sup>）的图像提取视觉的深度表征。然而，由于数据收集过程可能会受到人为或系统因素的影响，两个数据集之间的分布存在较大差异性，即 ImageNet 数据集中的图像属性不能对零样本图像分类数据集的不同图像属性进行刻画，从而存在跨数据集偏差问题<sup>[33]</sup>。例如，在细粒度鸟类数据集 CUB 上存在各种各样的鸟，且这些鸟的形状、局部外观千变万化；而 ImageNet 中鸟的图像只能对鸟的“概念”进行描述，不能对不同鸟的不同局部外观进行刻画。因此，跨数据集偏差限制了从 ImageNet 迁移知识到零样本图像分类数据集，使得在零样本数据集图像上提取的视觉特征与预定义的属性没有较好语义关联，导致视觉特征的表征能力不足，根本上限制了零样本图像分类中有效的视觉-语义交互（嵌入式模型：视觉 → 语义；生成式模型：语义 → 视觉）。本文发现，ImageNet 和零样本标准数据集的分布偏差越大，视觉特征表示能力越差。例如，在粗粒度数据集 AWA2 上提取的视觉特征比细粒度数据集上提取的视觉特征具有更强的判别性和迁移性，使得零样本图像分类方法能在 AWA2 上取得更好的分类效果。这一现象进一步验证了跨数据集偏差问题存在于基于深度表征的零样本图像分类方法，而且是现有方法忽略的一个重要研究问题。因此，现有零样本图像分类模型在已知类和未知类上的分类性能上限都受到限制。换句话说，本文认为通过增强零样本数据集的视觉特征表征能力以缓解跨数据集偏差问题，实现视觉-语义的有效交互，提高零样本图像分类性能。

尽管微调技术可以在一定程度上缓解跨数据集偏差问题，但它不可避免地会导致其他更严重的问题<sup>[34–38]</sup>。例如，微调技术使得模型极易过拟合于已知类，阻碍了零样本图像分类从已知类到未知类的知识迁移，不满足零样本图像分类高泛化性能

# 华 中 科 技 大 学 博 士 学 位 论 文

---

的需求。为此，探索新的视觉特征增强方法以缓解零样本图像分类中的跨数据集偏差问题是本根上提高零样本分类效果的必要研究。针对此问题，本文面向嵌入式和生成式模型提出基于视觉特征增强的零样本图像分类方法。面向嵌入式零样本图像分类模型，本文利用图指导双注意力网络融合局部视觉特征和显式全局视觉特征对视觉特征进行增强。面向生成式零样本图像分类模型，本文利用前置特征精细化和后置特征精细化算法实现对视觉特征的增强。大量的实验结果表明本文方法均能够有效地增强视觉特征的判别性和迁移性，从而显著提高零样本图像分类性能。

## 1.2.2 基于属性-视觉关键公共语义知识的零样本图像分类

在零样本图像分类中，一个未知类的图像与一组已知类的图像共享不同的局部信息，这些局部信息由丰富的语义属性刻画（例如，“黄颜色的嘴”，“红颜色的腿”等）。因此，挖掘视觉和属性特征之间关键公共语义知识是实现零样本图像分类从已知类到未知类语义知识迁移的关键。然而，深度表征是全局视觉特征表示，并不能有效地对这些语义属性进行刻画，造成视觉-语义特征的表示差异性问题。该问题直接限制了零样本图像分类从已知类到未知类的语义知识迁移。虽然有一些基于注意力机制的零样本图像分类方法<sup>[2,39-42]</sup>利用语义信息作为指导以挖掘具有判别性的局部/细粒度视觉特征，从而使得视觉特征更准确地映射到语义特征空间实现零样本图像分类。然而，它们只是简单地利用单向注意力机制（属性 → 视觉注意力模型），仅能探索视觉和属性特征之间有限的公共语义知识。因此，准确且充分地挖掘视觉特征和属性特征之间潜在的公共语义知识，对零样本图像分类的知识迁移尤为重要。

针对此问题，本文提出了基于属性-视觉关键公共语义知识的零样本图像分类方法。方法使用基于属性指导的 Transformer 网络，利用单向跨注意力机制学习具有属性定位的视觉特征，对属性-视觉特征之间的关键公共语义知识进行准确表示。该方法利用互语义蒸馏网络，通过属性 → 视觉和视觉 → 属性两个双向注意力子网络分别学习基于属性的视觉特征和基于视觉的属性特征，并在互语义蒸馏学习机制的指导下，两个子网络更充分地挖掘关键语义特征。该方法将这两个网络集成为一个统一的模型，提高视觉-语义特征的语义一致性，实现零样本图像分类从已知类到未知类有效的语义知识迁移。

# 华中科技大学博士学位论文

---

## 1.2.3 基于层次语义-视觉适应的零样本图像分类

公共子空间学习是零样本图像分类的一种经典方法，它将视觉域和语义域特征映射到公共子空间，从而实现视觉-语义交互并进行知识迁移。然而，现有的公共子空间式零样本图像分类方法只通过单步适应（one-step adaptation）对视觉和语义域的特征分布进行对准<sup>[3,43-46]</sup>，忽略了语义-视觉域的异构特征表示（即异构性）同时存在特征分布差异和特征流形结构差异<sup>[47,48]</sup>，使得视觉和语义特征映射到不同的子流形空间上，未能实现视觉和语义特征在子空间中真正对准。如果此时采用欧几里德距离或流形距离<sup>[49,50]</sup>对不同类别之间的关系进行度量，分类器不可避免地对一些样本进行错误分类，从而导致零样本图像分类性能根本上受到限制。

针对此问题，本文提出一种基于层次语义-视觉适应的零样本图像分类方法。不同于现有方法使用两种不同的映射模型和分布对齐约束只进行特征分布对准，本文采用层次适应同时进行结构对准和分布对准以学习视觉和语义特征真正对准的公共子空间。该方法将结构适应和分布适应统一在两个局部共享的变分自编码器里面。在结构适应模块中，本文使用有监督的对抗差异性学习促使视觉和语义特征流形相互靠近，从而实现两种异构特征的流形结构对准。在分布适应模块中，本文使用一个公共编码器将结构对准的视觉和语义特征映射到分布对准的公共子空间，该编码器通过最小化分布对准子空间中的视觉和语义特征的多元高斯分布之间的 Wasserstein 距离实现分布对准。因此，本文提出的基于层次语义-视觉适应的零样本图像分类方法消除视觉-语义特征的异构性，实现视觉和语义特征在公共子空间的真正对准，有效地提高零样本图像分类性能。

## 1.3 论文章节安排与组织结构

本文的章节安排与组织结构概括如下：

第一章对零样本图像分类的研究背景及选题意义做了相应介绍，对零样本图像分类当前面临的关键难点进行了分析，并对本文的研究内容做了简要的概述。最后对本文的结构及贡献进行了梳理和总结。

第二章重点回顾了零样本图像分类国内外的研究进展，对嵌入式零样本图像分类、生成式零样本图像分类、以及公共子空间式零样本图像分类做了详细阐述，并分析了现有工作的不足及存在的主要问题。最后介绍了零样本图像分类的常用数据

# 华 中 科 技 大 学 博 士 学 位 论 文

---

集及评价指标。

第三章介绍了基于视觉特征增强的零样本图像分类。首先分析了基于深度表征的零样本图像分类存在的跨数据集偏差问题对分类效果的影响，随后面向嵌入式和生成式零样本图像分类提出两种解决方案应对该问题进行解决，包括基于图指导双注意力网络的嵌入式零样本图像分类和基于特征精细化的生成式零样本图像分类。

第四章介绍了基于属性-视觉关键公共语义知识的零样本图像分类。首先分析了基于深度表征的零样本图像分类存在视觉-语义特征的表示差异性问题，随后提出基于属性指导的 Transformer 网络和互语义蒸馏网络分别对视觉和属性特征之间关键公共语义知识的准确和充分挖掘，提高视觉和语义特征的一致性。

第五章介绍了基于层次语义-视觉适应的零样本图像分类。首先分析基于深度表征的零样本图像分类中视觉-语义特征的异构性问题，随后提出一种基于层次语义-视觉适应的零样本图像分类方法同时考虑异构特征之间的特征流形结构差异性和特征分布差异性，学习本真的公共子空间有效实现两种异构特征的对准。

第六章对本文做了全面的总结和归纳，指出本文的主要贡献和创新点，并指出未来研究趋势。

## 2 零样本图像分类的研究概述

### 2.1 国内外相关研究工作综述

零样本图像分类为实现从已知类到未知类的知识迁移，需要以额外的语义信息（例如，属性描述）作为桥梁连接已知类和未知类的关系。为实现这一目标，零样本图像分类的关键任务是进行有效的、准确的视觉-语义交互<sup>[9,49,50]</sup>。视觉特征可通过手工设计特征（例如，尺度不变特征变换特征<sup>[51]</sup>（Scale Invariant Feature Transform, SIFT），加速鲁棒特征特征<sup>[52]</sup>（Speeded Up Robust Features, SURF），梯度方向直方图特征<sup>[53]</sup>（Histogram of Oriented Gradients, HOG），局部自相似性特征<sup>[54]</sup>（Local Self-Similarity, LSS）等）或者基于深度神经网络学习的特征（即深度表征，如VGG<sup>[55]</sup>, GoogleNet<sup>[56]</sup>, ResNet<sup>[31]</sup>等），表示为视觉空间（Visual Space）。基于手工特征零样本图像分类存在以下两点不足：（1）手工特征对人的经验和专业知识的依赖性强，不能充分刻画图像的视觉外观（特别是局部细节），和属性特征之间的语义表示存在较大的鸿沟<sup>[57]</sup>；（2）为增强手工特征的表示能力，通常将多种手工特征进行组合，造成特征维度较高且存在较多的冗余特征<sup>[58]</sup>。为此，基于深度表征的零样本图像分类成为近期研究的主流<sup>[59]</sup>。语义信息通常由人工标注的属性值（即类语义向量）或者利用语言模型对属性名称学习相应的属性特征（例如，GloVe<sup>[60]</sup>），构成语义空间。根据视觉空间与语义空间的特征表示，可以采用不同的方式进行视觉-语义交互。

零样本图像分类根据不同的设置，可以进一步细分为不同方法。根据测试时候的分类范围，可将零样本图像分类方法分为传统的零样本图像分类（Conventional Zero-Shot Learning, CZSL）和广义的零样本图像分类（Generalized Zero-Shot Learning, GZSL）<sup>[6,27,61]</sup>。CZSL 只对未知类进行测试，而 GZSL 同时对已知类和未知类进行测试。由于 GZSL 更符合真实场景设置且更具挑战性，大量工作只做了 GZSL 设置的实验。根据训练时是否使用无标签的未知类图像，可将零样本图像分类分为归纳式零样本图像分类（Inductive ZSL）和直推式零样本图像分类（Transductive ZSL）<sup>[6,27,62,63]</sup>。由于直推式零样本图像分类训练时使用了未知类图像样本，并不完全符合零样本图像分类的设置。为此，归纳式零样本图像分类是主流方向，本文的所有研究方法均

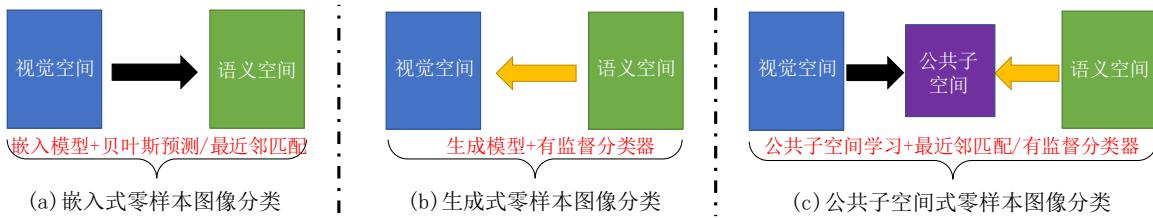


图 2.1 基于不同视觉-语义交互方式的零样本图像分类。

是基于这种方法。根据不同的视觉-语义交互方式，当前零样本图像分类方法可划分为嵌入式零样本图像分类、生成式零样本图像分类、公共子空间式零样本图像分类，如图2.1所示。本文将对这几种基于不同视觉-语义交互方式的零样本图像分类方法进行详细回顾和阐述。

### 2.1.1 嵌入式零样本图像分类

如图2.1(a) 所示，嵌入式零样本图像分类先将视觉特征嵌入到语义空间，再通过贝叶斯概率预测/最近邻匹配的方式实现对未知类的预测。早期的嵌入式零样本图像分类模型基于贝叶斯概率预测策略实现对未知类分类<sup>[22]</sup>。该方法使用属性语义表示作为链接视觉特征和标签的桥梁，在已知类数据上学习出属性分类器表示属性的先验概率，并根据属性分类器的概率预测值来推断未知类样本的标签，从而泛化到未知类数据。经典的贝叶斯概率预测模型有两种：直接属性预测模型（Direct Attribute Prediction, DAP）和间接属性预测模型（Indirect Attribute Prediction, IAP），如图2.2所示。图中  $\{a_1, \dots, a_A\}$  是属性集合表示类别中是否有相应的属性（即  $a_i \in \{0, 1\}$ ，可由人工专家标注<sup>[22]</sup>）、概念本体<sup>[64]</sup>（Concept Ontology）或者类和属性概率的语义关系<sup>[65]</sup>。 $\{y^{s,1}, \dots, y^{s,M}\}$  表示包含  $M$  个已知类数据标签集合， $\{y^{u,1}, \dots, y^{u,N}\}$  表示包含  $N$  个未知类数据标签集。在训练阶段，DAP 模型通过已知类训练数据学习属性分类器；在测试阶段，DAP 先使用属性分类器得到未知类样本的属性概率，然后根据贝叶斯概率推断测试样本标签的后验概率并进行标签预测：

$$p(y | f(x)) = \sum_{a \in \{0,1\}^A} p(y | a)p(a | f(x)) = \frac{p(y)}{p(a^y)} \prod_{i=1}^A p(a_i | f(x)) \quad (2.1)$$

其中  $A$  是属性个数， $p(y)$  和  $p(a^y)$  分别是标签和属性语义的先验概率， $p(a_i^y | f(x))$  是样本  $x$  经过第  $i$  个属性的概率输出值。

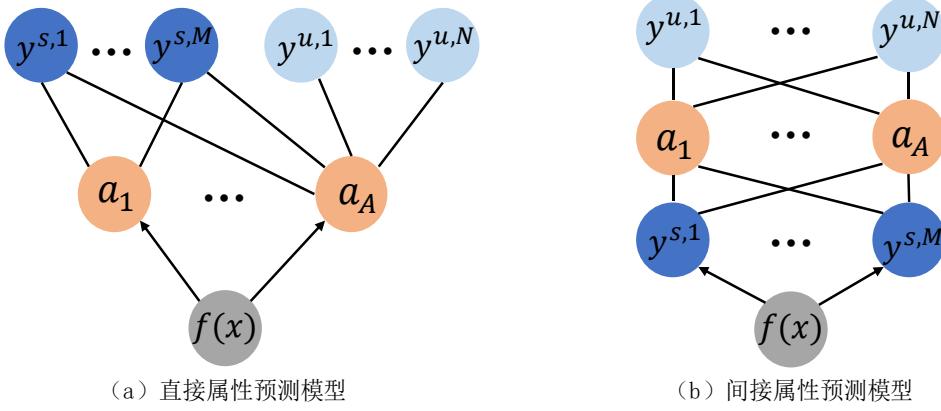


图 2.2 直接属性预测和间接属性预测模型结构示意图。

对于 IAP，其在训练阶段利用已知类的训练数据学习一个已知类的分类器，输出为  $p(y | f(x))$ ；测试阶段其先将未知类的测试数据输入到训练好的分类器，然后根据分类器输出值预测属性语义的后验概率，例如第  $i$  个属性的后验概率表示为：

$$p(a_i | f(x)) = \sum_{m=1}^M p(a_i | y^m) p(y^{s,m} | f(x)). \quad (2.2)$$

当得到未知类样本的属性预测概率之后，根据公式2.1推断标签的后验概率并进行标签预测。但基于贝叶斯概率预测的模型存在以下几点不足<sup>[6,66,67]</sup>：(1) 这种模型都是基于两步任务的处理过程，受限于中间任务和目标任务不能取得一致的优化结果；(2) 这种模型需要使用所有的类进行学习，它不能应用于增量学习的场景；(3) 这种模型过度依赖于标注好的属性。

针对此问题，近期的嵌入式零样本图像分类研究尝试直接学习一个由视觉空间到属性空间的一个映射，然后通过最近邻匹配的方式实现零样本图像分类<sup>[21,67-78]</sup>。ALE<sup>[67]</sup> 模型利用一个排序损失学习视觉和属性语义之间的对比函数语义嵌入。Frome 等人<sup>[68]</sup> 提出 DeVISE 模型学习一个视觉和语义空间的线性映射，并在大规模 ImageNet 数据集<sup>[18]</sup> 上进行实验。Akata 等人<sup>[69]</sup> 提出 SJE 学习一个双线性对比函数。Romera-Paredes 等人<sup>[70]</sup> 提出的 ESZSL 模型，利用平方损失去学习双线性对比函数并显式的正则化目标函数。Socher 等人<sup>[79]</sup> 提出的 CMT 模型使用包含两个隐藏层的神经网络去学习视觉空间到语义空间的非线性映射。为解决视觉-语义域的域偏差问题，Wang 等人<sup>[80]</sup> 提出 DPPN 同时进行学习视觉-语义嵌入学习和语义向量的动态更新优化。

# 华中科技大学博士学位论文

---

然而，这些模型都是直接将这种全局视觉特征直接嵌入到相应的语义空间，使得视觉特征和语义特征之间不具备较好的一致性，零样本图像分类不能取得较理想的结果<sup>[2,39,41,81]</sup>。例如，全局视觉特征没有很好的表示局部的细粒度图像外观，不具备更好的属性语义表示。因此，近期一些工作尝试使用注意力机制以学习局部视觉特征表示，从而提高视觉-语义的匹配一致性<sup>[2,39-42,81,82]</sup>。Yu 等人<sup>[82]</sup>首次使用堆栈式语义指导的注意力机制使得模型学习更具判别性的局部视觉特征。Xie 等人<sup>[2]</sup>提出 AREN 模型，在注意力和对比损失的指导下，挖掘更多局部特征表示。考虑到 AREN 模型缺乏局部特征之间的关系表示，Xie 等人<sup>[39]</sup>进一步使用区域图学习的方式增强视觉特征的局部推理。Zhu 等人<sup>[40]</sup>在没有额外监督信息的指导下，使用多个注意力机制学习具有局部定位的视觉特征。相反，Huynh 等人<sup>[81]</sup>利用属性特征信息，对视觉特征进行属性定位，从而提高视觉特征的局部细粒度信息表示。Liu 等人通过引入额外的“human gaze”监督信息，进一步增强基于注意力机制的局部特征增强。Xu 等人<sup>[41]</sup>通过联合局部和全局视觉特征，提高视觉语义的交互。然而，这些模型存在以下缺陷：(1)使用的注意力机制不能有效地对局部属性进行定位，对局部特征的表示能力不足；(2)这些模型使用的单向注意力机制（即属性 → 视觉注意力）不能充分挖掘视觉-属性之间的关键公共语义知识。

## 2.1.2 生成式零样本图像分类

由于零样图像分类中未知类样本的缺失，嵌入式零样本图像分类方法在学习分类的过程中只能使用已知类样本学习相应的分类器，不可避免地造成分类器很容易过拟合到已知类。虽然研究者通过应用自校准机制<sup>[46]</sup>，人工地给未知类提高分类置信度，但还是不能解决未知样本缺失造成的已知类过拟合问题。针对此问题，生成式零样本图像分类方法被进一步提出，如图2.1(b)所示。它们基于生成模型（例如，变分自编码器<sup>[83]</sup>（Variational Auto-Encoder, VAE），生成对抗网络<sup>[84]</sup>（Generative Adversarial Network, GAN），标准化流<sup>[85,86]</sup>（Normalizing Flows），VAEGAN<sup>[61,87]</sup>等）在已知类数据上学习一个语义 → 视觉的映射函数  $G$ ，从而使用学习的  $G$  对未知类生成大量的伪图像样本<sup>[88]</sup>或者伪特征样本<sup>[89-92]</sup>，随后将这些未知类伪样本和已知类训练样本一起训练一个有监督的分类器（例如，Softmax 等），从而将零样本图像分类问题转化为有监督的分类问题。

Arora 等人<sup>[88]</sup>提出了 SE-GZSL 模型，在类语义向量的指导下，利用 VAE 模型生

# 华中科技大学博士学位论文

---

成未知类的伪图像样本，进而通过训练一个有监督分类器实现零样本分类。由于生成高维的图像样本对生成器的要求较高，Xian 等人<sup>[89]</sup> 提出基于特征生成的 f-CLSGAN 模型，利用 CLSGAN 模型生成未知类的伪视觉特征样本，并利用未知类伪视觉特征样本和已知类的训练视觉特征样本训练一个有监督的分类器。Felix 等人<sup>[93]</sup> 认为 CLSWGAN 生成的视觉特征不具备强的语义表示，进一步加入视觉 → 语义的映射，与语义 → 视觉的映射够成一个环，并用环一致性损失对模型约束。实验结果表明提出 Felix 等人提出的 cycle-CLSWGAN 能显著提高零样本图像分类效果。Narayan 等人认为 cycle-CLSWGAN 的生成只在模型训练时候进行语义约束，而在特征生成和分类过程中没有进行约束，从而不利于整个生成模型的语义增强。为此，他们提出了提出 TF-VAEGAN 模型将特征模型训练、特征生成、分类同时进行语义约束。Xian 等人<sup>[61]</sup> 指出 CLSWGAN<sup>[89]</sup> 使用 WGAN 生成模型使得训练过程不够稳定，进一步提出 f-VAEGAN 模型同时兼顾 VAE 模型的稳定性和 GAN 模型的生成能力。后续大量的研究人员基于 f-VAEGAN 模型做进一步的扩展和优化。为实现生成的伪视觉特征具有更强的表征能力，Han 等人<sup>[94]</sup> 提出 RFF-GZSL 学习去除冗余特表示的视觉特征，Li 等人<sup>[95]</sup> 提出 LisGAN 学习稳定不变的视觉特征，Chen 等人<sup>[96]</sup> 通过提出 SDGZSL 进行特征解构。

## 2.1.3 公共子空间式零样本图像分类

如图2.1(c) 所示，公共子空间式零样本图像分类同时将视觉特征和语义特征映射到一个公共子空间，并在这个公共子空间里面进行最近邻匹配进行分类或者训练分类器进行有监督分类。这种子空间可以是具有实际可解释意义的空间<sup>[45,97-99]</sup>，也可以是任意的隐空间<sup>[66]</sup>。

早期的公共子空间式零样本学习通常在公共子空间中使用最近邻匹配的方式实现零样本分类。Zhang 等人<sup>[97]</sup> 提出 SSE 模型将已知类的不同类混合表示作为一个公共空间，测试时根据未知类样本是否具有相同的混合模型实现未知类分类。随后，Zhang 等人<sup>[43]</sup> 进一步提出 JLSE 模型，通过将视觉和语义映射到两个不同的子空间，再学习一个双线性对比函数度量视觉和语义的相似性。Changpinyo 等人<sup>[98]</sup> 指出这些子空间学习方法在子空间中没有考虑类间和类内差异性，提出双向学习机制进行结构约束以提高子空间特征的语义表示。为提高子空间具有更强的语义表示和视觉-语义一致性，Wang 等人<sup>[45]</sup> 提出双向子空间学习方法。类似于嵌入式零样本图像分类

方法，这些方法都面临明显的已知类过拟合问题。

针对此问题，研究者们利用未知类语义信息在子空间中生成大量的伪特征样本，并将已知类视觉特征也映射到子空间中，随后将子空间中未知类的伪特征和已知类特征一起训练一个有监督的分类器，类似于生成式零样本图像分类方法将零样本分类问题转化为有监督的分类问题<sup>[3,44]</sup>。Tsai 等人<sup>[44]</sup>提出 ReViSE 模型，利用自编码器将视觉和语义映射到公共空间，并通过分布一致性约束对视觉和语义域特征在子空间中对准。Schnfeld 等人<sup>[3]</sup>提出 CADA-VAE 模型，利用变分自编码器和特征分布约束将视觉和语义特征映射到公共子空间进行分类。得益于这些方法使用 VAE 模型可以同时对已知类和未知类进行数据增强，从而对小样本数据集更为有效（例如，SUN 数据集<sup>[5]</sup>）。由于视觉特征和语义特征属于异构域特征，当前方法都不容易将这两种特征进行对准，使得这个研究方法没能得到较好的发展。

## 2.2 现有工作的不足及存在的主要问题

随着深度学习技术的发展，基于深度表征的零样本图像分类方法取得的显著进步，包括嵌入式零样本图像分类、生成式零样本图像分类、公共子空间式零样本图像分类。然而，由于零样本图像分类属于人工智能中的一个基础型任务，且其发展还没有足够长的时间积累，仍有较多问题还未解决，离实际应用的需求依然存在较大的距离，这些问题主要体现在以下三个方面：

(1) 当前的零样本图像分类模型通常使用在 ImageNet<sup>[18]</sup> 上预先训练的卷积神经网络对零样本图像分类标准数据集的图像提取深度视觉表征<sup>[61,67,72,73,75,89,100,101]</sup>。然而，数据收集过程可能会受到人为或系统因素的影响，导致两个数据集之间的分布较明显差异，即 ImageNet 数据集中的图像属性不能对零样本图像数据集的图像属性进行刻画，从而存在跨数据集偏差问题。例如，在细粒度鸟类数据集 CUB 存在各种各样的鸟，且这些鸟形状和局部外观千变万化；而 ImageNet 中鸟的图像只能对鸟这个“概念”进行描述，不能对不同鸟的不同局部外观进行刻画。因此，跨数据集偏差限制了从 ImageNet 迁移知识到零样本图像数据集，使得在零样本数据集图像上提取的视觉特征与预定义的属性没有较好语义关联，导致视觉特征的表征能力不足，根本上限制了零样本图像分类进行有效的视觉-语义交互。

(2) 深度表征是全局视觉特征表示，其不能有效地对这些语义属性进行刻画，造

成视觉-语义表示的差异性问题。该问题直接限制了零样本图像分类从已知类到未知类的语义知识迁移。虽然有一些基于注意力机制的零样本图像分类方法<sup>[2,39-42]</sup>利用语义信息作为指导以挖掘具有判别性的局部/细粒度视觉特征，从而使得视觉特征更准确地映射到语义特征空间实现零样本图像分类。然而，它们只是简单地利用单向注意力机制(属性 → 视觉注意力模型)，仅能探索视觉和属性特征之间有限的公共语义知识。因此，如何有效地挖掘视觉和属性特征之间准确的、完整的公共语义知识以缓解视觉-语义的表示差异性问题是基于深度表征的零样本图像分类中的一个重要研究工作。

(3) 现有的公共子空间式零样本图像分类方法只使用单步适应对视觉和语义域的特征分布进行对准<sup>[3,43-46]</sup>，忽略了视觉-语义特征的异构性造成视觉和语义同时存在特征分布差异性和特征流形结构差异性，使得视觉和语义特征映射到不同的子流形空间上，未能实现视觉和语义特征在子空间中真正对准。如果此时我们采用欧几里德距离或流形距离来度量不同类别之间的关系，分类器不可避免地对一些样本进行错误分类，从而导致零样本图像分类的分类效果根本上受到限制。

## 2.3 零样本图像分类数据集及评价指标

本节将介绍零样本图像分类中常用的公开数据集和评价指标。

### 2.3.1 零样本图像分类数据集

零样本图像分类研究中较为常用的数据集包括 AWA1 (Animals with Attributes 1)<sup>[9]</sup>、AWA2 (Animals with Attributes 2)<sup>[6]</sup>、CUB (Caltech UCSD Birds 200)<sup>[4]</sup>、SUN (SUN Attribute)<sup>[5]</sup>。这些数据集的相关数据统计如表2.1所示，图2.3展示了数据集中的部分样本图像。数据集有两种不同的划分方式：旧划分和新划分。旧划分方式主要是早期基于传统图像特征的零样本图像分类方法使用的划分方式。近期的基于深度表征的零样本图像分类研究主要使用在 ImageNet 数据集上预训练的深度神经网络提取的视觉特征，由于 ImageNet 数据集会包含一些零样本标准数据集的未知类，造成数据泄露。为此，Xian 等人<sup>[6]</sup>提出了新的零样本数据集划分方式(即新划分)，并成为零样本图像分类主流的标准数据划分方式。此外，旧划分在测试时只包含未知类样本，使得其只能做传统的零样本图像分类。而新划分在测试集中既包含已知类



图 2.3 CUB、SUN、AWA2 等三个主流数据集部分样本展示图。

又包含未知类，可以同时做传统的零样本图像分类和广义的零样本图像分类。

表 2.1 四个最常用的零样本图像分类属的数据信息统计。

数据集	属性数量	类别数量			图像数量					
		总计	训练	测试	总计	旧划分		新划分		
						训练	测试	训练	测试 1	测试 2
AWA1 <sup>[9]</sup>	85	50	40	10	30475	24295	6180	19832	4958	5685
AWA2 <sup>[6]</sup>	85	50	40	10	37322	30337	6985	23527	5882	7913
CUB <sup>[4]</sup>	312	200	150	50	11788	8855	2933	7057	1764	2967
SUN <sup>[5]</sup>	102	717	645	72	14340	12900	1440	10320	2580	1440

注：“测试 1”表示测试集中已知类的图像数量，“测试 2”表示测试集中未知类的测试数量。

**AWA1<sup>①</sup>数据集：**AWA1 是动物图像数据集，由 50 个类共 20475 张图像组成，每个类具有 85 个属性。AWA1 是一个粗粒度数据集，其图像数据主要从 Google、Mocrospt、Yahoo、和 Flickr 等搜索引擎上搜索获取。数据集包含 50 个类别的图像，其中 40 个类别被划分为已知类，另外 10 个类别被划分为未知类。

**AWA2<sup>②</sup>数据集：**AWA2 数据集是数据集 AWA1 数据集的一个扩充，大部分研究工作只使用 AWA2 进行粗粒度数据集的实验。Xian<sup>[6]</sup> 等人按照 AWA1 中的 50 个类重新收集数据，一共包含 37,322 的图像数据，每个类包含的属性和数据集类别划分和 AWA1 保持一致。

**CUB<sup>③</sup>数据集：**CUB 是由 200 个鸟类共 11788 个图像构成细粒度数据集，每个类具有 312 个属性。数据集被划分为 150 个已知类用于模型训练，另外 50 个类作为未知

① <https://cvml.ist.ac.at/AwA/>

② <https://cvml.ist.ac.at/AwA2/>

③ [https://www.vision.caltech.edu/datasets/cub\\_200\\_2011/](https://www.vision.caltech.edu/datasets/cub_200_2011/)

类用于模型测试。

**SUN<sup>④</sup>数据集：**SUN 是一个细粒度场景分类数据集，由 717 个类别共 14340 张图像构成，且每个类别具有 102 个属性。数据集被划分为 645 个已知类用于模型训练，另外 72 个类作为未知类用于模型测试。由于 SUN 数据集中的每个类大约只包含 16 张图像，数据集较为不足。为此，相较于嵌入式零样本图像分类方法，生成式/公共子空间式零样本图像分类进行数据增强能在 SUN 数据集上取得更好的效果。

## 2.3.2 零样本图像分类评价指标

在零样本图像分类中，常用的定量评价指标是分类精度。传统的零样本图像分类和广义的零样本图像分类的具体评价指标有所区别。

**传统的零样本图像分类评价指标：**对于传统的零样本图像分类方法，因为其只对未知类进行测试，通常采用未知类的平均 Top-1 识别率 (acc) 作为评价指标，定义为：

$$acc = \frac{1}{\|\mathcal{Y}_u\|} \sum_{y \in \mathcal{Y}_u} \frac{y\text{-th 未知类中正确分类样本数量}}{y\text{-th 未知类样本总数量}},$$

$\mathcal{Y}_u$  为未知类样本标签集合。

**广义的零样本图像分类评价指标：**对于广义的零样本图像分类方法，其同时使用已知类的平均 Top-1 识别率 (S)、未知类的平均 Top-1 识别率 (U) 以及它们的调和均值 (Harmonic Mean, H) 对模型进行评估。这些指标具体定义如下：

$$S = \frac{1}{\|\mathcal{Y}_s\|} \sum_{y \in \mathcal{Y}_s} \frac{y\text{-th 已知类中正确分类样本数量}}{y\text{-th 已知类样本总数量}}, \quad (2.3)$$

$$U = \frac{1}{\|\mathcal{Y}_u\|} \sum_{y \in \mathcal{Y}_u} \frac{y\text{-th 未知类中正确分类样本数量}}{y\text{-th 未知类样本总数量}}, \quad (2.4)$$

$$H = \frac{2 \times S \times U}{S + U}, \quad (2.5)$$

$\mathcal{Y}_s$  为已知类样本标签集合。

## 2.4 本章小结

本章根据零样本图像分类中的视觉-语义交互方式将其划分为三种基本方法进行了系统的回顾，包括嵌入式零样本图像分类、生成式零样本图像分类、公共子空间

---

④ <https://cs.brown.edu/~gmpatter/sunattributes.html>

# 华 中 科 技 大 学 博 士 学 位 论 文

---

式零样本图像分类。通过深刻分析现有方法，总结出当前零样本图像分类方法亟待解决的三大难点：跨数据集偏差、如何挖掘属性-视觉的关键公共语义知识、如何实现语义-视觉的异构特征对准，并对这三个难点进行了简要分析。

### 3 基于视觉特征增强的零样本图像分类

#### 3.1 引言

随着研究者们投入大量精力来解决零样本图像分类中的视觉-语义域偏移问题和已知类偏差问题<sup>[9,39,45,67,69,102]</sup>，零样本图像分类研究取得了较为显著的进展<sup>[7,81,89,103–107]</sup>。不同类型的方法不断被提出，包括嵌入式零样本图像分类<sup>[46,104,108–110]</sup>、生成式零样本图像分类<sup>[2,88,89,109,111]</sup>、公共子空间式零样本图像分类<sup>[3,44–46,68]</sup>。相较于手工视觉特征提取，深度神经网络对图像提取的深度表征具有更强的表征能力，使得以上主流方法均以深度表征作为其视觉特征的表示。考虑到当前的零样本数据集的样本量均相对较少，不易利用这些数据训练一个深度神经网络用于视觉特征提取。为此，他们通常在 ImageNet 数据集<sup>[18]</sup> 上预训练一个卷积神经网络模型（Convolutional Neural Network, CNN）<sup>①</sup>，并利用预训练好的 CNN Backbone 对零样本数据集提取相应的视觉深度表征。然而，由于数据收集过程会受到人为或系统因素的影响，导致两个数据集之间的分布存在差异性，使得零样本数据集上提取的视觉特征是低质量的（即视觉特征不具备好的判别性表示，如图3.1(a) 所示）。这些低质量的视觉特征与预定义的属性没有强的语义关联，根本上限制了零样本图像分类中有效的视觉-语义交互。本文将此问题定义为跨数据集偏差问题。例如，细粒度鸟类不同类的外观属性可以在 CUB 数据集<sup>[4]</sup> 中找到，但在 ImageNet 中无法找不到。因此，如果不进行任何进一步的视觉特征增强使得视觉特征具有属性相关的信息表示，而是直接将视觉知识从 ImageNet 迁移到新的零样本数据集，必然导致零样本分类方法在已知类和不未知类上的识别性能表现不佳。换句话说，本文认为通过增强零样本数据集的视觉特征以解决零样本图像分类中的跨数据集偏差问题，将有效提高零样本图像分类性能。

为了直观地度量 ImageNet 数据集和零样本数据集之间的偏差，本文使用最大平均差异性<sup>[112,113]</sup>（Maximum Mean Discrepancy, MMD）对跨数据集偏差进行定量的度量。度量结果显示：ImageNet 数据集和 ImageNet 数据集的偏差量为 0.006（即  $MMD_{ImageNet-ImageNet} = 0.006$ ），ImageNet 数据集和 AWA2 数据集<sup>[6]</sup> 的偏差量为

---

① 本文将预训练好的卷积神经网络模型命名为 CNN Backbone。

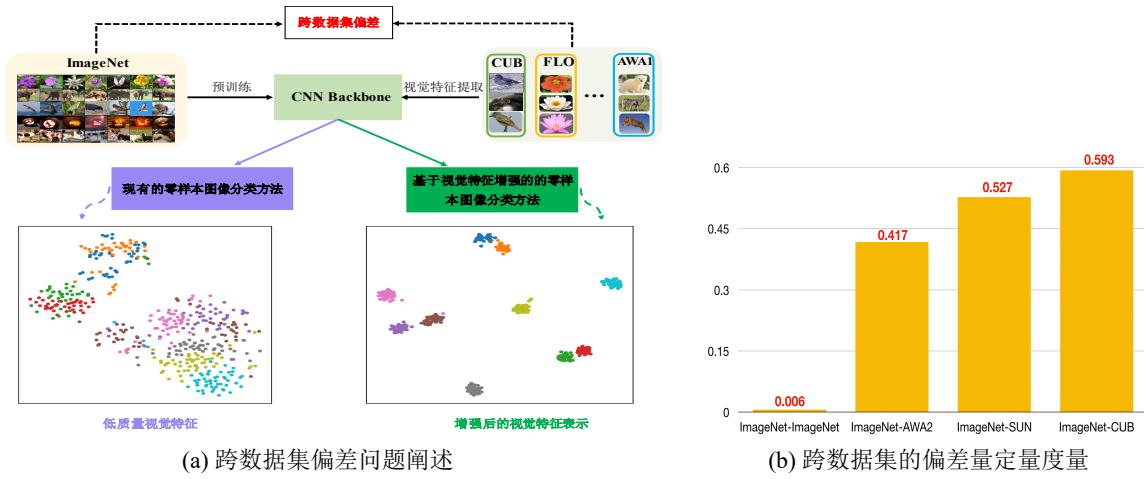


图 3.1 跨数据集偏差问题阐述与偏差量定量度量。

0.417（即  $MMD_{ImageNet-AWA2} = 0.417$ ），ImageNet 和 SUN 数据集<sup>[5]</sup> 的偏差量为 0.527（即  $MMD_{ImageNet-SUN} = 0.527$ ），ImageNet 数据集和 CUB 数据集<sup>[4]</sup> 的偏差量为 0.593（即  $MMD_{ImageNet-CUB} = 0.593$ ），如图3.1(b) 所示。可以直观地看出，MMD 可真实地、准确地反应 ImageNet 数据集和不同零样本数据集的偏差量。例如：(1) ImageNet 数据集自身的偏差量几乎为零；(2) 由于 ImageNet 数据集和粗粒度数据集（例如 AWA2）都是对粗粒度“类概念”的刻画，它们存在较小的跨数据集偏差；(3) 细粒度数据集需要对不同的细粒度“属性”对类别图像信息进行刻画，使得 ImageNet 数据集和细粒度数据集（例如，CUB 和 SUN 数据集）存在较大跨数据集偏差。因此，解决零样本图像分类中的跨数据集偏差问题是亟待解决的重要难点。本文首次将跨数据集偏差问题定义为零样本图像分类中的一个开放性问题。

在别的领域也存在类似问题，例如医学图像处理。利用微调对跨数据集偏差问题进行了缓解，但不可避免地会导致其他更严重的问题<sup>[34-38]</sup>：(1) 在小数据集上进行微调，模型很容易过拟合于已知类，这不利于零样本图像分类的知识迁移；(2) 由于微调需要对 CNN Backbone 的部分参数进行重新训练，降低了零样本图像分类模型的效率。为此，本文面向嵌入式和生成式模型提出了基于视觉特征增强的零样本图像分类方法解决跨数据集偏差这一挑战性问题，分别是基于图指导双注意力网络的嵌入式零样本图像分类（第3.2章节）、基于视觉特征精细化的生成式零样本图像分类（第3.3章节）。下文将对本章提出的这两种基于视觉特征增强的零样本学习方法进行详细阐述。

## 3.2 基于图指导双注意力网络的嵌入式零样本图像分类

### 3.2.1 研究动机

嵌入式零样本图像分类是零样本图像分类中的重要方法之一，其通过将视觉特征嵌入到语义空间，再利用最近邻匹配实现零样本图像分类。当前大部分嵌入式零样本图像分类方法<sup>[67,72,73,75,100,101]</sup>直接利用 ImageNet 数据集上预训练的 CNN Backbone (例如，ResNet-101<sup>[31]</sup>) 对零样本数据集的图像提取相应的隐式全局视觉特征，如图3.2(a)所示。然而，这种隐式全局特征不能对局部语义进行刻画，使得零样本图像分类方法不能进行有效地语义知识迁移。近期基于注意力机制的零样本图像分类方法<sup>[2,39–42,81]</sup>通过学习局部视觉特征提高其局部语义表示，克服基于隐式全局特征的零样本图像分类方法的不足，如图3.2(b)所示。但这种局部视觉特征不能对不同局部特征之间的显式关系进行表示，使得模型不能对同一类别的不同姿态、不同视角的图像目标进行有效的表示。本文认为，当前方法的这些不足均受限于章节3.1指出跨数据集偏差问题，使得当前的嵌入式零样本图像分类方法学习的视觉特征判别性严重不足，根本上限制了视觉和语义之间的有效交互。

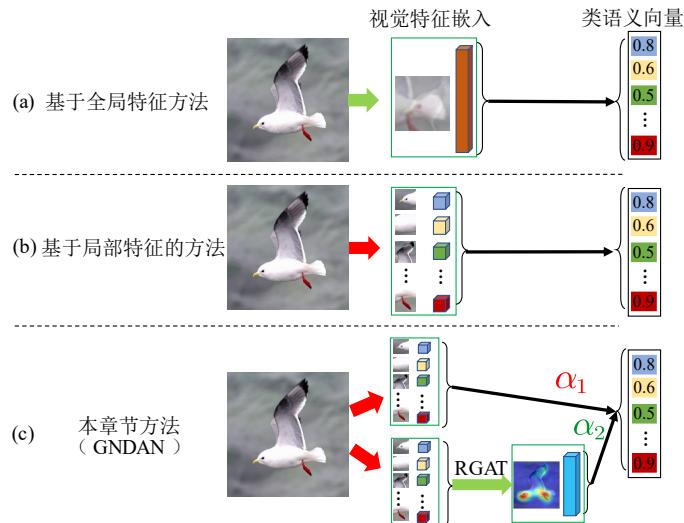


图 3.2 不同嵌入式零样本图像分类方法的对比。

针对此问题，本文提出了基于图指导双注意力网络的嵌入式零样本图像分类方法 (Graph Navigated Dual Attention Network for Zero-Shot Image Classification, GNDAN) 同时学习并融合局部视觉特征和显示全局视觉特征，增强视觉特征的判别性，如

图3.2(c) 所示。GNDAN 由区域指导的注意力网络（Region-guided Attention Network, RAN）和区域指导的图注意力网络（Region-guided Graph Attention Network, RGAT）组成。其中，RAN 采用软空间注意力机制<sup>[2]</sup>（Soft Space Attention），通过抑制背景和冗余前景区域挖掘具有可区分性的局部视觉特征，并使用局部视觉特征嵌入编码器（Local Visual Embedding Encoder）将其编码为局部视觉特征嵌入向量。不同于现有的全局特征学习方法直接隐式地学习全局视觉特征表示，GNDAN 的 RGAT 通过学习局部域之间的关系显示地表示全局特征。RGAT 首先采用基于属性的注意力机制（Attribute-Based Attention）获取基于属性的局部特征，其中每个属性聚焦于最相关的图像区域。为了进一步表示这些基于属性的局部特征关系，RGAT 进一步将它们表示为属性区域图（Attribute Region Graph），并采用图注意网络（例如，Chebyshev<sup>[114]</sup>, GCN<sup>[115]</sup>, GraphSAGE<sup>[116]</sup> 以及 GAT<sup>[117]</sup> 等）挖掘属性区域图中基于属性的区域特征之间的关系对显式全局特征进行表示，并使用全局视觉特征嵌入编码器（Global Visual Embedding Encoder）将其编码为全局视觉特征嵌入向量。最后，GNDAN 融合 RAN 学习的局部特征和 RGAT 学习的显示全局特征以增强视觉特征表示，缓解跨数据集偏差问题并提高零样本图像分类的效果。在三个主流标准数据集（例如，CUB 数据集<sup>[4]</sup>、SUN 数据集<sup>[5]</sup> 和 AWA2 数据集<sup>[6]</sup>）上的大量实验结果验证了 GNDAN 的有效性。

### 3.2.2 基于图指导的双注意力网络

在介绍方法之前，本章节先对一些基本的符号和问题定义进行介绍。由已知类  $C^s$  构成的训练集  $\mathcal{D}_{tr} = \{(x_{tr}^s, y^s, c_{y^s}) \mid x_{tr}^s \in \mathcal{X}^s, y^s \in \mathcal{Y}^s, c_{y^s} \in \mathcal{C}^s\}$  作为训练数据，其  $x_{tr}^s \in \mathcal{X}$  表示训练集图像样本，而  $y^s \in \mathcal{Y}^s$  是相应的类标签。测试集  $\mathcal{D}_{te} = \{\mathcal{D}_{te}^s, \mathcal{D}_{te}^u\}$  由两个子集构成，其中  $\mathcal{D}_{te}^s = \{(x_{te}^s, y^s, c_{y^s}) \mid x_{te}^s \in \mathcal{X}^s, y^s \in \mathcal{Y}^s, c_{y^s} \in \mathcal{C}^s\}$  是已知类测试子集， $\mathcal{D}_{te}^u = \{(x_{te}^u, y^u, c_{y^u}) \mid x_{te}^u \in \mathcal{X}^u, y^u \in \mathcal{Y}^u, c_{y^u} \in \mathcal{C}^u\}$  是未知类测试子集。 $x_{te}^u \in \mathcal{X}$  是未知类图像， $y^u \in \mathcal{Y}^u$  是相应的标签。特别指出， $\mathcal{Y}^s \cap \mathcal{Y}^u = \emptyset$ 。具有  $A$  个属性的类语义向量表示为  $z^c = [z_1^c, \dots, z_A^c]^\top = \phi(y)$ ，其中  $c \in \mathcal{C}^s \cup \mathcal{C}^u = \mathcal{C}$ . 属性特征  $\mathcal{V}_A = \{v_a\}_{a=1}^A$  通过语言模型（例如，GloVe 模型<sup>[60]</sup>）学习属性名称的词向量表示。

本文提出的 GNDAN 由一个区域指导的注意力网络（RAN）和一个区域指导的图注意力网络（RGAT）构成，如图3.3所示。RAN 采用软空间注意力机制<sup>[2]</sup> 抑制背景和冗余前景区域挖掘具有可区分性的局部视觉特征。RGAT 首先使用基于属性的

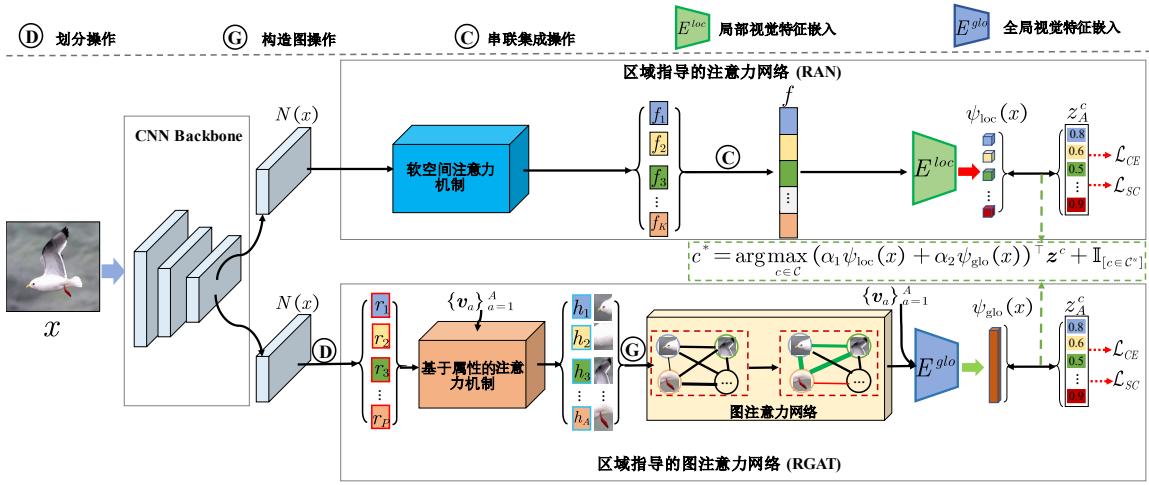


图 3.3 本章提出的 GNDAN 模型结构示意图。

注意力机制获取基于属性的局部特征，并通过构造属性区域图表示这些基于属性的局部特征关系，随后并采用图注意网络学习显式全局特征。本文使用基于属性的交叉熵损失和自校准损失对 GNDAN 进行模型优化。

**区域指导的注意力网络：**由于语义信息为零样本图像分类从已知类到未知类进行有效的知识迁移提供了局部辨别性视觉线索，因此局部视觉特征学习可以对视觉特征进行增强<sup>[40,40,81,82]</sup>。本文的 GNDAN 首先利用区域指导的注意力网络（RAN）基于软空间注意机制和局部视觉嵌入编码器学习局部视觉特征并将其嵌入到具有可区分性的语义空间中。

**软空间注意力机制：**软空间注意机制基于 CNN Backbone（例如，ResNet-101<sup>[31]</sup>）的最后一个卷积的特征表示  $N(x) \in \mathbb{R}^{H \times W \times C}$ ，将图像  $x$  表示为  $K$  个区域/局部特征。 $N(x)$  具有  $C$  个大小为  $H \times W$  的特征图。首先，在  $N(x)$  上使用  $1 \times 1$  卷积 ( $Conv_{1 \times 1}(\cdot)$ ) 和 Sigmoid 激活函数 ( $Sigmoid(\cdot)$ ) 做进一步学习，可以获得  $K$  个注意力掩码  $\{M_i(x)\}_{i=1}^K$ ：

$$M = \text{Sigmoid}\left(Conv_{1 \times 1}(N(x))\right). \quad (3.1)$$

因此，第  $i$ -个注意力掩码  $M_i(x) = M[:, :, i] \in \mathbb{R}^{H \times W}$  可以从  $M$  中划分出来。然后，RAN 根据这些注意力掩码可以获得  $K$  个对应的局部特征图  $\{T_i(x)\}_{i=1}^K$  w.r.t.  $N(x)$ ：

$$T_i(x) = N(x) \odot R(M_i(x)), \quad (3.2)$$

其中， $R(\cdot)$  将输入扩展为与  $N(x)$  相同的大小， $\odot$  是哈达玛积 (Hadamard Product)。最

后, RAN 进一步应用全局最大池化层 (Global Max-Pooling) 对每个  $T_i(x) \in \mathbb{R}^{H \times W \times C}$  进行特征压缩, 获得  $K$  个局部视觉特征  $\{f_i(x)\}_{i=1}^K$ , 且  $f_i(x) \in \mathbb{R}^C$ 。这些局部特征被串联为一个局部特征向量  $f \in \mathbb{R}^{KC}$ 。

**局部视觉特征嵌入编码器:** 将局部视觉特征嵌入到语义空间时, 考虑到特征维数的极度下降而导致的信息丢失, RAN 使用一个局部视觉特征嵌入编码器 ( $E^{loc}$ ) 实现有效的视觉 → 语义的映射。 $E^{loc}$  本质上是一个多层感知机 (Multi-Layer Perceptron, MLP), 结构为  $KC$  维-4096 维- $A$  维。 $A$  为类语义向量的维度, 即数据集属性的个数 (例如, 在 AWA2 数据集上,  $A = 85$ )。基于串联的局部视觉特征向量  $f$ , 我们得到局部视觉特征嵌入向量  $\psi_{loc}(x)$ :

$$\psi_{loc}(x) = E^{loc}(f). \quad (3.3)$$

通过实验发现, 当类语义向量的维数较低时,  $E^{loc}$  非常有效, 例如在 AWA2 数据集 ( $A = 85$ ) 和 SUN 数据集 ( $A = 102$ )。

**区域指导的图注意力网络:** 本文章节3.2.1分析发现: (1) 尽管局部视觉特征对零样本图像分类的局部属性表示很重要, 但局部视觉特征之间的关系也应充分利用并提高视觉特征的鲁棒性; (2) 当前隐式的全局视觉特征不能准确表示这些局部视觉特征之间的关系。为此, GNDAN 进一步使用区域指导的图注意力网络 (RGAT), 基于语义信息学习显式的全局视觉特征表示以进一步增强视觉特征。和 Huynh 等人<sup>[81]</sup>提出的 DAZLE 模型一样使用同等大小单元格划分方法<sup>[118]</sup>, RGAT 首先利用划分操作将视觉特征图  $N(x)$  划分为一组区域视觉特征  $\mathcal{R}_P = \{r_p\}_{p=1}^P$ ,  $P = H \times W$ 。然后, RGAT 使用基于属性的注意机制学习基于属性的局部特征, 其中每个属性聚焦于最相关的图像区域。因此, 可以基于这些基于属性的局部特征构建属性区域图表示局部特征之间的关系。由于基于注意机制的图网络有利于捕获一组节点的全局信息<sup>[117,119–121]</sup>, RGAT 进一步利用图注意力网络 (例如, GAT<sup>[117]</sup>) 挖掘属性区域图上基于属性的局部特征之间的关系, 并用于表示显式的全局特征。最后, 全局视觉特征嵌入编码器将这些全局特征表示嵌入到类语义空间中。

**基于属性的注意力机制:** RGAT 使用基于属性的注意力机制定位与属性最相关的图像区域, 学习基于属性的局部特征。基于属性特征  $\mathcal{V}_A$ , RGAT 将第  $a$  个属性的

基于属性的局部特征  $h_a$  定义为:

$$h_a = \sum_{p=1}^P \beta(r_p, v_a) r_p \quad (3.4)$$

$$= \sum_{p=1}^P \frac{\exp(v_a^\top W_\beta r_p)}{\sum_i \exp(v_a^\top W_\beta r_i)} r_p, \quad (3.5)$$

其中  $\beta(r_p, v_a)$  是关注图像  $x$  中第  $a$  个属性区域的注意力权重,  $W_\beta$  是可学习的矩阵用于测量每个属性特征  $v_a$  和每个图像区域  $r_p$  之间的关联度。本质上来看,  $h_a$  表示与第  $a$  个属性最相关的图像区域。因此, 给定图像  $x$ , RGAT 可以获得一组基于属性的局部特征  $\{h_a\}_{a=1}^A$ 。

**图注意力网络:** 当学习到基于属性的局部特征之后, RGAT 使用图注意网络 (GAT<sup>[117]</sup>) 对这些局部特征的关系进行挖掘并表示为显式的全局视觉特征。首先, RGAT 为每个输入图像构造一个属性区域图  $\Gamma$ 。 $\Gamma$  的节点是基于属性的局部特征, 边是它们之间的余弦相似度  $\Gamma_{i,j} = \langle h_i(x), h_j(x) \rangle$ 。因此,  $\Gamma$  是一个完全连通的无向图, 具有  $A$  节点和  $\frac{A^2-A}{2}$  条边。

由于单层图注意力网络是图注意力网络的基本结构, 本文以单层图注意力网络展开介绍。图注意力层首先使用一个可学习的线性变换层为每个节点学习跟更强的表达能力, 例如  $\{h_1, h_2, \dots, h_a\}_{a=1}^A \rightarrow \{W_g h_1, W_g h_2, \dots, W_g h_a\}_{a=1}^A$ 。该线性变换的权重矩阵为  $W_g \in \mathbb{R}^{C' \times C}$ , 本文将  $C'$  设置为 256。随后, 一个共享的自注意机制作用于节点上为每条边  $(j, i)$  计算一个注意力值, 这个注意力值表示近邻点特征  $j$  与特征节点  $i$  的关联程度:

$$e(h_i, h_j) = \text{LeakyReLU}(\gamma^\top \cdot [W_g h_i \| W_g h_j]), \quad (3.6)$$

其中  $\gamma$  表示单层前馈网络的权重向量,  $\|$  表示向量串联。使用 softmax 对所有近邻点  $j \in \mathcal{N}_i$  的注意力值进行归一化, 注意力函数公式如下:

$$\tau_{ij} = \text{softmax}(e(h_i, h_j)) = \frac{\exp(e(h_i, h_j))}{\sum_{j' \in \mathcal{N}_i} \exp(e(h_i, h_{j'}))}. \quad (3.7)$$

基于归一化后的注意力值, RGAT 计算相邻节点特征的加权平均值并进行非线性变换 ( $\delta$ ), 作为节点  $i$  的新特征表示:

$$h'_i = \delta \left( \sum_{j \in \mathcal{N}_i} \tau_{ij} \cdot W_g h_j \right), \quad (3.8)$$

# 华 中 科 技 大 学 博 士 学 位 论 文

---

与 Vaswani 等人<sup>[32]</sup> 的方法类似，GNDAN 也可以利用多头注意力将图注意层扩展为图多头注意力层（Graph Multi-Head Attention Layer），以稳定自注意机制的学习过程。具体来说， $T$  个独立注意机制执行等式3.8的转换，然后将其特征串联为节点  $i$  的输出特征：

$$h'_i = \left\|_{t=1}^T \delta \left( \sum_{j \in \mathcal{N}_i} \tau_{ij}^t W_g^t h_j \right) \right\|. \quad (3.9)$$

本文使用公式3.9去更新 RGAT 中属性区域图的节点。

为了避免图学习过程中的过平滑问题，本文进一步将属性区域图中每个节点的所有层特征串联起来。因此，属性图中节点  $i$  的最终输出表示为：

$$h_i^{output} = \left\|_{l=1}^L h_i'^{(l)} \right\|. \quad (3.10)$$

因此，GNDAN 得到了显式的全局视觉特征  $\{h_a^{output}\}_{a=1}^A$ 。

**全局视觉特征嵌入编码器：**不同于局部视觉特征嵌入编码器 ( $E^{loc}$ ) 直接使用 MLP 将高维视觉特征映射到低维类语义特征空间，RGAT 使用全局视觉特征嵌入编码器 ( $E^{glo}$ ) 并基于属性特征  $\mathcal{V}_A$  将显式的全局视觉特征进行语义嵌入：

$$\psi_{glo}(x)_a = E^{glo}(h_a^{output}) = v_a^\top W_{glo} h_a^{output}, \quad (3.11)$$

其中  $W_{glo}$  是一个可学习的嵌入矩阵，将视觉特征  $h_a^{output}$  嵌入到第  $a$  个属性语义空间。给定一组属性特征  $\{v_a\}_{a=1}^A$ ，RGAT 将获得整个全局视觉特征嵌入的语义向量  $\psi_{glo}(x) = [\psi_{glo}(x)_1, \psi_{glo}(x)_2, \dots, \psi_{glo}(x)_A]^\top$ 。

**模型优化：**为了稳定地优化 GNDAN 模型，本文使用相似的损失函数联合优化 RAN (如公式3.15) 和 RGAT (如公式3.15)。

**基于属性的交叉熵损失：**当一个属性直观地出现在图像中时，相关的视觉特征嵌入会投影到其类语义向量  $z^c$  附近。因此，基于属性的交叉熵损失  $\mathcal{L}_{ACE}$  被用于优化 GNDAN 模型。具体而言，视觉特征的语义嵌入（即  $\psi_{loc}(x)$  或  $\psi_{glo}(x)$ ）与每个类语义向量之间的点积，并形成相关类预测值。因此，给定一批量  $nb$  的训练图像  $\{x_i^s\}_{i=1}^{nb}$  及其相关的类语义向量  $z^c$ ， $\mathcal{L}_{CE}$  定义为：

$$\mathcal{L}_{ACE}(F(x_i)) = \frac{1}{n_b} \sum_{i=1}^{n_b} \log \frac{\exp(F(x_i) \times z^c)}{\sum_{\hat{c} \in \mathcal{C}} \exp(f(x) \times z^{\hat{c}})}, \quad (3.12)$$

其中，在 RAN 中  $F(x) = \psi_{loc}(x_i)$ ；而在 RGAT 中， $f(x) = \psi_{glo}(x_i)$ .

自校准损失：当使用基于属性的交叉熵损失在已知类进行模型优化时，GNDAN 不可避免地过拟合于已知类。为了应对这一挑战，本文进一步采用自校准损失<sup>[40,41,81]</sup>，使得模型在训练期间将一些预测概率从已知类转移到未知类，因此， $\mathcal{L}_{SC}$  的表示如下：

$$\mathcal{L}_{SC}(f(x)) = -\frac{1}{n_b} \sum_{i=1}^{n_b} \sum_{c'=1}^{\mathcal{C}^u} \log \frac{\exp(f(x_i) \times z^{c'} + \mathbb{I}_{[c' \in \mathcal{C}^u]})}{\sum_{\hat{c} \in \mathcal{C}} \exp(f(x_i) \times z^{\hat{c}} + \mathbb{I}_{[\hat{c} \in \mathcal{C}^u]})}, \quad (3.13)$$

其中  $\mathbb{I}_{[c \in \mathcal{C}^u]}$  是一个指示符函数（当  $c \in \mathcal{C}^u$  时，函数值为 1，否则为-1）。直观上来看， $\mathcal{L}_{SC}$  鼓励在训练期间将非零概率分配给未知类，使得 GNDAN 可以产生更大的非零概率给未知类测试样本。同时， $\mathcal{L}_{SC}$  可以通过控制其权重平衡已知类和未知类之间预测。

最后，结合基于属性的交叉熵损失和自校准损失函数联合优化 RAN 和 RGAT。因此，GNDAN 的总体优化函数为：

$$\mathcal{L}_{total} = \mathcal{L}_{RAN}(\psi_{loc}(x)) + \mathcal{L}_{RGAT}(\psi_{glo}(x)), \quad (3.14)$$

其中，

$$\mathcal{L}_{RAN}(\psi_{loc}(x)) = \mathcal{L}_{CE}(\psi_{loc}(x)) + \lambda_{SC} \mathcal{L}_{SC}(\psi_{loc}(x)), \quad (3.15)$$

$$\mathcal{L}_{RGAT}(\psi_{glo}(x)) = \mathcal{L}_{CE}(\psi_{glo}(x)) + \lambda_{SC} \mathcal{L}_{SC}(\psi_{glo}(x)), \quad (3.16)$$

$\lambda_{SC}$  为损失函数权重用于控制  $\mathcal{L}_{SC}$  平衡已知类和未知类之间预测。

**零样本图像分类：**为进一步增强视觉特征，GNDAN 融合 RAN 学习的局部特征和 RGAT 学习的全局特征进行零样本图像分类预测。GNDAN 首先利用训练好的 RAN 和 RGAT 获得测试样本  $x$  的语义嵌入向量，即  $\psi_{loc}(x)$  和  $\psi_{glo}(x)$ 。然后，GNDAN 使用两个组合系数 ( $\alpha_1, \alpha_2$ ) 对 RAN 和 RGAT 的预测进行融合，并通过自校准预测测试样本  $x$  的标签，表示为：

$$c^* = \arg \max_{c \in \mathcal{C}} (\alpha_1 \psi_{loc}(x) + \alpha_2 \psi_{glo}(x))^T z^c + \mathbb{I}_{[c \in \mathcal{C}^u]}. \quad (3.17)$$

### 3.2.3 实验结果与分析

为验证 GNDAN 方法的有效性，本章节在三个主流的零样本图像分类标准数据集上同时进行 CZSL 和 GZSL 实验，包括两个细粒度数据集（CUB<sup>[4]</sup>，SUN<sup>[5]</sup>）和

# 华 中 科 技 大 学 博 士 学 位 论 文

---

一个粗粒度数据集 (AWA2<sup>[6]</sup>)。本文依据 Xian 等人<sup>[6]</sup>最新的数据集划分方式进行模型训练和测试，数据集的详细介绍见章节2.3.1。本文利用在 ImageNet 上预训练的 ResNet-101 作为 CNN Backbone 对图像样本（分辨率裁剪为  $448 \times 448$ ）提取相应的视觉特征。模型的学习率和批量大小分别设置为 0.0001 和 50。根据在验证集上的实验，RAN 分支上的特征区域块数目  $K$  在所有数据集上均被设置为 10；对于  $\lambda_{SC}$ ，在 CUB 和 AWA2 数据集上被设置为 0.1，SUN 数据集上被设置为 0.01；对于组合系数  $(\alpha_1, \alpha_2)$ ，在 CUB 和 SUN 数据集上被设置为 (0.5, 0.5)，在 AWA2 数据集上被设置为 (0.8, 0.2)。

本章节通过以下实验验证 GNDAN 的有效性：

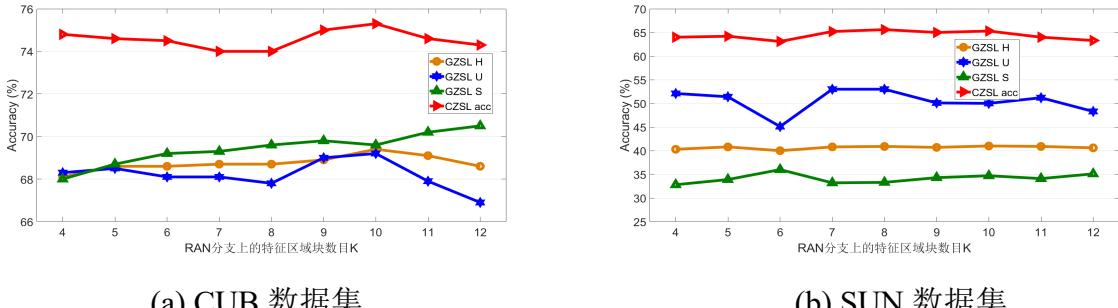
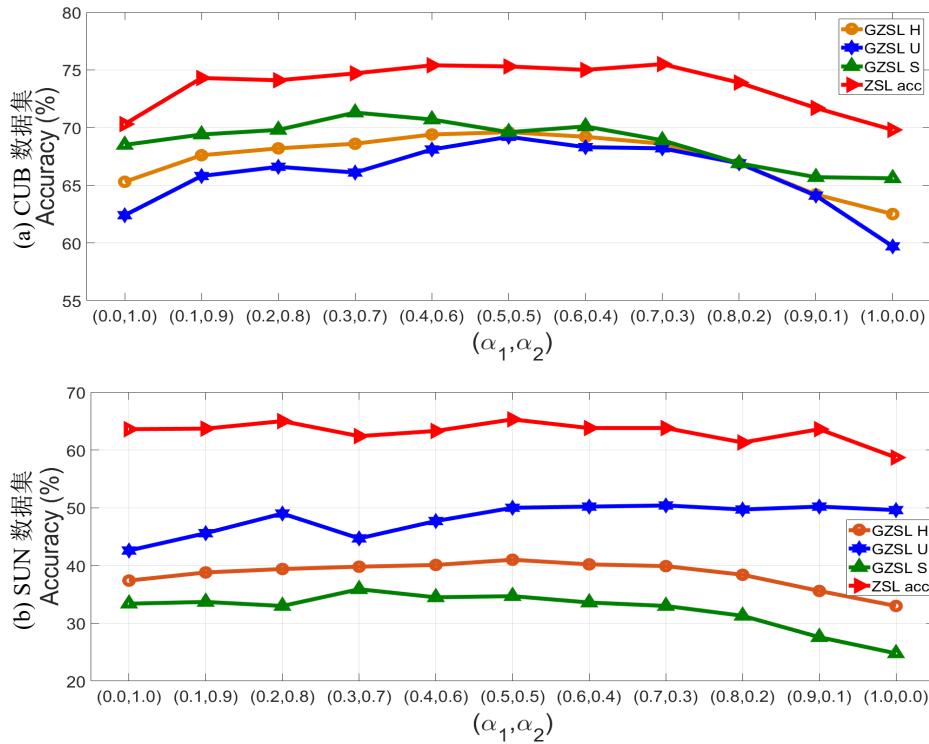
- 超参实验分析；
- 消融实验分析；
- 定性实验分析；
- GNDAN 和其他经典且先进的零样本图像分类方法在 CZSL/GZSL 不同设置下的实验结果对比。

接下来将对不同的实验进行详细阐述和分析。本章节方法 GNDAN 的源代码地址：  
<https://github.com/shiming-chen/GNDAN>。

**超参实验分析：** GNDAN 的主要超参数是 RAN 分支上的特征区域块数目  $K$ 、模型的组合系数  $(\alpha_1, \alpha_2)$ 、自校准损失权重  $\lambda_{SC}$ 。本文将对这三个超参数在 CUB 数据集和 SUN 数据集做相应的实验分析，并进行超参数设置。

RAN 分支上的特征区域块数目  $K$ : RAN 分支上的特征区域块数目  $K$  是 GNDAN 的一个重要超参数。我们对  $K$  进行一定范围内的取值（例如  $K = \{4, 5, 6, 7, 8, 9, 10, 11, 12\}$ ）并在 CUB 和 SUN 数据集上进行实验，如图3.4所示。实验结果显示，GNDAN 总体上对超参  $K$  不敏感。当  $K$  设置为 10 时，GNDAN 在 CUB 和 SUN 数据集上获得较好的结果。特别指出，当  $K$  的值太大时，已知类的识别精度会增加，而未知类的识别精度会降低。这意味着已知类和未知类识别精度之间的平衡被打破，导致 GNDAN 的泛化能力较差。因此，本文将  $K$  在所有数据集上均设置为 10。

模型的组合系数  $(\alpha_1, \alpha_2)$  : 为了确定 RAN 和 RGAT 子网络之间合适的组合系数  $(\alpha_1, \alpha_2)$ ，有效地增强 GNDAN 学习的特征表示。本文在一定范围内取值进行实验，例如， $\{(0.1, 0.9), (0.2, 0.8), (0.3, 0.7), (0.4, 0.6), (0.5, 0.5), (0.6, 0.4), (0.7, 0.3), (0.8, 0.2), (0.9, 0.1)\}$ 。


 图 3.4 RAN 分支上的特征区域块数目  $K$  对 GNDAN 的影响。

 图 3.5 模型的组合系数  $(\alpha_1, \alpha_2)$  对 GNDAN 的影响。

当  $\alpha_1, \alpha_2 = (1.0, 0.0)$  时，表示 GNDAN 只使用 RAN 学习的局部特征进行分类；当  $\alpha_1, \alpha_2 = (0.0, 1.0)$  时，表示 GNDAN 只使用 RGAT 学习的全局特征进行分类。如图3.5所示，当  $\alpha_1/\alpha_2$  被设置得太小或太大时，GNDAN 的性能很差，因为局部视觉特征和全局视觉特征的语义嵌入对增强视觉特征表示都很重要；当  $\alpha_1/\alpha_2$  非常大时，CZSL 和 GZSL 的所有评估指标均会下降，因为 RGAT 学习的显式全局视觉特征对

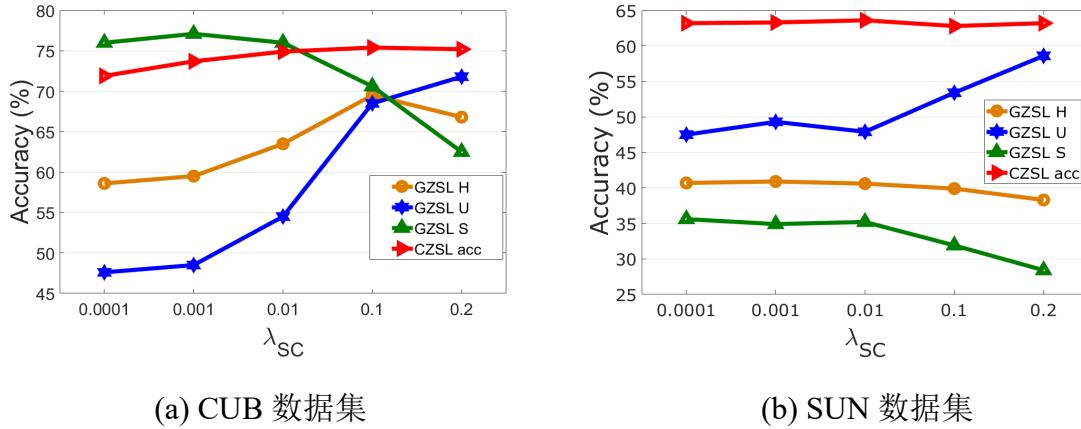


图 3.6 自校准损失权重 ( $\lambda_{SC}$ ) 对 GNDAN 的影响。

表 3.1 不同模型成分设置下，GNDAN 在 CUB 数据集<sup>[4]</sup> 和 SUN 数据集<sup>[5]</sup> 上的性能表现。

不同设置下的 GNDAN	CUB			SUN				
	acc	U	S	H	acc	U	S	H
GNDAN (无 RGAT)	69.8	59.7	65.6	62.5	58.7	49.6	24.8	33.0
GNDAN (无 RAN)	70.3	62.4	68.5	65.3	62.8	42.6	33.4	37.4
GNDAN (无图网络)	71.9	69.0	64.9	66.9	63.2	49.2	30.3	37.5
GNDAN (无自校准损失)	73.2	46.7	78.6	58.6	63.6	49.3	34.7	40.7
GNDAN	75.1	69.2	69.6	69.4	65.3	50.0	34.7	41.0

零样本图像分类中的知识迁移起到重要作用。

**自校准损失权重  $\lambda_{SC}$ :** 为了缓解已知类偏差问题<sup>[40,81]</sup>, 本文使用自校准损失进行模型优化, 并利用损失权重  $\lambda_{SC}$  平衡 GNDAN 的已知类和未知类识别精度。如图所示, 当  $\lambda_{SC}$  增加时, 已知类识别精度下降, 而未知类识别精度上升。这表明自校准损失有效缓解 GNDAN 的已知类偏差问题。然而,  $\lambda_{SC}$  也不能设置太大, 因为它会过度惩罚已知类的识别精度。

**消融实验分析:** 为验证 GADAN 的不同模型成分和 RGAT 中使用的不同图模型的有效性, 本文在 CUB 和 SUN 数据集上进行消融实验分析。

GADAN 的不同模型成分：本文通过评估 GNDAN 在没有 RAN 子网络、RGAT 子网络、图形注意网络、自校准损失约束等模型成分情况下，验证 GNDAN 的性能表现。如表3.1所示，当 GNDAN 只使用 RAN 学习的局部视觉特用于零样本图像分类时，其性能比其完整模型差，例如，CUB 数据集上的指标  $acc/H$  下降 5.3%/6.9%，

表 3.2 GNDAN 使用不同图模型的情况下，其在 CUB 数据集<sup>[4]</sup> 和 SUN 数据集<sup>[5]</sup> 上的分类性能表现。

GNDAN 使用不同的图模型	CUB 数据集				SUN 数据集			
	acc	U	S	H	acc	U	S	H
GNDAN（不使用图模型）	71.9	69.0	64.9	66.9	63.2	49.2	30.3	37.5
GNDAN+Chebyshev	74.9	68.4	70.4	69.4	63.5	52.5	30.2	38.4
GNDAN+GCN	75.0	68.8	69.4	69.1	64.2	47.7	35.2	40.5
GNDAN+GraphSAGE	73.9	65.6	71.9	68.6	64.8	51.4	33.8	40.8
GNDAN+GAT	75.1	69.2	69.6	69.4	65.3	50.0	34.7	41.0

SUN 数据集上的指标  $acc/H$  下降 6.6%/8.0。类似地，当只考虑 RGAT 学习到的全局视觉特征时，GNDAN 的性能也会下降。这些结果直观地表明，不同于现有方法只使用隐式的全局视觉特征局部特征进行零样本图像学习，GNDAN 联合局部视觉特征和显式全局视觉特征可以很好的缓解跨数据集偏差问题，从而增强视觉特征并促进模型准确的视觉-语义交互。此外，GNDAN 的自校准约束有助于解决在 GZSL 设置下的已知类偏差问题，提高从已知类到未知类的知识转移。图注意网络可以有效地利用属性区域图上基于属性的区域特征之间的关系对显式全局视觉特征进行表示，它可以帮助 DNGAN 在 CUB 和 SUN 数据集上的指标  $acc/H$  提升 3.2%/2.5% 和 2.1%/3.5%。

GADAN 使用不同的图模型：本文评估了多种典型图模型对 GNDAN 的影响，包括 Chebyshev<sup>[114]</sup>、GCN<sup>[115]</sup>、GraphSAGE<sup>[116]</sup> 和 GAT<sup>[117]</sup>。在实验中，本文统一将这些图模型设为两个隐藏层，层节点数为 256。结果如表3.2所示。实验结果直观地显示，所有的图模型都可以在一定程度上提高 GNDAN 的分类性能。例如，GNDAN 在分别使用 Chebyshev、GCN、GraphSAGE 和 GAT 的情况下，其性能指标  $acc/H$  分别提高了 2.5%/0.9%、2.2%/3.0%、1.7%/3.3% 和 2.7%/3.5%。这表明图模型有效地指导 GNDAN 挖掘基于属性的局部特征之间的显式关系，从而学习具有判别性的全局视觉特征，有效缓解跨数据集偏差造成视觉特征表示能力不足的问题。此外，当使用 GAT 时，GNDAN 在所有数据集上都表现最好。因此，本文将 GAT<sup>[117]</sup> 作为 GNDAN 的图模型。

**定性实验分析：**为直观地验证 GNDAN 的有效性，本文通过特征图可视化和视觉特征的 t-SNE 可视化<sup>[1]</sup> 进行定性实验分析。

**特征图可视化：**图3.7展示了 GNDAN 的 RAN 和 RGAT 子网络学习的特征图可

可视化。结果显示，RAN 可以通过抑制背景和冗余前景区域学习与类相关的局部特征，有效增强视觉特征。因此，当 GNDAN 只使用 RAN 时，GNDAN 在 CUB 上取得  $H = 62.5\%$  (如表3.1所示)。同时，RGAT 有效地学习基于属性的局部特征。GNDAN 基于这些基于属性的局部特征构造属性区域图，并利用图注意网络进行挖掘它们之间的关系，从而显式地学习具有判别性的全局视觉特征。为此，当 GNDAN 只使用 RGAT 子网络时，GADAN 在 CZSL 和 GZSL 设置中均取得了具有竞争性的结果 (例如，在 CUB 上的性能为  $acc/H = 70.3\%/65.3\%$ )。通过融合 RAN 学习的局部视觉特征和 RGAT 学习的全局视觉特征，有效缓解跨数据集偏差问题并增强视觉，进一步提高零样本图像分类从已知类到未知类有效的知识迁移，使得 GNDAN 取得更优异的零样本分类效果。

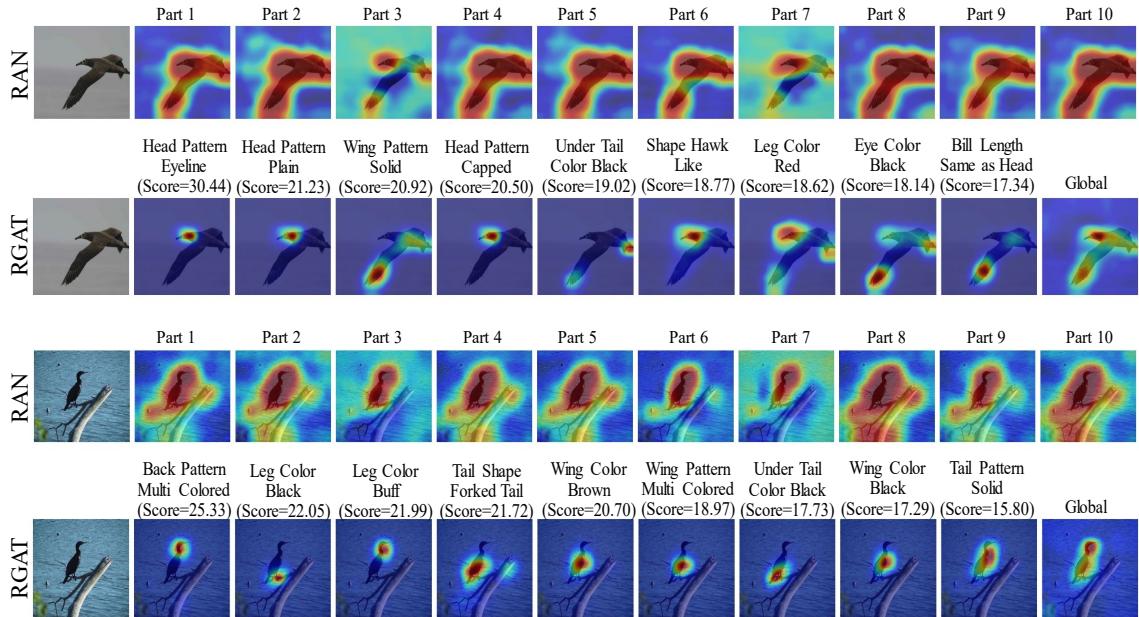


图 3.7 GNDAN 中 RAN 和 RGAT 子网络学习的特征图可视化。

视觉特征的 t-SNE 可视化：为直观地验证 GNDAN 缓解数据集偏差以增强视觉特征的有效性，本文对不同模型（例如，GNDAN 的两个子网络 RAN、RGAT 以及 GNDAN 的完整模型）在 CUB 数据集上随机选取的 20 个已知类和未知类学习的视觉特征进行 t-SNE<sup>[1]</sup> 可视化。结果显示，RAN 子网络可以学习具有良好类内紧致性的局部视觉特征，而 RGAT 子网络学习具有良好类间可分性的显式全局视觉特征。GNDAN 最终将这些互补的局部视觉特征和全局视觉特征的语义

嵌入向量融合，进一步增强视觉特征的表示，实现有效的视觉-语义交互。

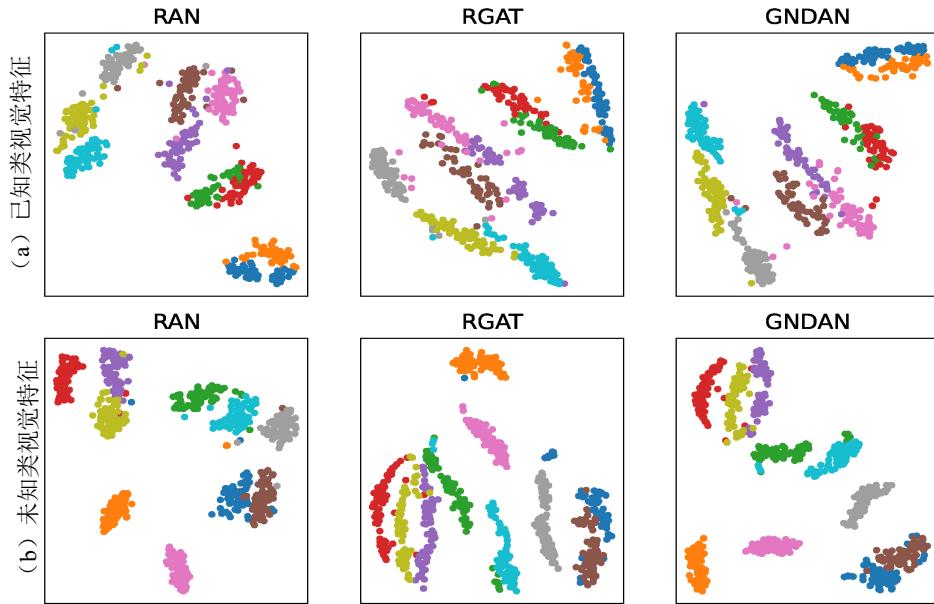


图 3.8 由 GNDAN 的子网络 RAN、RGAT 以及其完整模型在 CUB 数据集上学习的视觉特征 t-SNE<sup>[1]</sup>可视化。

**GNDAN 和当前先进的零样本图像分类方法的实验结果对比：**本文将 GNDAN 和其他先进的零样本图像分类方法在 CUB<sup>[4]</sup>、SUN<sup>[5]</sup>、AWA2<sup>[6]</sup> 数据集上的 CZSL/GZSL 实验结果进行对比，包括端到端的方法（例如，QFSL<sup>[72]</sup>、LDF<sup>[73]</sup>、AREN\*<sup>[2]</sup>、SGMA(NeurIPS'19)<sup>[40]</sup> 等）和非端到端的方法（例如，生成式方法（f-CLSWGAN<sup>[89]</sup>、cycle-CLSWGAN<sup>[93]</sup>、LsrGAN<sup>[102]</sup>、OCD-CVAE<sup>[125]</sup>、GCM-CF<sup>[123]</sup>、MGA-GAN<sup>[122]</sup> 等）和嵌入式方法（CMT<sup>[79]</sup>、ALE<sup>[67]</sup>、DEVISE<sup>[68]</sup>、ESZSL<sup>[70]</sup>、LATEM<sup>[100]</sup>、SP-AEN<sup>[108]</sup>、PQZSL<sup>[124]</sup>、DVBE<sup>[107]</sup>、DAZLE<sup>[81]</sup>、APN<sup>[41]</sup> 等））。

本文首先将 GNDAN 和当前先进的零样本图像分类方法在 CZSL 设置下进行比较。如表3.3所示，本文提出的嵌入式方法 GNDAN 在 CUB 和 SUN 数据集上的识别精度分别比当前先进的非端到端嵌入式方法至少高出 9.3% 和 5.8%。结果表明，得益于 GNDAN 基于图指导的双注意力网络融合的局部和显式全局视觉特征具有局部判别性，有效地区分各种细粒度类。对于粗粒度数据集（例如，AWA2），GNDAN 仍取得极高的识别性能，识别精度为 71.0%，明显优于其他非生成式方法。与端到端和生成方法相比，GNDAN 在 CUB 和 SUN 数据集上仍然获得最好的结果，准确率分别为 75.1% 和 65.3%，表明特征增强方法比微调技术表现更加。这进一步显示了

# 华 中 科 技 大 学 博 士 学 位 论 文

---

表 3.3 GNDAN 和和其他先进的零样本图像分类方法在 CUB<sup>[4]</sup>、SUN<sup>[5]</sup>、AWA2<sup>[6]</sup> 数据集上的 CZSL/GZSL 实验结果对比。

先进的零样本图像分类方法	CUB 数据集				SUN 数据集				AWA2 数据集			
	CZSL		GZSL		CZSL		GZSL		CZSL		GZSL	
	acc	U	S	H	acc	U	S	H	acc	U	S	H
<b>端到端的方法</b>												
QFSL <sup>[72]</sup>	58.8	33.3	48.1	39.4	56.2	30.9	18.5	23.1	63.5	52.1	72.8	60.7
LDF <sup>[73]</sup>	67.5	26.4	<b>81.6</b>	39.9	—	—	—	—	65.5	9.8	87.4	17.6
SGMA <sup>[40]</sup>	71.0	36.7	71.3	48.5	—	—	—	—	68.8	37.6	87.1	52.5
AREN <sup>[2]</sup>	71.8	38.9	78.7	52.1	60.6	19.0	38.8	25.5	67.9	15.6	92.9	26.7
LFGAA <sup>[74]</sup>	67.6	36.2	80.9	50.0	61.5	18.5	<b>40.0</b>	25.3	68.1	27.0	<b>93.4</b>	41.9
<b>非端到端的方法</b>												
<b>生成式方法</b>												
f-CLSWGAN <sup>[89]</sup>	57.3	43.7	57.7	49.7	60.8	42.6	36.6	39.4	68.2	57.9	61.4	59.6
cycle-CLSWGAN <sup>[93]</sup>	58.4	45.7	61.0	52.3	60.0	49.4	33.6	40.0	66.3	56.9	64.0	60.2
f-VAEGAN <sup>[61]</sup>	61.0	48.4	60.1	53.6	64.7	45.1	38.0	41.3	71.1	57.6	70.6	63.5
LsrGAN <sup>[102]</sup>	60.3	48.1	59.1	53.0	62.5	44.8	37.7	40.9	—	53.1	68.8	60.0
RFF-GZSL <sup>[94]</sup>	—	52.6	56.6	54.6	—	45.7	38.6	41.9	—	—	—	—
MGA-GAN <sup>[122]</sup>	58.9	46.6	58.5	51.8	61.8	45.6	37.3	41.0	70.6	59.3	67.7	63.2
GCM-CF <sup>[123]</sup>	—	61.0	59.7	60.3	—	47.9	37.8	42.2	—	<b>60.4</b>	75.1	67.0
<b>嵌入式方法</b>												
CMT <sup>[79]</sup>	34.6	7.2	49.8	12.6	39.9	8.1	21.8	11.8	37.9	0.5	90.0	1.0
ALE <sup>[67]</sup>	54.9	23.7	62.8	34.4	58.1	21.8	33.1	26.3	62.5	14.0	81.8	23.9
DEVISE <sup>[68]</sup>	52.0	23.8	53.0	32.8	56.5	16.9	27.4	20.9	59.7	17.1	74.7	27.8
ESZSL <sup>[70]</sup>	53.9	12.6	63.8	21.0	54.5	11.0	27.9	15.8	58.6	5.9	77.8	11.0
LATEM <sup>[100]</sup>	49.3	15.2	57.3	24.0	55.3	14.7	28.8	19.5	55.8	11.5	77.3	20.0
SP-AEN <sup>[108]</sup>	55.4	34.7	70.6	46.6	59.2	24.9	38.6	30.3	58.5	23.3	90.9	37.1
PQZSL <sup>[124]</sup>	—	43.2	51.4	46.9	—	35.1	35.3	35.2	—	31.7	70.9	43.8
IIR <sup>[75]</sup>	63.8	30.4	65.8	41.2	63.5	22.0	34.1	26.7	67.9	17.6	87.0	28.9
TCN <sup>[76]</sup>	59.5	52.6	52.0	52.3	61.5	31.2	37.3	34.0	<b>71.2</b>	61.2	65.8	63.4
DAZLE <sup>[81]</sup>	66.0	56.7	59.6	58.1	59.4	<b>52.3</b>	24.3	33.2	67.9	60.3	75.7	67.1
APN <sup>[41]</sup>	72.0	65.3	69.3	67.2	61.6	41.9	34.0	37.6	68.4	57.1	72.4	63.9
<b>GNDAN (本文方法)</b>	<b>75.1</b>	<b>69.2</b>	<b>69.6</b>	<b>69.4</b>	<b>65.3</b>	50.0	34.7	41.0	71.0	60.2	80.8	<b>69.0</b>

注：符号“—”表示相应结果缺失。

GNDAN 在有效缓解跨数据集偏差问题，提高零样本图像分类从已知类到未知类的知识迁移效果。

表3.3显示了 GZSL 设置中不同方法在 CUB<sup>[4]</sup>、SUN<sup>[5]</sup>、AWA2<sup>[6]</sup> 数据集上的实验结果。结果表明，在 CUB 和 AWA2 数据集上，所有方法的未知类识别率 ( $U$ ) 通常低于已知类识别率 ( $S$ )，即  $U < S$ 。同时，由于 SUN 数据集上的已知类的类别数量远远大于未知类的类别数量，使得  $U > S$ 。

表3.3也展示了 GNDAN 和其他先进的零样本图像分类方法在 GZSL 设置下的结

果对比。结果显示：在 CUB 和 AWA 数据集上，当前先进的非生成式方法在已知类上取得不错的性能，但在未知类上且取得较差的结果，而本文提出的 GNDAN 却同时在已知类和未知类取得很好的零样本分类性能。例如，GNDAN 在 CUB 和 AWA2 数据集上的比当前最先进的方法（例如 DAZLE<sup>[81]</sup>）在调和均值 ( $\mathbf{H}$ ) 这一指标上提高了 11.3% 和 1.9%。这表明 GNDAN 可以学习具有更具判别性和迁移性的视觉特征，促进零样本图像分类进行有效的知识迁移。此外，端到端的方法对 CNN Backbone 进行微调能提高零样本数据集视觉特征的判别性，然而会面临过拟合等问题（见章节3.1的讨论分析）；而本文基于特征增强零样本图像分类方法更有效地解决跨数据集偏差问题，从而取得更好的零样本分类效果。然而，GNDAN 的性能在 SUN 数据集上不如当前先进的生成式方法（例如，OCD-CVAE<sup>[125]</sup>、GCM-CF<sup>[123]</sup>），这是因为 SUN 数据集每个类平均只包含 16 个训练样本，严重限制了模型的学习。因此，生成式方法生成大量的伪样本进行数据增强，有效缓解 SUN 数据集数据量不足的问题，使得大多数生成方法的性能优于非生成式方法。

### 3.3 基于视觉特征精细化的生成式零样本图像分类

#### 3.3.1 研究动机

嵌入式零样本图像分类方法取得了一定的进展，但由于其只在已知类数据上学习相应的分类器，造成模型极易过拟合于已知类<sup>[88,89]</sup>。虽然研究者通过使用自校准机制<sup>[46]</sup> 人工地给未知类提高分类置信度，但还是不能解决未知样本缺失造成这种严重的已知类过拟合问题。针对此问题，生成式零样本图像分类方法被进一步提出，并逐渐成为当前的主流方法。它们基于生成模型（例如，变分自编码器<sup>[83]</sup>（Variational Auto-Encoder, VAE），生成对抗网络<sup>[84]</sup>（Generative Adversarial Network, GAN），标准化流<sup>[85,86]</sup>（Normalizing Flows）等）在已知类数据上学习一个语义 → 视觉的映射函数（即生成器），从而使用学习的生成器基于类语义向量（即类原型）对未知类生成大量的伪图像样本<sup>[88]</sup> 或者伪特征样本<sup>[89,90]</sup>，随后将这些未知类伪样本和已知类训练样本一起训练一个有监督的分类器，从而将零样本图像分类问题转化为有监督的分类问题。由于生成图像样本比特征样本的难度更大，使得特征生成式零样本图像分类方法成为当前主要的研究方向。本章节也研究特征生成式零样本图像分类方法。

由于跨数据集偏差问题，现有生成式零样本图像分类方法使用表示能力不足

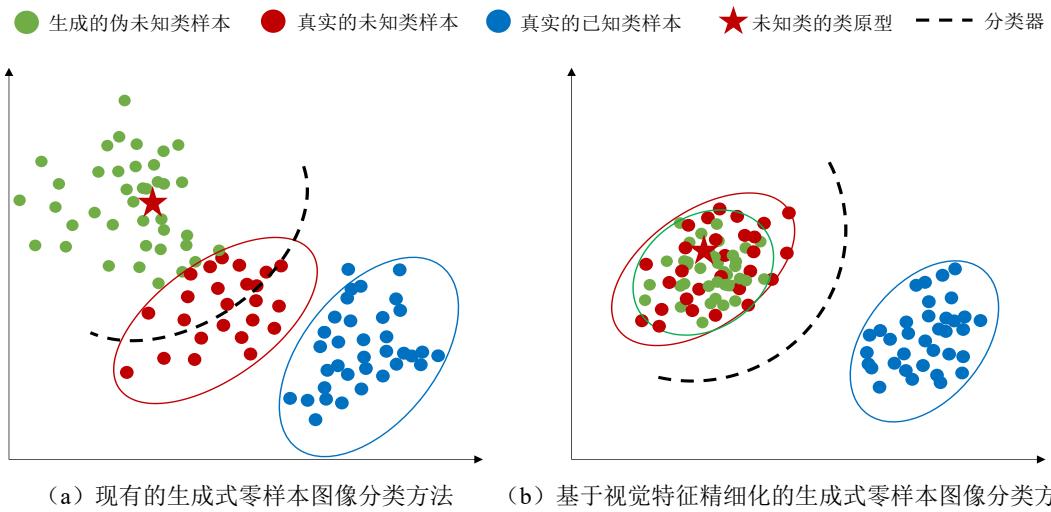


图 3.9 现有的生成式零样本图像分类方法和本文基于视觉特征精细化的生成式零样本图像分类方法对比。

的全局视觉特征学习一个不准确的生成器（语义  $\rightarrow$  视觉映射函数），造成生成的伪未知类样本和真实的未知类样本具有明显的分布偏差，使得基于生成的未知类样本和真实的已知类样本训练的分类器在测试时对大部分真实的未知类样本（即测试样本）错误分类，如图3.9(a) 所示。因此，现有的大部分生成式零样本学习方法<sup>[7,61,89,93–95,102,106,122,125,126]</sup> 在已知类取得较高的分类精度，在未知类上取得明显较低的分类精度。为此，跨数据集偏差问题在生成式零样本图像分类中也是一个急需解决的重要问题。

针对此问题，本文提出一种基于视觉特征精细化的生成式零样本图像分类方法（Visual Feature Refinement for Zero-Shot Image Classification, ViFR）。给定一个现有的生成模型，ViFR 引入了前置特征细化模块（Pre-Feature-Refinement, Pre-FR）和后置特征细化模块（Post-Feature-Refinement, Post-FR），同时精细化视觉特征实现特征增强。在 Pre-FR 模块中，在基于属性的交叉熵损失指导下，ViFR 使用属性引导的注意机制进行基于属性的视觉定位，学习属性相关的局部视觉特征作为前置精细化视觉特征。随后将前置精细化视觉特征输入到生成模型（即语义  $\rightarrow$  视觉映射）中学习视觉特征生成。在 Post-FR 模块中，ViFR 引入一个视觉  $\rightarrow$  语义映射函数，该映射和生成模型中的语义  $\rightarrow$  视觉映射合进行联合优化，形成环一致性学习。在自适应间隔中心损失与语义循环一致性损失的共同约束下，Post-FR 学习具有类相关和语义相关的视觉特征表示，并通过融合 Post-FR 中不同层的特征（包括输入层的视觉特征、中

间层的隐含特征、输出层的语义特征)提取完全精细化的视觉特征,实现视觉特征增强。为此,VIFR生成的伪视觉样本与真实的未知类样本具有更一致的特征分布,从而学习更有效的有监督分类器实现准确的零样本分类,如图3.9(b)所示。在三个主流标准数据集(例如,CUB数据集<sup>[4]</sup>、SUN数据集<sup>[5]</sup>和AWA2数据集<sup>[6]</sup>)上的大量实验表明,本文的ViFR有效解决生成式零样本图像分类中的跨数据集偏差问题,并取得领先的零样本分类性能。

### 3.3.2 视觉特征精细化学习

本章节使用的基本符号定义见章节3.2.2。ViFR的模型结构如图3.10所示。ViFR由前置特征细化模块(Pre-FR)、特征生成模型(f-VAEGAN<sup>[61]</sup>)、后置特征细化模块(Post-FR)和分类器组成。ViFR使用CNN Backbone(例如,ResNet-101<sup>[31]</sup>)为每个图像提取一组区域视觉特征 $\mathcal{R}_P = \{r_p\}_{p=1}^P$ ,并使用语言模型(例如,GloVe<sup>[60]</sup>为 $A$ 个属性提取属性特征 $\mathcal{V}_A = \{v_a\}_{a=1}^A$ ,并将它们作为模型输入。ViFR学习增强的视觉特征进行零样本图像分类包含三个学习阶段。第一阶段,Pre-FR在基于属性的交叉熵损失函数的约束下,使用属性指导的注意机制进行视觉的属性定位学习属性增强视觉特征进行视觉特征的前置精细化,从而实现视觉特征的预增强。第二阶段,基于Pre-FR的前置精细化视觉特征,f-VAEGAN学习语义→视觉映射(即特征生成器 $G$ )进行视觉特征生成。同时,在自适应间隔中心损失与语义循环一致性损失的共同约束下,Post-FR学习准确的视觉→语义映射函数,使得Post-FR的不同层具有类相关和属性语义相关的特征表示,并通过融合Post-FR中不同层的特征提取完全精细化的视觉特征,实现视觉特征的进一步增强。特别指出,f-VAEGAN和Post-FR在第二阶段进行统一训练,有效促进生成模型的学习。第三阶段,使用 $\mathcal{D}^{tr}$ 中完全精细化的已知类视觉特征和完全精细化的伪未知类特征样本(使用 $G$ 进行特征生成)一起训练一个有监督的分类器(例如,softmax分类器),而 $\mathcal{D}^{te}$ 中已知类和未知类的完全精细化视觉特征用于模型测试。ViFR在模型优化时不再对CNN Backbone和语言模型中的参数做进一步训练,所以ViFR是一个非端到端方法。此外,ViFR在训练过程中没有使用真实的未知类视觉特征,因此它是一种归纳式方法。接下来将对具体模型和模型优化过程进行详细介绍。

**前置特征精细化模块:**考虑到学习细粒度属性信息可有效提高视觉特征表示<sup>[2,39-41]</sup>,ViFR在学习生成模型之前先将视觉特征进行前置精细化,实现视觉特征的预增强。

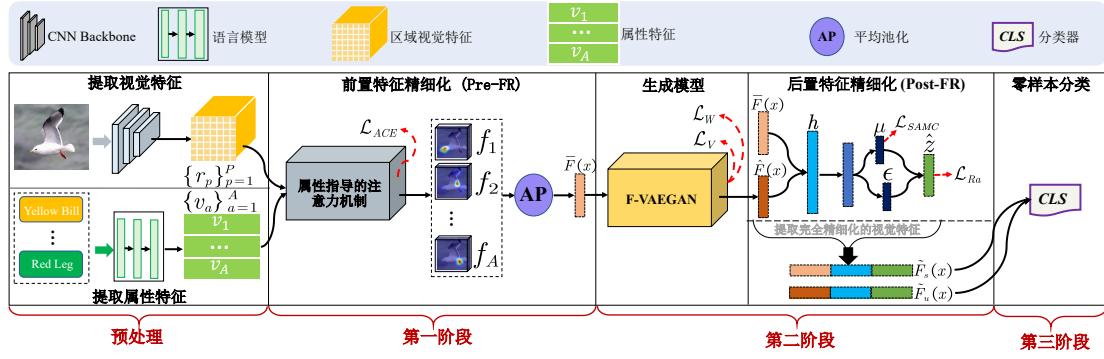


图 3.10 本章提出的 ViFR 模型结构示意图。

为了实现此目的，ViFR 采用了一种基于属性指导的注意力机制，将最相关的属性在相应的图像区域进行定位，从而学习每个属性的属性增强视觉特征。具体而言，Pre-FR 在属性特征  $\mathcal{V}_A = \{v_a\}_{a=1}^A$  的指导下，使模型关注每个属性相应的视觉区域特征  $\mathcal{R}_P = \{r_p\}_{p=1}^P$ ，并将每个属性与相应关注的视觉区域特征进行比较，实现属性在视觉中的定位。Pre-FR 对第  $a$  个属性进行视觉的特征表示定义为：

$$f_a(x) = \sum_{p=1}^P \frac{\exp(v_a^\top W_1 r_p)}{\sum_{p=1}^P \exp(v_a^\top W_1 r_p)} r_p, \quad (3.18)$$

其中  $W_1$  是一个可学习的矩阵， $\frac{\exp(v_a^\top W_1 r_p)}{\sum_{p=1}^P \exp(v_a^\top W_1 r_p)}$  表示图像中第  $p$  个区域与第  $a$  个属性的注意力权重（即相关关系）。如果图像的相关区域具有突出的视觉属性  $a$ ，Pre-FR 将为  $a$  属性分配一个较高的注意力值，否则，将分配一个较小的注意力值。为此，Pre-FR 根据给定一组属性为图像  $x$  学习一组属性增强的视觉特征  $F(x) = \{f_a(x)\}_{a=1}^A$ 。

为了使 Pre-FR 能够有效地学习这些属性增强的视觉特征，ViFR 将它们进一步嵌入到类语义空间中进行优化。具体而言，ViFR 首先将每个属性增强的视觉特征  $f_a(x)$  与其相关的属性特征  $v_a$  进行匹配实现视觉特征的语义嵌入：

$$f'_a(x) = v_a^\top W_2 f_a(x), \quad (3.19)$$

其中， $W_2$  是一个可学习的映射矩阵。因此，ViFR 可获得图像  $x$  类似于类语义向量  $z^c = [z_1^c, \dots, z_A^c]^\top$  的语义嵌入特征  $F'(x) = \{f'_a(x)\}_{a=1}^A$ 。然后，利用基于属性的交叉熵损失<sup>[41,81]</sup> 实现对 Pre-FR 的优化：

$$\mathcal{L}_{ACE} = -\frac{1}{n_b} \sum_{i=1}^{n_b} \log \frac{\exp(F'(x_i) \times z^c)}{\sum_{\hat{c} \in \mathcal{C}} \exp(F'(x_i) \times z^{\hat{c}})}, \quad (3.20)$$

其中  $n_b$  表示批量大小。

当 Pre-FR 的进行训练之后, ViFR 使用平均池化 (Average Pooling, AP) 将所有训练样本和测试样本的属性增强视觉特征压缩为具有判别性的前置精细化视觉特征  $\bar{F}(x)$ , 并用于后续模型的训练和分类测试, 表示为:

$$\bar{F}(x) = AP(F(x)). \quad (3.21)$$

**特征生成模型:** ViFR 的框架较为灵活, 它可以基于任何现有的生成式零样本图像分类模型 (例如, f-VAEGAN<sup>[61]</sup>, TF-VAEGAN<sup>[7]</sup> 等) 进行视觉特征生成。考虑到 f-VAEGAN<sup>[61]</sup> 取得了很好的性能, 并已成为生成式零样本图像分类方法的重要基准方法。为此, ViFR 使用 f-VAEGAN 作为特征生成器学习语义  $\rightarrow$  视觉映射, 为未知类生成伪特征样本。为了使生成模型稳定且有效, f-VAEGAN 将 VAE 和 GAN 合并到统一的生成模型中。f-VAEGAN 由特征生成 VAE (f-VAE) 和特征生成 WGAN (f-WGAN) 组成。f-VAE 包括编码器  $E(\bar{F}(x), z^c)$  (表示为  $E$ ) 和解码器  $G$  (与 f-WGAN 共享, 作为以类语义向量  $z^c$  作为指导的条件生成器  $G(o, z^c)$ )。编码器  $E(\bar{F}(x), z^c)$  将前置精细化的已知类视觉特  $\bar{F}(x)$  编码为隐含编码  $o$ , 而解码器  $G(o, z^c)$  从  $z$  重建视觉特征  $\hat{F}(x)$ 。f-VAE 首先通过 VAE 损失  $\mathcal{L}_V$  进行模型优化:

$$\begin{aligned} \mathcal{L}_V &= \mathcal{L}_{KL} + \mathcal{L}_{R\_x} \\ &= \text{KL}(E(\bar{F}(x), z^c) \| p(o|z^c)) - \mathbb{E}_{E(\bar{F}(x), z^c)}[\log G(o, z^c)], \end{aligned} \quad (3.22)$$

其中,  $\mathcal{L}_{KL}$  是 Kullback-Leibler 散度,  $p(o|z^c)$  是假定为  $\mathcal{N}(0, 1)$  的先验分布,  $\mathcal{L}_{R\_x}$  是视觉特征重建损失并表示为  $-\log G(o, z^c)$ 。另一方面, f-WGAN 由生成器  $G(o, z^c)$  和鉴别器  $D(x, z^c)$  (表示为  $D$ ) 组成。生成器  $G$  通过随机输入噪声  $o$  生成视觉特征  $\hat{F}(x)$ , 而鉴别器  $D$  将真实视觉特征  $\bar{F}(x)$  或生成的伪视觉特征  $\hat{F}(x)$  作为输入, 并输出一个实数指示输入特征判别为真或假。 $G$  和  $D$  都以类语义向量  $z^c$  为条件指导, 使用 WGAN 损失进行优化:

$$\begin{aligned} \mathcal{L}_W &= \mathbb{E}[D(\bar{F}(x), z^c)] - \mathbb{E}[D(\hat{F}(x), z^c)] \\ &\quad - \lambda \mathbb{E}\left[\left(\|\nabla D(F^\flat(x), z^c)\|_2 - 1\right)^2\right], \end{aligned} \quad (3.23)$$

其中,  $F^\flat(x) = \tau \bar{F}(x) + (1 - \tau) \hat{F}(x)$ ,  $\tau \sim U(0, 1)$ ,  $\lambda$  是惩罚系数并设置为 10。为了解 f-VAEGAN 细节, 请参考 Xian 等人<sup>[61]</sup> 的原文。

**后置特征精细化模块:** 为了进一步增强视觉特征缓解跨数据集偏差，并促进零样本图像分类有效的知识迁移，ViFR 集成了一个后置特征精细化模块（Post-FR）。在自适应间隔中心损失与语义循环一致性损失的共同约束下，Post-FR 学习准确的视觉 → 语义映射函数，使得 Post-FR 的不同层具有类相关和属性语义相关的特征表示，并通过融合 Post-FR 中不同层的特征（包括输入层的视觉特征、中间层的隐含特征、输出层的语义特征）提取完全精细化的视觉特征，实现视觉特征的进一步增强。特别指出，f-VAEGAN 和 Post-FR 进行联合训练，促进生成器学习更具语义表示的视觉特征。

**自适应间隔中心损失:** 为了促使 Post-FR 有效地学习类相关的视觉特征表示，本章节提出了自适应边缘中心损失（Self-Adaptive Margin Center Loss,  $\mathcal{L}_{SAMC}$ ）对 Post-FR 进行优化。 $\mathcal{L}_{SAMC}$  的提出主要考虑到以下四个因素：(1) 由于类标签信息可用， $\mathcal{L}_{SAMC}$  显式地鼓励类内紧性和类间可分性，从而引导 Post-FR 学习判别性的类相关特征；(2)  $\mathcal{L}_{SAMC}$  同时具有中心损失<sup>[127]</sup> 和三元组损失<sup>[128]</sup> 的优点，例如，避免人工构造样本对且同时学习类内紧性和类间可分性；(3) 考虑到类内紧致性和类间可分性对各种粒度数据集（例如，粗粒度和细粒度数据集）的敏感性不同， $\mathcal{L}_{SAMC}$  采用一个平衡因子 ( $\gamma$ ) 自适应地平衡视觉特征的类间可分性和类内紧致性。较大的  $\gamma$  值使得  $\mathcal{L}_{SAMC}$  在细粒度数据集上（例如，CUB 数据集和 SUN 数据集）主要通过调整类内距离对细粒度相似类别加以区分，较小的  $\gamma$  值使得  $\mathcal{L}_{SAMC}$  在粗粒度数据集上（例如，AWA2 数据集）主要通过调整类间距对混淆的类进行区分；(4)  $\mathcal{L}_{SAMC}$  直接作用于 Post-FR 中的中间层编码特征  $\mu$ ，这有利于提高 Post-FR 中间层特征的判别性表示。 $\mathcal{L}_{SAMC}$  的形式化定义如下：

$$\begin{aligned} \mathcal{L}_{SAMC} = \max & \left( 0, \Delta + \gamma \|\mu - c_y\|_2^2 \right. \\ & \left. - (1 - \gamma) \|\mu - c_{y'}\|_2^2 \right), \end{aligned} \quad (3.24)$$

其中  $c_y$  是类别  $y$ （即已知类样本  $x$  的标签）的类中心， $c_{y'}$  是类别  $y'$ （即不属于标签  $y$  的其他随机选择的标签）的类中心， $\Delta$  用于控制类内距离和类间距离的间距， $\mu$  是 Post-FR 中的编码特征， $\gamma \in [0, 1]$  用于平衡类内紧性和类间可分性。

**语义环一致性损失:** Post-FR 的最后一层使用重参数化<sup>[83]</sup> 将生成的伪视觉特征  $\hat{F}(x)$  或真实的视觉特征  $\bar{F}(x)$  重建为相应的类语义特征表示  $\hat{z}$ 。为了进一步指导 Post-FR 有效地学习语义相关的表示实现语义重构，本文进一步使用语义循环一致性

# 华 中 科 技 大 学 博 士 学 位 论 文

---

损失<sup>[7,93]</sup> (Semantic Cycle-Consistency Loss,  $\mathcal{L}_{Ra}$ ) 对 Post-FR 进行优化。不同于 Felix 等人<sup>[93]</sup> 使用  $\mathcal{L}_{Ra}$  只学习语义 → 视觉映射 ( $G$ ) 以生成未知类的伪视觉特征, 本文使用  $\mathcal{L}_{Ra}$  同时学习  $G$  和 Post-FR 用于生成未知类的伪视觉特征和视觉特征精细化。语义环一致性损失  $\mathcal{L}_{Ra}$  使用  $\ell_1$  重构函数实现:

$$\mathcal{L}_{Ra} = \mathbb{E} [\|\hat{z}_{real}^c - z^c\|_1] + \mathbb{E} [\|\hat{z}_{syn}^c - z^c\|_1], \quad (3.25)$$

其中,  $\hat{z}_{real}^c$  是 Post-FR 从真实的视觉特征  $\bar{F}(x)$  重构的类语义特征,  $\hat{z}_{syn}^c$  是 Post-FR 从生成的伪视觉特征  $\hat{F}(x)$  重构的类语义特征,  $z^c$  是与视觉特征  $\bar{F}(x)$  或  $\hat{F}(x)$  相关的类语义向量.

提取完全精细化的视觉特征: Post-FR 训练完成后, ViFR 使用 Post-FR 为真实的已知类视觉特征  $\bar{F}(x_s)$  和真实/合成的未知类视觉特征  $\bar{F}(x_u)/\hat{F}(x_u)$  提取相应完全精细化的视觉特征  $\tilde{F}(x_s)$  和  $\tilde{F}(x_u)$ , 从而缓解跨数据集偏差问题, 实现视觉特征增强。考虑到以下两个因素: (1) 高维视觉特征映射到低维语义特征空间, 不可避免地造成信息损失; (2) 由于视觉概念特征 (即从图像中学习的视觉表示) 对粗粒度数据集的样本分类产生更大的作用, 而语义属性对细粒度数据集的样本分类产生更大的作用。ViFR 利用残差学习思想<sup>[31]</sup> 将前置精细化的真实视觉特征/生成模型生成的伪视觉特征 (例如,  $\bar{F}(x_s)/\hat{F}(x_u)$ )、Post-FR 中相应的中间层特征 ( $h_s/h_u \in \mathcal{H}$ ) 和重构的类语义特征 ( $\hat{z}_s/\hat{z}_u \in \mathcal{Z}$ ) 进行融合, 从而得到具有类相关和语义相关的完全精细化视觉特征:

$$\tilde{F}(x_s^{tr}) = \bar{F}(x_s^{tr}) \oplus h_s^{tr} \oplus \hat{z}_s^{tr}, \quad (3.26)$$

$$\tilde{F}(x_s^{te}) = \bar{F}(x_s^{te}) \oplus h_s^{te} \oplus \hat{z}_s^{te}, \quad (3.27)$$

$$\tilde{F}(x_u^{te}) = \bar{F}(x_u^{te}) \oplus h_u^{te} \oplus \hat{z}_u^{te}, \quad (3.28)$$

$$\tilde{F}(x_u^{syn}) = \hat{F}(x_u^{syn}) \oplus h_u^{syn} \oplus \hat{z}_u^{syn}, \quad (3.29)$$

其中,  $\oplus$  是特征融合操作,  $h_s^{tr} \cup h_s^{te} = h_s$ ,  $h_u^{syn} \cup h_u^{te} = h_u$ ,  $\hat{z}_s^{tr} \cup \hat{z}_s^{te} = \hat{z}_s$ ,  $\hat{z}_u^{syn} \cup \hat{z}_u^{te} = \hat{z}_u$ ,  $\tilde{F}(x_s^{tr}) \cup \tilde{F}(x_s^{te}) \cup \tilde{F}(x_u^{te}) \cup \tilde{F}(x_u^{syn}) = \tilde{\mathcal{F}}$ ,  $\tilde{\mathcal{F}}$  是完全精细化的视觉特征域, 用于零样本图像分类。完全精细化的真实已知类视觉特征  $\tilde{F}(x_s^{tr})$  和完全精细化的伪未知类视觉特征  $\tilde{F}(x_u^{syn})$  用于训练分类器, 而测试集样本的已知类完全精细化的视觉特征  $\tilde{F}(x_s^{te})$  和未知类精细化视觉特征  $\tilde{F}(x_u^{te})$  用于零样本图像分类测试。

**模型优化:** ViFR 的模型优化包含三个阶段: 前置特征精细化模块优化 (第一阶段)、

# 华中科技大学博士学位论文

生成模型与后置特征精细化模块优化（第二阶段）、零样本图像分类器优化（第三阶段）。

在第一阶段，ViFR 使用基于属性的交叉熵损失对 Pre-FR 进行训练优化：

$$\mathcal{L}_{\text{Stage1}}(\text{Pre-FR}) = \mathcal{L}_{\text{ACE}}. \quad (3.30)$$

随后，将完成训练的 Pre-FR 对训练集和测试集中的特征样本进行前置特征精细化为  $\bar{F}$  并用于第二阶段的模型学习。

在第二阶段，ViFR 联合训练生成模型和 Post-FR 模块，具体优化目标为：

$$\begin{aligned} \mathcal{L}_{\text{Stage2}}(\mathbf{E}, \mathbf{G}, \mathbf{D}, \text{Post-FR}) &= \mathcal{L}_V + \mathcal{L}_W \\ &\quad + \lambda_{SAMC} \mathcal{L}_{SAMC} + \lambda_{Ra} \mathcal{L}_{Ra}, \end{aligned} \quad (3.31)$$

其中， $\lambda_{SAMC}$  和  $\lambda_{Ra}$  为损失权重，用于控制相应的损失项对整个 ViFR 模型的影响。完成训练后，ViFR 使用生成器  $G$  根据未知类样本的类语义向量生成足够的未知类伪视觉特征，即  $\tilde{F}(x_u^{syn} = G(o, z^u))$ 。随后，ViFR 使用 Post-FR 对前置精细化的真实视觉特征和生成的未知类伪视觉特征再次进行特征精细化，提取完全精细化的视觉特征用于零样本图像分类。

在第三阶段，使用完全精细化的已知类视觉特征和生成的未知类伪视觉特征训练一个有监督分类器（例如，Softmax）。

$$\mathcal{L}_{\text{Stage3}}(f_{CZSL}/f_{GZSL}) = -\frac{1}{n_b} \sum_{i=1}^{n_b} \sum_{c=1}^{C^u/C} \mathbb{I}_{[c_y=y_i]} \log p(c_y | \tilde{F}(x_i)). \quad (3.32)$$

$f_{CZSL}/f_{GZSL}$  分别表示传统的零样本图像分类器和广义的零样本图像分类器。最后将训练好的分类器对测试集样本的已知类/未知类完全精细化视觉特征  $\tilde{F}(x_s^{te})/\tilde{F}(x_u^{te})$  进行测试。

### 3.3.3 实验结果与分析

为验证 ViFR 的有效性，本文在三个主流的零样本图像分类标准数据集上进行实验，包括两个细粒度数据集（CUB<sup>[4]</sup>, SUN<sup>[5]</sup>）和一个粗粒度数据集（AWA2<sup>[6]</sup>）。本文依据 Xian 等人<sup>[6]</sup> 最新的数据集划分方式进行模型训练和测试，数据集的具体介绍见章节2.3.1。在 ViFR 中，编码器 ( $E$ )、生成器 ( $G$ ) 和鉴别器 ( $D$ ) 均是多层感知机，包含一个 4096 个单元的隐藏层，并跟随一个 LeakyReLU 激活函数<sup>[129]</sup>。Post FR 也

# 华中科技大学博士学位论文

是 MLP 结构，它有两个隐藏层（单元数分别为 4096 和  $2 \times A$ ）且都是有 LeakyReLU 激活函数进行特征非线性变换，第二个隐藏层被编码为两个大小均为  $A$  的隐特征向量（分别是均值  $\mu$  和方差  $\delta$ ）。它的输出层通过重参数化<sup>[83]</sup>对隐特征向量重构为类语义向量。本文使用 Adam 优化器<sup>[130]</sup> ( $\beta_1=0.5$ ,  $\beta_2=0.999$ ) 对模型进行优化。根据实验结果，本文将公式3.31中的损失权重  $\lambda_{SAMC}$  和  $\lambda_{Ra}$  在细粒度数据集（例如，CUB 和 SUN）/粗粒度数据集（例如，AWA2）分别设置为 0.1/0.05 和 0.1/ $1e^{-5}$ 。对于细粒度和粗粒度数据集，公式3.24中的平衡因子  $\gamma$  分别设置为 0.9 和 0.3。

本章节通过以下实验验证 ViFR 的有效性：

- 超参实验分析；
- 消融实验分析；
- 定性实验分析；
- 通用性实验分析；
- ViFR 和其他经典且先进的零样本图像分类方法在 CZSL/GZSL 不同设置下的实验结果对比。

接下来将对不同的实验进行详细阐述和分析。本章节方法 ViFR 的源代码地址：  
<https://github.com/shiming-chen/FREE>, <https://github.com/shiming-chen/ViFR>。

**超参实验分析：**模型 ViFR 的主要超参数有五个，分别为输入图像的分辨率大小、公式3.24中的平衡因子  $\gamma$ 、公式3.31中的损失权重  $\lambda_{SAMC}$  和  $\lambda_{Ra}$ 、未知类伪视觉特征样本的数量  $N_{syn}$ 。本章节将对这四个超参数在 CUB 数据集和 AWA2 数据集做相应的实验分析，并进行超参数设置。

表 3.4 不同的输入图像分辨率对 ViFR 的影响。

基于不同输入图像分别率的 ViFR	CUB 数据集			AWA2 数据集		
	CZSL	GZSL		CZSL	GZSL	
		U	S		U	S
f-VAEGAN <sup>224[61]</sup>	61.0	48.3	58.9	53.1	71.2	53.7 76.2 63.0
ViFR <sup>224</sup>	69.1 <sup>↑7.9</sup>	57.8	62.7	60.1 <sup>↑7.0</sup>	73.7 <sup>↑2.6</sup>	58.4 81.4 68.0 <sup>↑5.0</sup>
f-VAEGAN <sup>448[61]</sup>	64.0	48.1	61.7	54.1	75.0	56.9 84.2 67.9
ViFR <sup>448</sup>	74.5 <sup>↑10.5</sup>	63.9	72.0	67.7 <sup>↑13.6</sup>	77.8 <sup>↑2.8</sup>	68.2 78.9 73.2 <sup>↑5.3</sup>

注：“224”和“448”分别表示模型的输入图像分辨率为  $224 \times 224$  和  $448 \times 448$ 。

**输入图像分辨率大小：**现有先进方法通常会使用将原始图像重新裁剪为不同分辨率的图像进行模型学习，例如  $224 \times 224$  和  $448 \times 448$ 。为此，本文对这两种分辨率

的图像作为模型输入在 CUB 和 AWA2 数据集上进行实验。如表3.4所示，输入图像为  $448 \times 448$  时，基准模型 f-VAEGAN<sup>[61]</sup> 和 ViFR 均能取得更好的零样本图像分类性能识别性能。然而，在这两种不同的分辨率图像作为输入的情况下，ViFR 均能对基准模型 f-VAEGAN<sup>[61]</sup> 产生显著的性能提升。例如，在图像分辨率输入为  $224 \times 224$  时，ViFR 相较于 f-VAEGAN<sup>[61]</sup> 在 CUB 和 AWA2 数据集分别在性能指标  $acc/H$  上提升 7.9%/7.0% 和 2.6%/5.0%；当输入图像分辨率为  $448 \times 448$  时，ViFR 相较于 f-VAEGAN<sup>[61]</sup> 在 CUB 和 AWA2 数据集分别在性能指标  $acc/H$  上提升 10.5%/13.6% 和 2.8%/5.3%。这结果表明，ViFR 通过特征精细化有效增强视觉特征，从而缓解跨数据集偏差问题并取得较好的零样本图像分类结果。为此，本文将输入图像的分辨率重新裁剪为  $448 \times 448$ 。

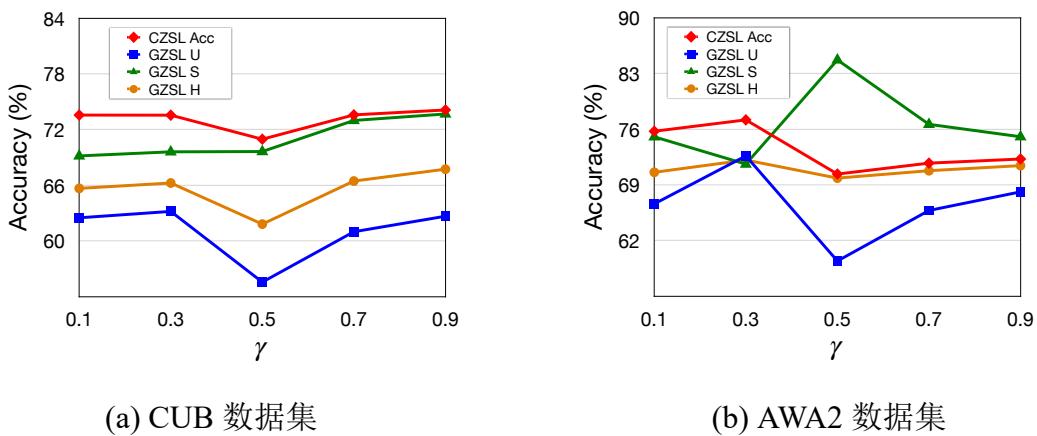
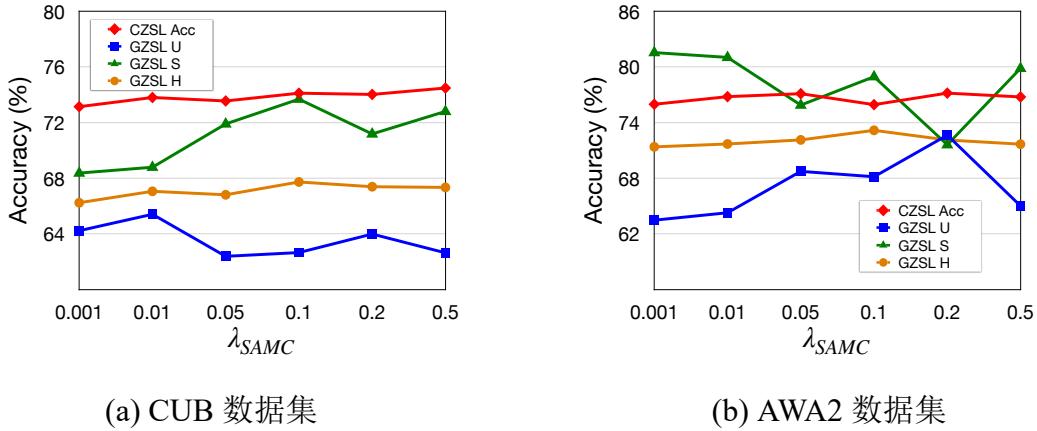
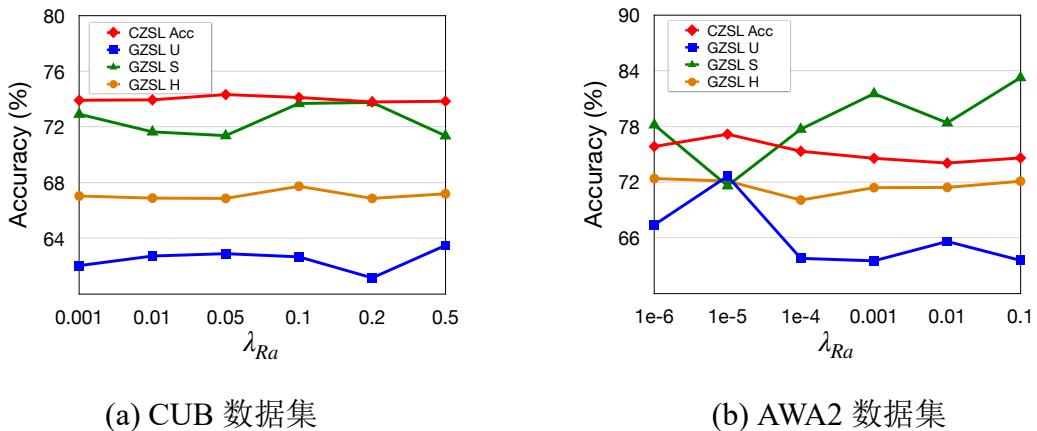


图 3.11 平衡因子  $\gamma$  对 ViFR 模型的影响。

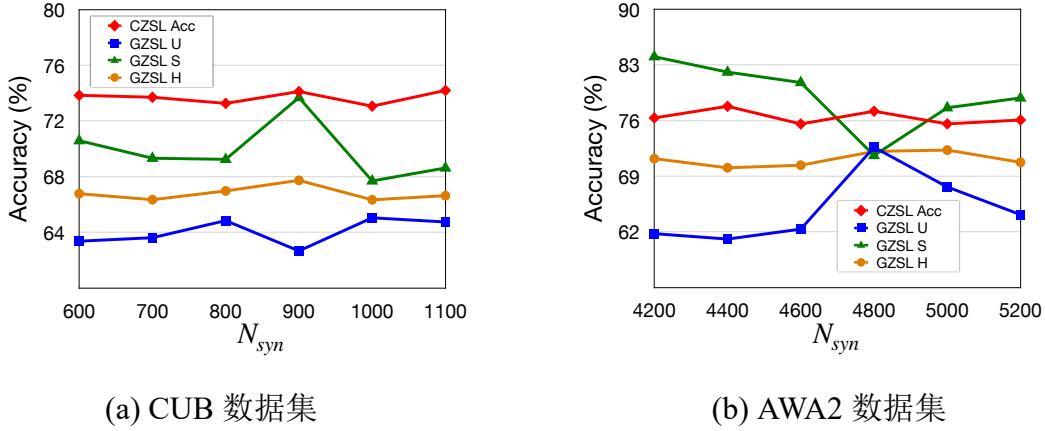
平衡因子  $\gamma$ : 本文通过研究公式3.24中的平衡因子  $\gamma$ ，以确定其对 ViFR 的影响。如图3.11所示，随着  $\gamma$  的增长，零样本图像分类性能指标  $acc$  和  $H$  在细粒度数据集（例如，CUB 数据集）上获得了一致的提高，而在粗粒度数据集（例如 AWA2）上的性能将持续下降。这些结果表明：(1) 在细粒度数据集上，当类混淆时，增加类内紧性可以更有效地对这些类进行区分；(2) 在粗粒度数据集上，增加类间可分性显著提高视觉特征判别性。根据实验结果，本文将 ViFR 的超参数  $\gamma$  在细粒度和粗粒度数据集上分别设置为 0.9 和 0.3。

损失权重  $\lambda_{SAMC}$  和  $\lambda_{Ra}$ : 本章节分析了公式3.31中的损失权重  $\lambda_{SAMC}$  和  $\lambda_{Ra}$  对 ViFR 的影响。如图3.12和图3.13所示，ViFR 的评估指标  $acc$  和  $H$  对损失权重  $\lambda_{SAMC}$


 图 3.12 损失权重  $\lambda_{SAMC}$  对 ViFR 模型的影响。

 图 3.13 损失权重  $\lambda_{Ra}$  对 ViFR 模型的影响。

和  $\lambda_{Ra}$  不敏感，表明 ViFR 对损失权重的设置具有较好的鲁棒性。根据实验结果，ViFR 的损失权重  $\lambda_{SAMC}$  和  $\lambda_{Ra}$  在细粒度数据集（例如，CUB 和 SUN 数据集）/粗粒度数据集（例如，AWA2 数据集）分别设置为 0.1/0.05 和 0.1/1e<sup>-5</sup>。

未知类伪视觉特征样本的数量  $N_{syn}$ : 本章节分析了未知类伪视觉特征样本的数量  $N_{syn}$  对 ViFR 的影响。如图3.14所示，ViFR 的评估指标  $acc$  和  $H$  面对  $N_{syn}$  的变化相对稳定。当  $N_{syn}$  增大时，已知类的识别精度略有下降，而未知类的识别精度有所提高，使得调和平均值保持稳定。这是因为 ViFR 可以通过增加数据来缓解已知类偏差问题<sup>[94,106]</sup>。由于生成样本的多样性存在上界，所以  $N_{syn}$  的值不宜设置过大。根据实验结果，ViFR 将  $N_{syn}$  在 SUN 数据集、CUB 数据集和 AWA2 数据集分别设置为



(a) CUB 数据集

(b) AWA2 数据集

 图 3.14 未知类伪视觉特征样本的数量  $N_{syn}$  对 ViFR 模型的影响。

400、900 和 5000。

**消融实验分析：**为了进一步验证 ViFR 的有效性，本文使用基于不同模型成分、特征成分的 ViFR 进行了相关的消融实验。由于 ViFR 把基于 f-VAEGAN<sup>[61]</sup> 作为特征生成模型，本章节将此方法重新复现的实验结果作为基准（baseline）。

**基于不同模型成分的 ViFR：**如表3.5所示，当 ViFR 仅使用 Pre-FR 进行视觉特征的前置细化时，相较于基准方法 f-VAEGAN<sup>[61]</sup>，其在 CUB 数据集和 AWA2 数据集上的性能指标  $acc/H$  分别提升 8.9%/11.3% 和 4.0%/4.2%；当 ViFR 仅使用 Post-FR 进行视觉特征的后置精细化时，相较于基准方法 f-VAEGAN<sup>[61]</sup>，其在 CUB 数据集和 AWA2 数据集上的性能指标  $acc/H$  分别提升了 2.4%/4.0% 和 1.8%/5.3%。这表明本文提出的 Pre-FR 和 Post-FR 均可以有效地通过特征精细化实现视觉特征增强，从而缓解跨数据集偏差问题。当 ViFR 同时使用 Pre-FR 和 Post-FR 进行特征精细化时取得更好的性能。如章节3.1分析，细粒度数据集和 ImageNet 之间的偏差大于粗粒度数据集和 ImageNet 之间的偏差。为此，相较于粗粒度数据集（例如，AWA2），ViFR 可以在细粒度数据集（例如，CUB）上取得更明显的性能提升。

此外，ViFR 中不同的损失函数也起到了积极作用（例如， $\mathcal{L}_{Ra}$  和  $\mathcal{L}_{SAMC}$ ）。当 ViFR 在 Post-FR 中不使用  $\mathcal{L}_{SAMC}$  时，其在 CUB 数据集和 AWA2 数据集上的性能指标  $acc/H$  分别比其完整模型降低了 0.9%/2.0% 和 2.2%/2.7%；当 ViFR 在 Post-FR 中不使用  $\mathcal{L}_{Ra}$  损失进行模型优化时，其性能指标  $acc/H$  分别下降了 0.8%/1.0% 和 1.3%/2.2%。这表明了 ViFR 中不同函数均产生了一定的作用。完整版本的 ViFR 在所

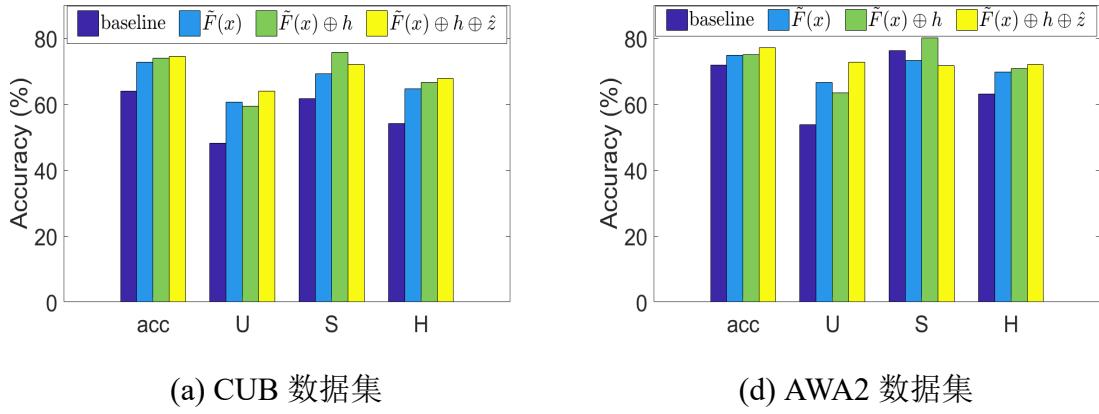


图 3.15 不同特征成分对 ViFR 的影响。

 表 3.5 基于不同模型成分的 ViFR 在 CUB 数据集<sup>[4]</sup> 和 AWA2 数据集<sup>[6]</sup> 上的实验结果。

基于不同模型成分的 ViFR	CUB 数据集				AWA2 数据集			
	CZSL	GZSL			CZSL	GZSL		
		acc	U	S		acc	U	S
基准方法 f-VAEGAN <sup>[61]</sup>	64.0	48.1	61.7	54.1	71.9	53.7	76.2	63.0
ViFR (无 Pre-FR)	66.4	52.6	64.9	58.1	73.7	61.3	77.2	68.3
ViFR (无 Post-FR)	72.9	61.6	69.6	65.4	75.9	61.0	74.7	67.2
ViFR (无 $\mathcal{L}_{SAMC}$ )	73.6	62.5	69.2	65.7	75.6	66.4	75.1	70.5
ViFR (无 $\mathcal{L}_{Ra}$ )	73.7	63.4	70.3	66.7	76.5	62.0	83.4	71.0
ViFR (full)	74.5	63.9	72.0	67.7	77.8	68.2	78.9	73.2

有数据集上都取得了最好的性能。相较于基准方法 f-VAEGAN<sup>[61]</sup>, ViFR 在 CUB 数据集和 AWA2 数据集上取得了非常高的识别性能提升，在性能指标  $acc/H$  分别提高；10.5%/13.6% 和 5.9%/10.2%。这些结果表明，ViFR 使用特征精细化实现视觉特征增强，有效地缓解了跨数据集偏差问题，从而提高零样本图像分类效果。

**基于不同特征成分的 ViFR:** 在进行后置特征精细化时，ViFR 融合了 Post-FR 中不同层的特征（例如，输入层的前置精细化视觉特征  $\tilde{F}(x)$ 、中间层特征 ( $h$ ) 和重构的类语义特征 ( $\hat{z}$ )）对视觉特征进行表示，从而实现视觉特征的增强。本文对不同的特征成分进行实验验证。如图3.15所示，Post-FR 后的所有特征成分均可显著提高 ViFR 在 CZSL 和 GZSL 设置情况下的识别性能。由于 Post-FR 和生成模型进行联合优化，促使语义 → 视觉映射函数（即生成器  $G$ ）学习语义相关的特征表示，从而能提高生成视觉特征的判别性。为此，当 ViFR 仅使用输入层的前置视觉

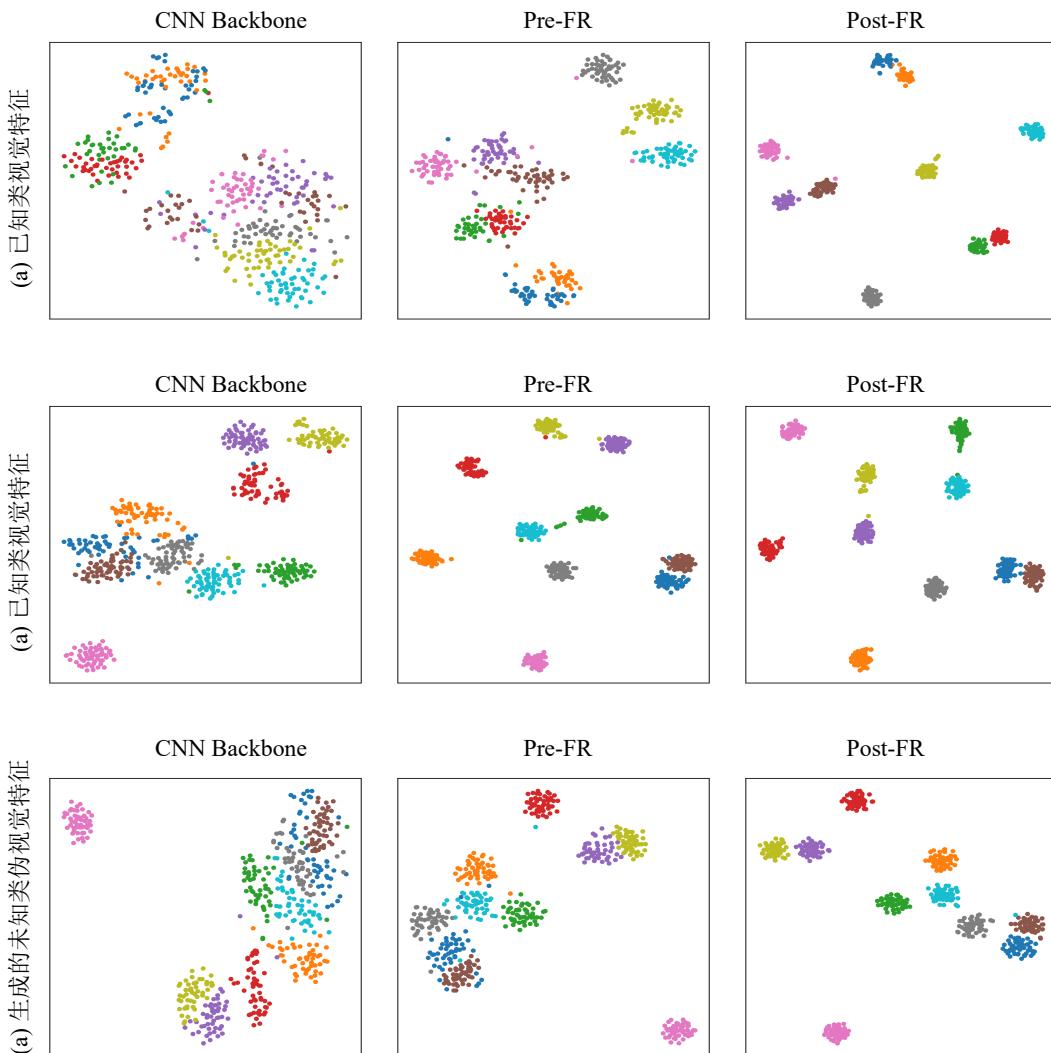


图 3.16 ViFR 和 CNN Backbone 在 CUB 数据集上学习的视觉特征 t-SNE<sup>[1]</sup> 可视化对比。

特征  $\bar{F}(x)$  进行后置特征精细化时，ViFR 一致提高了基准方法 f-VAEGAN<sup>[61]</sup> 性能指标  $acc/\mathbf{H}$ 。如果 ViFR 同时将 Post-FR 输入层的前置精细化视觉特征与隐藏特征融合（定义为  $\tilde{F}(x) \oplus h$ ），或者进一步融合 Post-FR 输出层的重构类语义特征（定义为  $\tilde{F}(x) \oplus h \oplus \hat{z}$ ），ViFR 的识别性能得到进一步提升。这些实验结果表明了 ViFR 中不同特征成为对视觉特征增强发挥了重要作用。

**定性实验分析：**如图3.1所示，跨数据集偏差问题使得零样本数据集的视觉特征判别性差，根本上限制了生成式零样本图像分类方法学习有效的语义 → 视觉映射，从而影响零样本图像分类效果。本文提出 ViFR 进行视觉特征精细化增强视觉特征表示，以缓解这种偏差问题。如图3.16所示，本章节展示了 CUB 数据集数据集上真实已知

# 华中科技大学博士学位论文

---

类和真实/生成未知类的 CNN Backbone 学习的视觉特征、Pre-FR 学习的前置特征精细化视觉特征、Post-FR 学习的完全特征精细化视觉的 t-SNE 可视化<sup>[1]</sup>。结果表明，ViFR 中的两种特征细化方法（Pre-FR 和 PostFR）一致地增强了已知类和未知类的视觉特征表示，避免不同类别之间的混淆。此外，图3.16显示了生成的未知类视觉特征与真实的未知类视觉特征具有一致的类关系，表明 ViFR 生成了更准确的未知类伪样本，可以学习一个更准确的有监督分类器实现正确的零样本图像分类（如图3.9）。因此，与当前最先进的方法及其基准方法相比，ViFR 实现了令人印象深刻的性能增益。

**通用性实验分析：**本章节提出的 ViFR 具有很强的通用性，其可以基于 f-VAEGAN<sup>[6]</sup> 生成模型，也可以基于其他先进的生成模型进行特征精细化，从而实现显著的性能提升。本文将 ViFR 中的 Pre-FR 模块集成到 TF-VAEGAN<sup>[7]</sup>、HSVA<sup>[8]</sup> 等两个生成式零样本图像分类模型中，实验结果如表所示。实验结果显示：与 TF-VAEGAN<sup>[7]</sup> 相比，当使用分辨率为  $224 \times 224$  的图像时， $\text{Pre-FR} \oplus \text{TF-VAEGAN}^{[7]}$  相较于  $\text{TF-VAEGAN}^{[7]}$  在 CUB 和 AWA2 数据集上分别取得的性能提升为  $\text{acc}/\mathbf{H} = 2.4\%/2.3\%$  和  $\text{acc}/\mathbf{H} = 1.9\%/1.9\%$ ， $\text{Pre-FR} \oplus \text{HSVA}^{[8]}$  相较于  $\text{HSVA}^{[8]}$  在 CUB 和 AWA2 数据集上分别取得的性能提升为  $\text{acc}/\mathbf{H} = 0.2\%/1.1\%$  和  $\text{acc}/\mathbf{H} = 1.3\%/2.9\%$ ；当使用分辨率为  $448 \times 448$  的图像时， $\text{Pre-FR} \oplus \text{TF-VAEGAN}^{[7]}$  相较于  $\text{TF-VAEGAN}^{[7]}$  在 CUB 和 AWA2 数据集上分别取得的性能提升为  $\text{acc}/\mathbf{H} = 7.4\%/8.3\%$  和  $\text{acc}/\mathbf{H} = 2.6\%/3.1\%$ ， $\text{Pre-FR} \oplus \text{HSVA}^{[8]}$  相较于  $\text{HSVA}^{[8]}$  在 CUB 和 AWA2 数据集上分别取得的性能提升为  $\text{acc}/\mathbf{H} = 8.3\%/10.2\%$  和  $\text{acc}/\mathbf{H} = 2.1\%/3.0\%$ 。这些结果表明本文提出的特征精细化方法能有效的实现视觉特征增强，从而解决生成式零样本图像分类方法面临的跨数据集偏差问题。本文特别指出，当使用较大分辨率图像时，特征精细化方法对视觉特征增强更为有效，从而相较于基准模型取得更大的性能提升。此外，本文提出的特征精细化方法在细粒度数据集比粗粒度数据集能取得更显著的性能提升，这是因为细粒度数据集与 ImageNet 数据集的偏差（例如， $MMD_{\text{ImageNet}-\text{SUN}} = 0.527$ ）比粗粒度数据集与 ImageNet 数据集的偏差（ $MMD_{\text{ImageNet}-\text{AWA2}} = 0.417$ ）更大，更多分析讨论详见章节3.1。

**ViFR 和其他先进的零样本图像分类方法的实验结果对比：**由于 ViFR 是一种归纳式方法，本文将 ViFR 与和其他经典且先进的零样本图像分类方法在 CUB<sup>[4]</sup>、SUN<sup>[5]</sup>、AWA2<sup>[6]</sup> 数据集上的 CZSL/GZSL 实验结果对比，包括嵌入式方法（例如，

表 3.6 Pre-FR 模块集成到 TF-VAEGAN<sup>[7]</sup>、HSVA<sup>[8]</sup> 等两个生成式零样本图像分类模型在 CUB<sup>[4]</sup> 数据集和 AWA2<sup>[6]</sup> 数据集上的实验结果。

不同的零样本图像分类方法	CUB 数据集			AWA2 数据集				
	CZSL	GZSL		CZSL	GZSL			
		acc	U		S	H		
TF-VAEGAN <sup>224[7]</sup>	64.9	52.8	64.7	58.1	72.2	59.8	75.1	66.6
Pre-FR⊕TF-VAEGAN <sup>224</sup>	67.3 <sup>↑2.4</sup>	56.4	65.1	60.4 <sup>↑2.3</sup>	74.1 <sup>↑1.9</sup>	56.9	85.9	68.5 <sup>↑1.9</sup>
TF-VAEGAN <sup>448[7]</sup>	67.7	53.6	67.2	59.6	75.0	56.9	84.2	67.9
Pre-FR⊕TF-VAEGAN <sup>448</sup>	75.1 <sup>↑7.4</sup>	63.7	72.5	67.9 <sup>↑8.3</sup>	77.6 <sup>↑2.6</sup>	60.4	86.2	71.0 <sup>↑3.1</sup>
HSVA <sup>224[8]</sup>	62.8	52.7	58.3	55.3	70.1	56.7	79.8	66.3
Pre-FR⊕HSVA <sup>224</sup>	63.0 <sup>↑0.2</sup>	53.4	59.9	56.4 <sup>↑1.1</sup>	71.4 <sup>↑1.3</sup>	61.2	79.6	69.2 <sup>↑2.9</sup>
HSVA <sup>448[8]</sup>	59.3	47.5	59.1	52.7	72.1	59.4	86.2	70.3
Pre-FR⊕HSVA <sup>448</sup>	67.6 <sup>↑8.3</sup>	57.3	69.6	62.9 <sup>↑10.2</sup>	74.2 <sup>↑2.1</sup>	66.7	81.3	73.3 <sup>↑3.0</sup>

注：“<sup>224</sup>”和“<sup>448</sup>”分别表示模型的输入图像分辨率分别为  $224 \times 224$  和  $448 \times 448$ 。

SP-AEN<sup>[108]</sup>, SGMA<sup>[40]</sup>, AREN<sup>[2]</sup>, LFGAA<sup>[74]</sup>, DAZLE<sup>[81]</sup>, APN<sup>[41]</sup> 等)、生成式方法(例如, f-CLSWG<sup>[89]</sup>, cycle-CLSWGAN<sup>[93]</sup>, f-VAEGAN<sup>[61]</sup>, LsrGAN<sup>[102]</sup>, E-PGN<sup>[111]</sup>, RFF-GZSL<sup>[94]</sup>, GCM-CF<sup>[123]</sup>, ZeroNAS<sup>[131]</sup> 等)和公共子空间式方法(例如, DeVISE<sup>[68]</sup>, DCN<sup>[46]</sup>, CADA-VAE<sup>[3]</sup>, SGAL<sup>[105]</sup> 等)。

表3.7展示了 ViFR 和其他先进的零样本图像分类方法在 CZSL 设置下的结果对比。结果表明, 本文提出的 ViFR 在 CUB、SUN、AWA2 数据集上均达到了最先进的性能, 分别为 74.5%、69.2% 和 77.8% 的分类精度。相较于现有先进的生成式零样本图像分类方法<sup>[61,111,133]</sup>, ViFR 在 CUB、SUN、AWA2 数据集上至少取得了 2.1%、4.5%、5.0% 的性能提升。这些结果表明 ViFR 通过特征精细化学习实现视觉特征的增强, 有效缓解跨数据集偏差问题, 从而提高从已知类到未知类的知识迁移并在未知类取得较好的分类表现。

表3.7展示了 ViFR 和其他先进的零样本图像分类方法在 GZSL 设置下的结果对比, 包括未知类识别精度 (**U**)、已知类识别精度 (**S**) 及它们的调和均值 (**H**)。结果显示, ViFR 在 CUB、SUN、AWA2 数据集上均取得最佳性能, 性能指标 **H** 分别为 67.7%、44.7%、72.1%。这些结果表明, ViFR 在没有自校准机制的情况下同时在已知类和未知类均取得较好的零样本分类性能。这主要得益于: (1) 使用 Pre-FR 进行前置特征精细化的视觉特征具有更强的判别性, 从而增强视觉特征并促进生成模型学习更准确的语义 → 视觉映射; (2) 生成模型和 Post-FR 在统一模型里面进行联合训练, 使得生成器学习具有更强语义的特征表示, 从而增强生成的未知类伪特征; (3)

表 3.7 ViFR 和其他先进的零样本图像分类方法在 CUB<sup>[4]</sup>、SUN<sup>[5]</sup>、AWA2<sup>[6]</sup> 数据集上的 CZSL/GZSL 实验结果对比。

当前先进的零样本图像分类方法	CUB 数据集				SUN 数据集				AWA2 数据集			
	CZSL		GZSL		CZSL		GZSL		CZSL		GZSL	
	acc	U	S	H	acc	U	S	H	acc	U	S	H
<b>嵌入式方法</b>												
SP-AEN <sup>[108]</sup>	55.4	34.7	70.6	46.6	59.2	24.9	38.6	30.3	58.5	23.3	90.9	37.1
SGMA <sup>[40]</sup>	71.0	36.7	71.3	48.5	—	—	—	—	68.8	37.6	87.1	52.5
AREN <sup>[2]</sup>	71.8	38.9	78.7	52.1	60.6	19.0	38.8	25.5	67.9	15.6	92.9	26.7
LFGAA <sup>[74]</sup>	67.6	36.2	<b>80.9</b>	50.0	61.5	18.5	<b>40.0</b>	25.3	68.1	27.0	<b>93.4</b>	41.9
TCN <sup>[76]</sup>	59.5	52.6	52.0	52.3	61.5	31.2	37.3	34.0	71.2	61.2	65.8	63.4
DAZLE <sup>[81]</sup>	66.0	56.7	59.6	58.1	59.4	52.3	24.3	33.2	67.9	60.3	75.7	67.1
APN <sup>[41]</sup>	72.0	<b>65.3</b>	69.3	67.2	61.6	41.9	34.0	37.6	68.4	57.1	72.4	63.9
<b>公共子空间式方法</b>												
DeViSE <sup>[68]</sup>	52.0	23.8	53.0	32.8	56.5	16.9	27.4	20.9	54.2	17.1	74.7	27.8
DCN <sup>[46]</sup>	56.2	28.4	60.7	38.7	61.8	25.5	37.0	30.2	65.2	25.5	84.2	39.1
CADA-VAE <sup>[3]</sup>	59.8	51.6	53.5	52.4	61.7	47.2	35.7	40.6	63.0	55.8	75.0	63.9
SGAL <sup>[105]</sup>	—	40.9	55.3	47.0	—	35.5	34.4	34.9	—	52.5	86.3	65.3
<b>生成式方法</b>												
f-CLSWGAN <sup>[89]</sup>	57.3	43.7	57.7	49.7	60.8	42.6	36.6	39.4	68.2	57.9	61.4	59.6
cycle-CLSWGAN <sup>[93]</sup>	58.4	45.7	61.0	52.3	60.0	49.4	33.6	40.0	66.3	56.9	64.0	60.2
f-VAEGAN <sup>[61]</sup>	61.0	48.4	60.1	53.6	64.7	45.1	38.0	41.3	71.1	57.6	70.6	63.5
E-PGN <sup>[111]</sup>	72.4	52.0	61.1	56.2	—	—	—	—	73.4	52.6	83.5	64.6
TF-VAEGAN <sup>[7]</sup>	64.9	52.8	64.7	58.1	66.0	45.6	40.7	43.0	72.2	59.8	75.1	66.6
LsrGAN <sup>[102]</sup>	60.3	48.1	59.1	53.0	62.5	44.8	37.7	40.9	—	53.1	68.8	60.0
Composer <sup>[132]</sup>	69.4	56.4	63.8	59.9	62.6	<b>55.1</b>	22.0	31.4	71.5	62.1	77.3	68.8
RFF-GZSL <sup>[94]</sup>	—	52.6	56.6	54.6	—	45.7	38.6	41.9	—	—	—	—
GCM-CF <sup>[123]</sup>	—	61.0	59.7	60.3	—	47.9	37.8	42.2	—	60.4	75.1	67.0
Chou et al. <sup>[133]</sup>	57.2	41.4	49.7	45.2	63.3	29.9	40.2	34.3	73.8	65.1	78.9	71.3
ZeroNAS <sup>[131]</sup>	66.4	56.0	63.8	59.6	68.3	47.1	<b>41.8</b>	44.3	73.2	75.3	61.4	67.6
<b>ViFR</b>	<b>74.5</b>	63.9	72.0	<b>67.7</b>	<b>69.2</b>	51.3	40.0	<b>44.7</b>	<b>77.8</b>	<b>68.2</b>	78.9	<b>73.2</b>

注：符号“—”表示相应结果缺失。

Post-FR 提取具有较强的视觉相关和语义相关的特征表示，使得 ViFR 在所有数据集上学习的完全精细化的视觉特征均具有更强的判别性和迁移性。为此，ViFR 是生成式零样本图像分类方法解决跨数据集偏差问题的有效方案。

### 3.4 本章小结

本章研究了基于深度表征的零样本图像分类中的跨数据集偏差问题，并面向嵌入式零样本图像分类和生成式零样本图像分类提出了相应的特征增强方法对该问题进行解决，分别为 GNDAN 和 ViFR。考虑到现有先进的嵌入式零样本图像分类方法受限于跨数据集偏差，它们学习的隐式全局视觉特征或者局部视觉特征表示能力不

# 华 中 科 技 大 学 博 士 学 位 论 文

---

足，限制了有效的视觉-语义交互。GNDAN 利用区域指导的注意力子网络和区域指导的图注意力网络分别学习局部和显式全局视觉特征，并将这两种特征融合为具有强判别性和迁移性的增强视觉特征，有效缓解跨数据集偏差问题，更准确地将视觉特征映射到语义空间实现零样本图像分类。面向生成式零样本图像分类方法，ViFR 使用前置特征精细化和后置特征精细化学习完全精细化特征实现视觉特征增强，并促使生成模型生成更真实的未知类视觉特征，从而训练更准确的有监督分类器实现零样本分类。大量的实验表明，本章提出的基于特征增强的方法有效解决面向嵌入式和生成式方法的跨数据集偏差问题。

## 4 基于属性-视觉关键公共语义知识的零样本图像分类

### 4.1 引言

如图4.1所示，一个未知类样本与一组已知类样本共享不同的关键局部信息，且这些关键局部信息由语义属性进行表示。未知类样本“Red legged Kittiwake”和一组已知类样本（例如，“Herring Gull”，“Parakeet Auklet”，“California Gull”，“Pigeon Guillemot”等）共享关键局部信息（例如，“Wing Color Gray”，“Breast Color White”，“Bill Color Yellow”，“Leg Color Red”等）。然而，深度表征是全局视觉特征表示，其不能有效地对这些语义属性进行刻画，造成视觉-语义的表示差异性问题。该问题直接限制了零样本图像分类从已知类到未知类的语义知识迁移。虽然最近一些基于注意力的零样本图像分类的方法<sup>[2,39,40]</sup>利用语义信息作为指导以挖掘具有判别性的局部/细粒度视觉特征，使得视觉特征更准确地映射到语义特征空间实现零样本图像分类。然而，这些方法局限于：(1) 它们只是学习图像中整体目标特征表示（例如，鸟的整个目标），而忽视了具有更强可区分性的属性定位知识；(2) 它们利用单向注意力机制，只能挖掘视觉特征和属性特征之间有限的潜在公共语义知识。这两个因素造成当前方法不能准确地、充分地挖掘视觉和属性特征之间的关键公共语义知识，从而不能实现有效的从已知类到未知类的语义知识迁移。因此，如何解决视觉-语义的表示差异性问题是实现准确的零样本图像分类的一个重要研究工作。

本文提出了基于属性-视觉关键公共语义知识的零样本图像分类方法。该方法使用基于属性指导的 Transformer 网络（章节4.2），利用单向跨注意力机制学习具有属性定位的视觉特征，对属性-视觉特征之间的关键公共语义知识进行准确表示。该方法利用互语义蒸馏网络（章节4.3），通过属性 → 视觉和视觉 → 属性两个双向注意力子网络分别学习基于属性的视觉特征和基于视觉的属性特征，并在互语义蒸馏学习机制的指导下，两个子网络更充分地挖掘关键语义特征。该方法将这两个网络集成成为一个统一的模型（章节4.4），准确且充分地挖掘关键公共语义知识，提高视觉-语义特征的语义一致性，实现零样本图像分类从已知类到未知类有效的语义知识迁移。

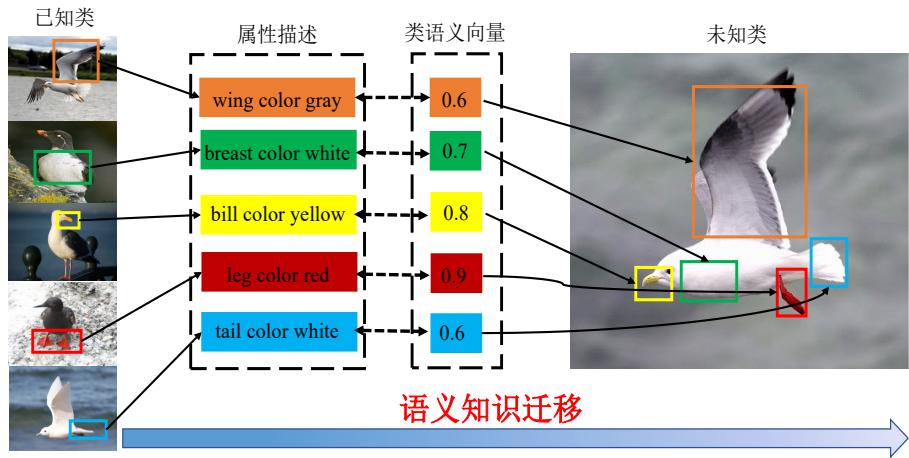


图 4.1 零样本图像分类的关键任务阐述：通过挖掘视觉特征和语义特征关键的公共语义知识实现有效地已知类到未知类的语义知识迁移。

## 4.2 基于属性指导 Transformer 的零样本图像分类

### 4.2.1 研究动机

大部分嵌入式零样本图像分类方法直接利用 ImageNet 数据集上预训练的 CNN Backbone（例如，ResNet-101<sup>[31]</sup>）对零样本数据集的图像进行全局视觉特征提取，使得视觉特征不能有效地对局部细粒度信息（例如，图像属性）进行刻画。近期有一些研究工作尝试利用注意力机制学习具有局部表示能力的视觉特征<sup>[2,39–42,81]</sup>，以增强视觉特征的属性语义表示，如图4.2(a)所示。然而，这些方法的局限性在于：i) 它们直接将局部特征和相应的先验几何关系用于零样本图像分类，这降低了视觉特征的迁移性；ii) 它们只是学习图像中整体目标特征表示（例如，鸟的整体目标），而忽视了具有更强可区分性的属性定位（例如，鸟的局部外观）的重要性。为此，现有方法不能准确地挖掘属性-视觉的关键公共语义知识以解决视觉-语义特征的表示差异性问题（章节4.1中指出），从而限制了零样本图像分类从已知类到未知类的语义知识迁移。

为了应对上述挑战，本章节提出了一种基于属性指导 Transformer 的零样本图像分类方法（Attribute-Guided Transformer for Zero-Shot Image Classification，简称为 TransZero），通过减少局部特征之间的先验几何关系以提高视觉特征的迁移性，并将图像属性进行视觉定位以表示具有可区分性的细粒度特征（即属性-视觉的关键公共语义知识），有效促进从已知类到未知类的知识迁移，提高零样本分类效果，如图所

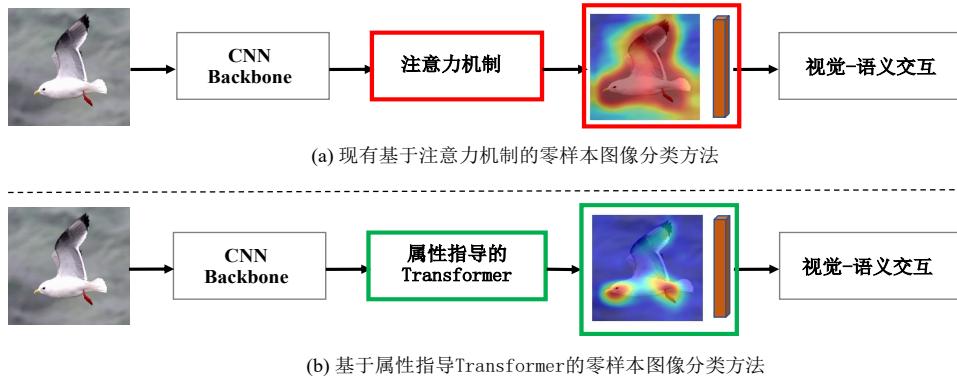


图 4.2 现有基于注意力机制的零样本图像分类方法和本章节提出的基于属性指导 Transformer 的零样本图像分类方法的模型对比。

示4.2(b)。具体而言，TransZero 由一个属性指导 Transformer 网络（Attribute-Guided Transformer, AGT）和一个视觉-语义嵌入网络（Visual-Semantic Embedding Network, VSEN）组成，前者学习具有属性-视觉关键公共语义知识的局部增强视觉特征，后者进行视觉语义-交互。在 AGT 中，TransZero 首先采用一个特征增强编码器，减少不同局部特征之间的相对几何关系，从而提高视觉特征的迁移性。在属性特征的指导下，TransZero 在 AGT 中使用属性 → 视觉解码器对图像中每个属性进行视觉定位，学习具有属性-视觉关键公共语义知识的局部增强视觉特征。在 VSEN 模块中，利用局部增强视觉特征和类语义向量实现有效的视觉-语义交互。充分的实验表明，TransZero 有效地挖掘属性-视觉关键公共语义知识，从而提高模型从已知类到未知类的知识迁移，并在三个主流的零样本标准数据集（例如 CUB<sup>[4]</sup>、SUN<sup>[5]</sup>、AWA2<sup>[6]</sup> 数据集）上取得了比其他基于注意力的零样本图像分类方法更好的分类效果。TransZero 是首个将 Transformer 引入到零样本图像分类领域的工作，充分展示了 Transformer<sup>[32]</sup> 挖掘不同模态特征的公共语义知识的潜力。

#### 4.2.2 属性指导的 Transformer

本章节提出的 TransZero 由一个属性指导的 Transformer 网络（AGT）和一个视觉-语义嵌入网络（VSEN），如图4.3所示。AGT 使用特征增强编码器消除局部特征间的几何先验关系以提高视觉特征迁移性，并使用属性 → 视觉解码器学挖掘属性-视觉特征的关键公共语义知识对局部增强视觉特征进行表示。VSEN 进行零样本图像分类的视觉-语义交互。TransZero 使用属性回归损失、基于属性的交叉熵损失和自校

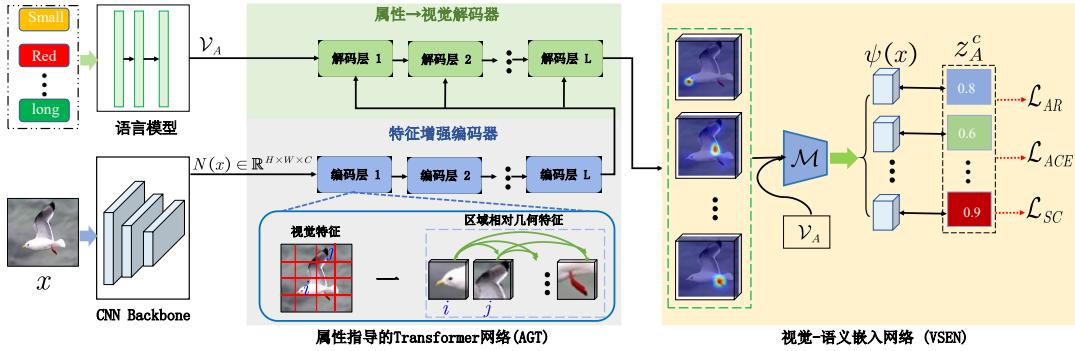


图 4.3 本章提出的 TransZero 模型结构示意图。

准损失进行模型优化。本章节使用的基本符号定义见章节3.2.2。

**属性指导 Transformer 的特征增强编码器：**现有的零样本图像分类方法通常将从 CNN Backbone 中提取的视觉特征  $N(x) \in \mathbb{R}^{H \times W \times C}$  通过池化层拉伸为特征向量，进一步用于生成模型或嵌入式学习。然而，这样的特征向量隐含地将图像中不同区域之间的视觉特征表示进行统一表示，增加了一些图像固有先验信息(例如，由于鸟的头部和翅膀比鸟的脚更靠近，使得鸟的头部和翅膀比头部和腿具有更强的视觉一致性)，这阻碍了它们从一个域到其他域(例如，从已知类到未知类)的迁移性<sup>[41,134,135]</sup>。因此，本文提出了一种特征增强的缩放点积注意机制(Feature Augmented Scale Dot-Product Attention)，通过减少不同区域网格特征之间的相对几何关系来进一步增强 AGT 的编码器。

为了学习不同区域网格特征相对几何表示<sup>[136,137]</sup>，TransZero 首先根据第  $i$  个网格的二维相对位置  $\{(v_i^{\min}, t_i^{\min}), (v_i^{\max}, t_i^{\max})\}$ ，计算其相对中心坐标  $(v_i^{\text{cen}}, t_i^{\text{cen}})$ :

$$(v_i^{\text{cen}}, t_i^{\text{cen}}) = \left( \frac{v_i^{\min} + v_i^{\max}}{2}, \frac{t_i^{\min} + t_i^{\max}}{2} \right), \quad (4.1)$$

$$w_i = (v_i^{\max} - v_i^{\min}) + 1, \quad (4.2)$$

$$h_i = (t_i^{\max} - t_i^{\min}) + 1, \quad (4.3)$$

其中  $(v_i^{\min}, t_i^{\min})$  和  $(v_i^{\max}, t_i^{\max})$  分别是视觉网格特征  $i$  左上角和右下角的相对位置坐标。随后，TransZero 构造第  $i$  个和  $j$  个视觉网格特征的区域相对集合表示:

$$G_{ij} = \text{ReLU}(w_g^T g_{ij}), \quad (4.4)$$

$$g_{ij} = FC(r_{ij}), \quad r_{ij} = \begin{pmatrix} \log\left(\frac{|v_i^{\text{cen}} - v_j^{\text{cen}}|}{w_i}\right) \\ \log\left(\frac{|t_i^{\text{cen}} - t_j^{\text{cen}}|}{h_i}\right) \end{pmatrix}, \quad (4.5)$$

其中,  $r_{ij}$  是第  $i$  个和  $j$  个视觉网格特征的相对集合关系,  $FC$  是一个全连接层并跟随一个  $RELU$  激活层,  $w_g^T$  是一组可学习的权重参数。

最后, TransZero 从编码器的缩放点积注意力中去除相应的区域几何关系表示, 以提供更准确的注意图, 定义为:

$$Q^e = U(x)W_q^e, K^e = U(x)W_k^e, V^e = U(x)W_v^e, \quad (4.6)$$

$$Z_{aug} = \text{softmax}\left(\frac{Q^e K^{e^\top}}{\sqrt{d^e}} - G\right) V^e, \quad (4.7)$$

$$U_{aug}(x) \leftarrow U + Z_{aug}, \quad (4.8)$$

其中,  $Q^e$ 、 $K^e$ 、 $V^e$  分别使查询矩阵、键矩阵和值矩阵,  $W_Q^e$ 、 $W_K^e$ 、 $W_V^e$  是可学习的权重矩阵,  $d^e$  是缩放因子,  $Z_{aug}$  是增强特征。另外,  $U(x) \in \mathbb{R}^{HW \times C}$  是由视觉特征  $N(x)$  通过全连接层、ReLU 层和 Dropout 层压缩的网格视觉特征 (Grid Feature)。

**属性指导 Transformer 的属性 → 视觉解码器:** 不同于标准的 Transformer<sup>[32]</sup>, TransZero 采用跨注意力机制 (Cross Attention) 和前馈网络 (Feed-Farward Network, FFN) 来构建解视觉-语义码器。属性指导 Transformer 的跨注意力机制和标准 Transformer 的自注意力机制如图4.4所示。在语义属性特征  $\mathcal{V}_A = \{v_a\}_{a=1}^A$  的指导下, 属性 → 视觉解码器可以有效地定位出图像中每个属性最相关的图像区域, 从而挖掘属性-视觉的关键公共语义知识。跨注意力层使用编码器中增强后的视觉特征  $U$  作为键矩阵 ( $K_t^d$ ) 和值矩阵 ( $V_t^d$ ) 以及语义属性特征  $\mathcal{V}_A$  作为查询矩阵 ( $Q_t^d$ )。形式化表示如下:

$$Q_t^d = \mathcal{V}_A W_{qt}^d, K_t^d = U_{aug}(x) W_{kt}^d, V_t^d = U_{aug}(x) W_{vt}^d, \quad (4.9)$$

$$\text{head}_t = \text{softmax}\left(\frac{Q_t^d K_t^{d^\top}}{\sqrt{d^d}}\right) V_t^d, \quad (4.10)$$

$$\hat{F} = \|_{t=1}^T (\text{head}_t) W_o, \quad (4.11)$$

其中,  $W_{qt}^d$ ,  $W_{kt}^d$ ,  $W_{vt}^d$  是可学习的权重,  $d^d$  是缩放因子,  $\parallel$  是串联操作。然后, 将特征  $\hat{F}$  输入到一个 FFN 进一步学习:

$$F = \text{ReLU}(\hat{F} W_1 + b_1) W_2 + b_2, \quad (4.12)$$

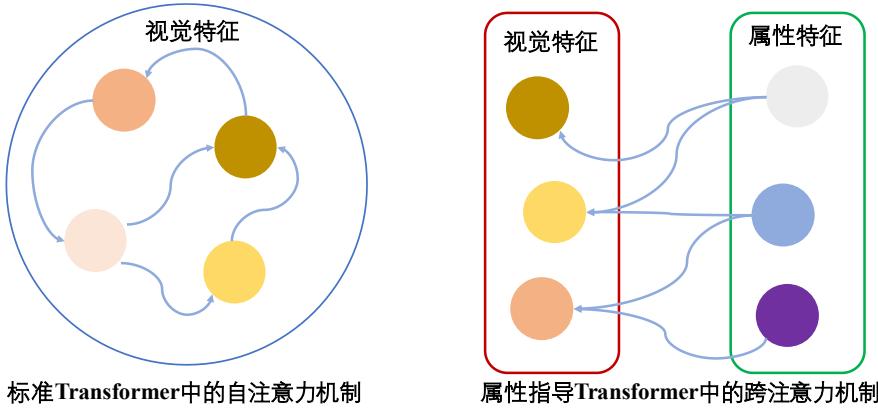


图 4.4 标准 Transformer 的自注意力机制和属性指导 Transformer 的跨注意力机制对比。

其中,  $W_1, W_2, b_1$  和  $b_2$  分别是线性层的权重和偏差,  $F$  是具有属性-视觉关键公共语义知识的局部增强视觉特征。

**视觉-语义嵌入网络:** 在学习具有属性-视觉关键公共语义知识的局部增强视觉特征后, 视觉-语义嵌入网络 (VSEN) 进一步将其映射到语义嵌入空间进行视觉-语义交互。为使映射更加准确, VSEN 以属性特征作为辅助, 使用视觉-语义映射器 ( $\mathcal{M}$ ) 将局部增强的视觉特征  $F$  映射到类语义空间:

$$\psi(x_i) = \mathcal{M}(F) = \mathcal{V}_A^\top W F, \quad (4.13)$$

其中,  $W$  是将  $F$  嵌入到语义属性空间的嵌入矩阵。本质上,  $\psi(x_i)[a]$  是一个属性得分, 表示在图像  $x_i$  中具有第  $a$ -个属性的置信度。给定一组语义属性特征  $\mathcal{V}_A = \{v_a\}_{a=1}^A$ , TransZero 即可获得相应视觉特征的语义嵌入  $\psi(x_i)$ 。

**模型优化:** 为了实现 TransZero 的有效优化, 本章节使用属性回归损失 (Attribute Regression Loss,  $\mathcal{L}_{AR}$ )、基于属性的交叉熵损失 (Attribute-Based Cross-Entropy Loss,  $\mathcal{L}_{ACE}$ ) 和自校准损失 (Self-Calibration Loss,  $\mathcal{L}_{SC}$ ) 进行模型优化。

**属性回归损失:** 为了鼓励 VSEN 将视觉特征准确地映射到相应的语义空间中, 本章节提出了属性回归损失对 TransZero 进行优化。本章节将视觉语-义映射视为一个回归问题, 并最小化批量样本  $\{x_i^s\}_{i=1}^{n_b}$  的类语义向量  $z^c$  和相应视觉特征的语义嵌入  $\psi(x_i)$  之间的均方误差:

$$\mathcal{L}_{AR} = -\frac{1}{n_b} \sum_{i=1}^{n_b} \|\psi(x_i^s) - z^c\|_2^2. \quad (4.14)$$

**基于属性的交叉熵损失:** 由于当属性在图像中出现时, 关联图像视觉特征的语

义嵌入在其相应类语义向量  $z^c$  附近，因此本章节采用基于属性的交叉熵损失  $\mathcal{L}_{\text{ACE}}$  优化 TransZero 模型的参数，即计算视觉嵌入和每个类语义向量之间的点积。这促使图像特征与其对应的类语义向量具有最高的相似性得分。给定一批  $n_b$  训练图像  $\{x_i\}_{i=1}^{n_b}$  及其相应的类语义向量  $z^c$ ,  $\mathcal{L}_{\text{ACE}}$  定义为：

$$\mathcal{L}_{\text{ACE}} = -\frac{1}{n_b} \sum_{i=1}^{n_b} \log \frac{\exp(\psi(x_i) \times z^c)}{\sum_{\hat{c} \in \mathcal{C}} \exp(\psi(x_i) \times z^{\hat{c}})}. \quad (4.15)$$

**自校准损失：**由于  $\mathcal{L}_{\text{AR}}$  和  $\mathcal{L}_{\text{ACE}}$  只在已知类上进行模型优化，使得 TransZero 不可避免地过拟合于这些已知类，如 Zhu 等人<sup>[40]</sup>、Huynh 等人<sup>[81]</sup>、Xu 等人<sup>[41]</sup> 所观察到结论的一样。为解决这一问题，本文进一步引入了自校准损失  $\mathcal{L}_{\text{SC}}$ ，显式地将未知类的预测概率增加置信度。 $\mathcal{L}_{\text{SC}}$  的形式化表示如下：

$$\mathcal{L}_{\text{SC}} = -\frac{1}{n_b} \sum_{i=1}^{n_b} \sum_{c'=1}^{\mathcal{C}^u} \log \frac{\exp(\psi(x_i) \times z^{c'} + \mathbb{I}_{[c' \in \mathcal{C}^u]})}{\sum_{\hat{c} \in \mathcal{C}} \exp(\psi(x_i) \times z^{\hat{c}} + \mathbb{I}_{[\hat{c} \in \mathcal{C}^u]})}, \quad (4.16)$$

其中  $\mathbb{I}_{[c \in \mathcal{C}^u]}$  是一个指示函数（即，当  $c \in \mathcal{C}^u$  时为 1，否则为-1）。直观上， $\mathcal{L}_{\text{SC}}$  鼓励在训练期间将非零概率分配给未知类，这允许 TransZero 对未知类样本进行测试时，为真正未知类提供更大的非零概率值。

最后，TransZero 的完整优化目标函数为：

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ACE}} + \lambda_{\text{AR}} \mathcal{L}_{\text{AR}} + \lambda_{\text{SC}} \mathcal{L}_{\text{SC}}, \quad (4.17)$$

其中  $\lambda_{\text{AR}}$  和  $\lambda_{\text{SC}}$  相应的损失权重。

**零样本图像分类：**训练好 TransZero 之后，先获得测试实例  $x_i$  在语义空间中语义嵌入  $\psi(x_i)$ 。然后，利用最近邻匹配与显式自校准预测  $x_i$  的标签：

$$c^* = \arg \max_{c \in \mathcal{C}^u / \mathcal{C}} \psi(x_i) \times z^c + \mathbb{I}_{[c \in \mathcal{C}^u]}, \quad (4.18)$$

$\mathcal{C}^u / \mathcal{C}$  分别对应于 CZSL/GZSL 设置下的零样本图像分类。

#### 4.2.3 实验结果与分析

为验证 TransZero 方法的有效性，本章节在三个主流的零样本图像分类标准数据集上进行实验，包括两个细粒度数据集（CUB<sup>[4]</sup>, SUN<sup>[5]</sup>）和一个粗粒度数据集（AWA2<sup>[6]</sup>）。本章节依据 Xian 等人<sup>[6]</sup> 最新的数据集划分方式进行模型训练和测试，数据集的具体介绍见章节2.3.1。本文利用在 ImageNet 上预训练的 ResNet-101 作为

# 华中科技大学博士学位论文

---

CNN Backbone 对图像样本（分辨率裁剪为  $448 \times 448$ ）提取相应的视觉特征。使用带有超参数（momentum = 0.9, weight decay = 0.0001）的 SGD 优化器来优化 TransZero 模型，学习率和批量大小分别设置为 0.0001 和 50。根据验证集上的实验结果，将 TransZero 的超参  $\lambda_{SC}$  在所有数据集上设置为 0.3，将  $\lambda_{AR}$  设置为 0.005，编码器和解码器层的层数均设置为 1，且注意力层只使用单头注意力。

本章节通过以下实验验证 TransZero 的有效性：

- 超参实验分析；
- 消融实验分析；
- 定性实验分析；
- TransZero 和其他经典且先进的零样本图像分类方法在 CZSL/GZSL 不同设置下的实验结果对比。

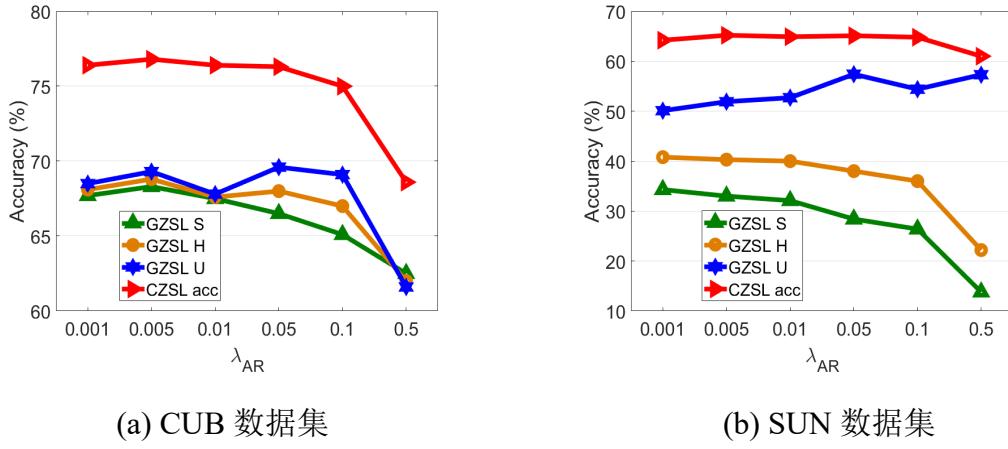
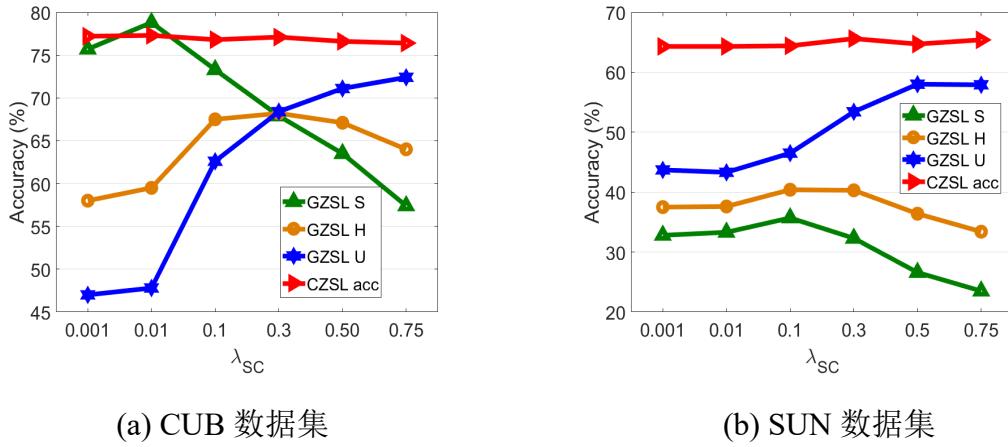
接下来将对不同的实验进行详细阐述和分析。本章节方法 TransZero 的源代码地址：  
<https://github.com/shiming-chen/TransZero>。

**超参实验分析：**TrasZero 的主要超参数是公式4.17的两个损失权重  $\lambda_{AR}$  和  $\lambda_{SC}$ 。本章节将对这两个超参数在 CUB 数据集和 SUN 数据集做相应的实验分析，并进行超参数设置。

**损失权重  $\lambda_{AR}$ ：** $\lambda_{AR}$  用于衡量属性回归损失对整体模型优化的影响，这将直接影响 VSEN 进行有效的视觉-语义交互。本文取用  $\lambda_{AR}$  一定范围的取值进行实验，例如  $\lambda_{AR} = \{0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$ 。结果如图4.5所示。当  $\lambda_{AR}$  设置为较大的值时，所有评估指标都会下降。这是因为属性回归损失的损失值较大，极易减轻了其他损失的对模型的贡献。当  $\lambda_{AR}$  设置为 0.005 时，TransZero 取到相对较好的性能。

**损失权重  $\lambda_{SC}$ ：** $\lambda_{SC}$  调整自校准损失的权重，这有效地缓解已知类偏差问题<sup>[41,81]</sup>。如图4.6所示，在 GZSL 设置情况下，当增加  $\lambda_{SC}$  时，已知类的识别率下降，而未知类的识别率上升。这表明白校准损失有效地缓解已知类偏差问题<sup>[89,106]</sup>，对已知类和未知类的识别精度做一定的平衡。而在 CZSL 设置情况下，TransZero 对自校准损失不敏感。根据实验结果，本章节将所有数据集的  $\lambda_{SC}$  设置为 0.3。

**消融实验分析：**为了进一步深入验证 TransZero 不同模型成分的有效性，本章节进行了充分的消融实验以评估特征增强编码器、不进行特征增强的编码器、属性 → 视觉解码器、自校准损失以及属性回归损失的影响。实验结果如表4.1所示。当未使用特征增强编码器时，TransZero 的性能明显低于其完整的模型，例如，CUB 数据集上


 图 4.5 损失权重  $\lambda_{AR}$  对 TransZero 模型的影响。

 图 4.6 损失权重  $\lambda_{SC}$  对 TransZero 模型的影响。

$acc/\mathbf{H}$  下降 9.5%/12.0%，SUN 数据集上  $acc/\mathbf{H}$  下降 4.4%/8.7%。如果编码器中不进行几何先验信息消除，TransZero 的结果也会有所下降，例如，在 CUB 数据集和 SUN 数据集上的  $acc/\mathbf{H}$  分别下降 2.8%/2.3% 和 1.8%/2.3%。这些结果表明编码器有效提高视觉特征的迁移性和判别性。当 TransZero 没有属性  $\rightarrow$  视觉解码器时，它在所有数据集上的性能都会急剧下降，表明学习具有属性-视觉关键公共语义知识的局部增强视觉特征对零样本图像分类从已知类到未知类的知识迁移尤为重要。此外，自校准机制可以有效缓解偏差问题，使 TransZero 在 CUB 数据集和 SUN 数据集上的调和均值分别提高 10.7% 和 3.4%。属性回归损失约束直接指导 VSEN 进行有效的视

觉-语义交互，进一步提高了 TransZero 的性能。

表 4.1 不同模型成分设置下，TransZero 在 CUB<sup>[4]</sup> 和 SUN<sup>[5]</sup> 数据集上的零样本图像分类性能表现。

不同设置下的 TransZero	CUB 数据集				SUN 数据集			
	acc	U	S	H	acc	U	S	H
TransZero (无特征增强编码器)	67.3	61.0	53.1	56.8	61.2	55.7	22.5	32.1
TransZero (编码器中不进行几何先验信息消除)	74.0	66.7	66.3	66.5	63.8	49.5	31.4	38.5
TransZero (无属性 → 视觉解码器)	62.3	53.3	54.1	53.7	58.3	35.0	28.8	31.6
TransZero (无自校准损失约束)	74.8	47.1	75.5	58.1	64.2	42.4	33.4	37.4
TransZero (无属性回归损失约束)	74.5	65.9	68.8	67.3	64.1	47.2	33.3	39.1
TransZero	76.8	69.3	68.3	68.8	65.6	52.6	33.4	40.8

**定性实验分析：**为直观地验证 TransZero 的有效性，本章节通过特征图可视化和视觉特征的 t-SNE 可视化<sup>[1]</sup>进行定性实验分析。

**特征图可视化：**为了直观地展示 TransZero 通过属性定位学习具有属性-视觉关键语义知识的局部增强视觉特征的有效性，本章节通过对现有基于注意力机制方法（例如，AREN<sup>[2]</sup>）和 TransZero 学习的特征图进行可视化。如图4.7所示，AREN 只学习图像中视觉表示的局部特征，例如整个鸟体，而忽略了细粒度语义属性信息。相比之下，本文提出的 Transzero 通过为关键属性（例如，图中 Acadian Flycatcher 的“bill shape all purpose”）进行视觉定位学习具有属性-视觉关键语义知识的局部增强视觉特征，缓解视觉-语义特征的表示差异性问题。因此，TransZero 有效提高零样本图像分类从已知类到未知类的知识迁移。

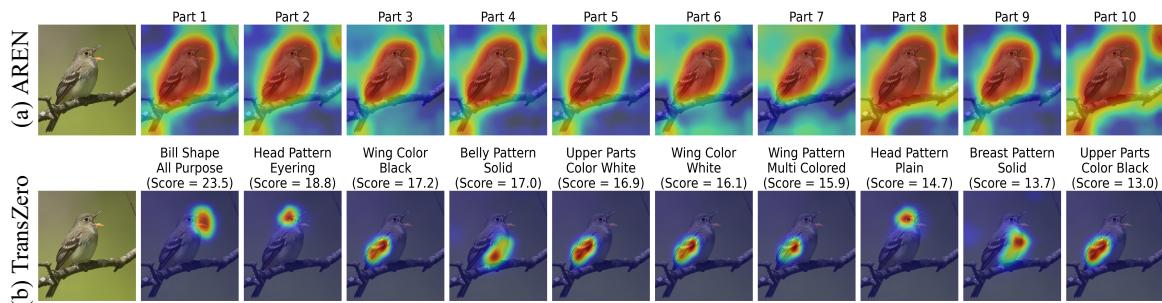


图 4.7 现有基于注意力机制方法（例如，AREN<sup>[2]</sup>）和 TransZero 学习的特征图可视化。

**视觉特征的 t-SNE 可视化：**如图4.8所示，本章节展示了不同的模型在 CUB 数据集的（a）已只类和（b）未知类学习的视觉特征 t-SNE<sup>[1]</sup> 可视化，例如，CNN

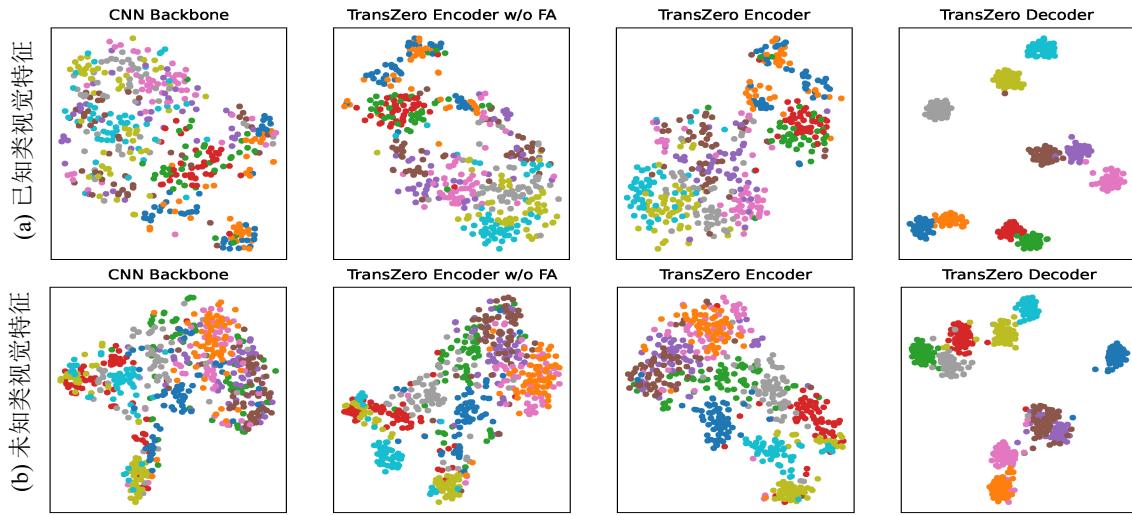


图 4.8 TransZero 和 CNN Backbone 在 CUB 数据集上学习的视觉特征 t-SNE<sup>[1]</sup> 可视化对比。

Backbone、不进行特征增强的 TransZero 编码器（TransZero w/o FA）、TransZero 编码器（TransZero Encoder）以及 TransZero 解码器（TransZero Decoder）。当 TransZero 使用标准编码器（即在编码器中不进行几何先验信息消除），编码器学习的视觉特征比从 CNN Backbone 提取的全局视觉特征在判别性方面有显著改善。当 TransZero 使用特征增强编码器增强视觉特征迁移性时，未知类的特征质量会进一步提高。此外，属性 → 视觉解码器学习视觉的属性定位表示属性-视觉关键公共语义知识，以提高视觉特征的局部判别性，使得 TransZero 在已知类和未知类上均能取得较好的分类效果。

**TransZero 和其他先进的零样本图像分类方法的实验结果对比：**本章节将 TransZero 和其他先进的零样本图像分类方法在 CUB<sup>[4]</sup>、SUN<sup>[5]</sup>、AWA2<sup>[6]</sup> 数据集上的 CZSL/GZSL 实验结果进行对比，包括端到端的方法（例如，QFSL<sup>[72]</sup>、LDF<sup>[73]</sup>、AREN<sup>[2]</sup>、LFGAA<sup>[74]</sup> 等）和非端到端的方法（例如，嵌入式方法（SP-AEN<sup>[108]</sup>、PQZSL<sup>[124]</sup>、DVBE<sup>[107]</sup>、DAZLE<sup>[81]</sup>、APN<sup>[41]</sup> 等）和生成式方法（f-CLSWGAN<sup>[89]</sup>、f-VAEGAN-D2<sup>[61]</sup>、Composer<sup>[132]</sup>、GCM-CF<sup>[123]</sup> 等））

本章节首先将 TransZero 与其他经典且先进的方法在 CZSL 设置中进行比较。如表4.2所示，本文提出的 TransZero 在 CUB 数据集和 SUN 数据集上分别达到了 76.8% 和 65.6% 的最佳分类精度。这表明 TransZero 可以有效地学习属性增强局部特征实现有效的知识迁移，从而在各种细粒度类上具有好的迁移性。对于粗粒度数据集（即

表 4.2 TransZero 和其他先进的零样本图像分类方法在 CUB<sup>[4]</sup>、SUN<sup>[5]</sup>、AWA2<sup>[6]</sup> 数据集上的 CZSL/GZSL 实验结果对比。

当前先进的零样本图像分类方法	CUB 数据集				SUN 数据集				AWA2 数据集			
	CZSL		GZSL		CZSL		GZSL		CZSL		GZSL	
	acc	U	S	H	acc	U	S	H	acc	U	S	H
<b>端到端</b>												
QFSL <sup>[72]</sup>	58.8	33.3	48.1	39.4	56.2	30.9	18.5	23.1	63.5	52.1	72.8	60.7
LDF <sup>[73]</sup>	67.5	26.4	81.6	39.9	—	—	—	—	65.5	9.8	87.4	17.6
SGMA* <sup>[40]</sup>	71.0	36.7	71.3	48.5	—	—	—	—	68.8	37.6	87.1	52.5
AREN* <sup>[2]</sup>	71.8	38.9	78.7	52.1	60.6	19.0	38.8	25.5	67.9	15.6	92.9	26.7
LFGAA* <sup>[74]</sup>	67.6	36.2	<b>80.9</b>	50.0	61.5	18.5	40.0	25.3	68.1	27.0	<b>93.4</b>	41.9
<b>非端到端</b>												
<b>生成式方法</b>												
f-CLSWGAN <sup>[89]</sup>	57.3	43.7	57.7	49.7	60.8	42.6	36.6	39.4	68.2	57.9	61.4	59.6
cycle-CLSWGAN <sup>[93]</sup>	58.4	45.7	61.0	52.3	60.0	49.4	33.6	40.0	66.3	56.9	64.0	60.2
f-VAEGAN-D2 <sup>[61]</sup>	61.0	48.4	60.1	53.6	64.7	45.1	38.0	41.3	71.1	57.6	70.6	63.5
LsrGAN <sup>[102]</sup>	60.3	48.1	59.1	53.0	62.5	44.8	37.7	40.9	—	53.1	68.8	60.0
E-PGN <sup>[111]</sup>	72.4	52.0	61.1	56.2	—	—	—	—	73.4	52.6	83.5	64.6
Composer <sup>[132]</sup>	69.4	56.4	63.8	59.9	62.6	<b>55.1</b>	22.0	31.4	<b>71.5</b>	62.1	77.3	68.8
GCM-CF <sup>[123]</sup>	—	61.0	59.7	60.3	—	47.9	37.8	42.2	—	60.4	75.1	67.0
<b>嵌入式方法</b>												
SP-AEN <sup>[108]</sup>	55.4	34.7	70.6	46.6	59.2	24.9	38.6	30.3	58.5	23.3	90.9	37.1
PQZSL <sup>[124]</sup>	—	43.2	51.4	46.9	—	35.1	35.3	35.2	—	31.7	70.9	43.8
IIR <sup>[75]</sup>	63.8	30.4	65.8	41.2	63.5	22.0	34.1	26.7	67.9	17.6	87.0	28.9
TCN <sup>[76]</sup>	59.5	52.6	52.0	52.3	61.5	31.2	37.3	34.0	71.2	61.2	65.8	63.4
DVBE <sup>[107]</sup>	—	53.2	60.2	56.5	—	45.0	37.2	40.7	—	<b>63.6</b>	70.8	67.0
DAZLE* <sup>[81]</sup>	66.0	56.7	59.6	58.1	59.4	52.3	24.3	33.2	67.9	60.3	75.7	67.1
APN* <sup>[41]</sup>	72.0	65.3	69.3	67.2	61.6	41.9	34.0	37.6	68.4	57.1	72.4	63.9
<b>TransZero (本文方法)</b>	<b>76.8</b>	<b>69.3</b>	68.3	<b>68.8</b>	<b>65.6</b>	52.6	33.4	40.8	70.1	61.3	82.3	<b>70.2</b>

注：符号“—”表示相应结果缺失，符号“\*”表示基于注意力机制的方法。

AWA2 数据集），TransZero 仍能取得极具竞争力的性能表现，最高精度为 70.1%。与其他基于注意的方法（如 SGMA<sup>[40]</sup>、AREN<sup>[2]</sup>、APN<sup>[41]</sup>）相比，TransZero 在 CUB 数据集和 SUN 数据集上分别获得了至少 4.8% 和 4.0% 的显著提升。这表明，TransZero 利用属性指导的 Transformer 学习具有属性-视觉关键公共语义知识的局部增强视觉特征比现有基于注意力机制方法<sup>[2,40,41,74,81]</sup>学习的局部特征更具判别性和迁移性。

从表4.2中也可以看出：大多数经典且先进的方法在已知类上取得了很好的结果，但在未知类上却取得较差的结果；而本文的 TransZero 方法可以很好地从已知类中迁移知识到未知类，同时在已知类和未知类上均取得不错的识别率。例如，TransZero 在 CUB 数据集和 AWA2 数据上分别取得了 68.8% 和 70.2% 的最优调和均值（H）。这归功于（1）TransZero 的特征增强编码器提高了视觉特征的转移性；（2）TransZero

中的属性 → 视觉解码器学习具有属性-视觉关键公共语义知识的局部增强视觉特征，有效促进零样本图像分类从已知类到未知类的知识迁移；(3) 自校准机制缓解了已知类偏差问题。此外，TransZero 也优于其他基于注意力的方法（例如，SGMA<sup>[40]</sup>、AREN<sup>[2]</sup>、LFGAA<sup>[74]</sup>、APN<sup>[41]</sup>，DAZLE<sup>[81]</sup> 等），在 CUB、SUN 和 AWA2 等三个数据集上分别取得了至少 1.6%、3.2% 和 3.1% 的调和均值领先。这表明了基于属性指导的 Transformer 在零样本图像分类任务中具有明显优势和巨大潜力。由于本章节方法是首次将 Transformer 引入到零样本图像分类领域，TransZero 可以作为零样本图像领域中的一个新基准方法，为后续工作提供新的研究思路。

## 4.3 基于互语义蒸馏网络的零样本图像分类

### 4.3.1 研究动机

如章节4.1分析，零样本图像分类方法的关键任务是在已知类的视觉特征和属性特征之间推断潜在的语义知识，实现已知类到未知类的语义知识迁移。早期基于全局视觉特征表示的零样本图像分类方法<sup>[67,72,73,75,89,100,101]</sup>不能对这些关键语义知识进行刻画，使得从已知类到未知类的语义知识迁移受阻。在属性信息或者其他额外语义信息的指导下，近期一些基于注意力的零样本图像分类方法<sup>[2,39,40]</sup>使用属性 → 视觉注意力网络尝试对这些公共局部知识进行挖掘，从而能够更准确地进行视觉-语义匹配实现语义知识迁移。然而，这些方法利用属性 → 视觉的单向注意力机制，只能挖掘视觉特征和属性特征之间有限的潜在公共语义知识。章节4.2提出的 TransZero 也是基于单向注意力机制的方法。因此，如何充分地挖掘视觉特征和属性特征之间关键的公共语义知识实现准确的零样本图像分类是仍然是一个重要的研究难题。

针对此问题，本章节工作提出基于互语义蒸馏网络的零样本图像分类方法（Mutually Semantic Distillation Network for Zero-Shot Image Classification，MSDN）。MSDN 由属性 → 视觉注意力子网络和视觉 → 属性注意力子网络组成，前者学习基于属性的视觉特征，后者学习基于视觉的属性特征。这两个子网络相互充当师生网络，并在语义蒸馏损失的约束下，它们在整个优化过程中彼此协作学习和相互指导。通过语义蒸馏，MSDN 可以学习更一致的基于属性的视觉特征和基于视觉的属性特征，从而有效地挖掘属性-视觉一致的语义知识（即关键公共语义知识），促进零样本图像分类进行有效的语义知识迁移。在三个主流标准数据集（例如，CUB 数据集<sup>[4]</sup>、

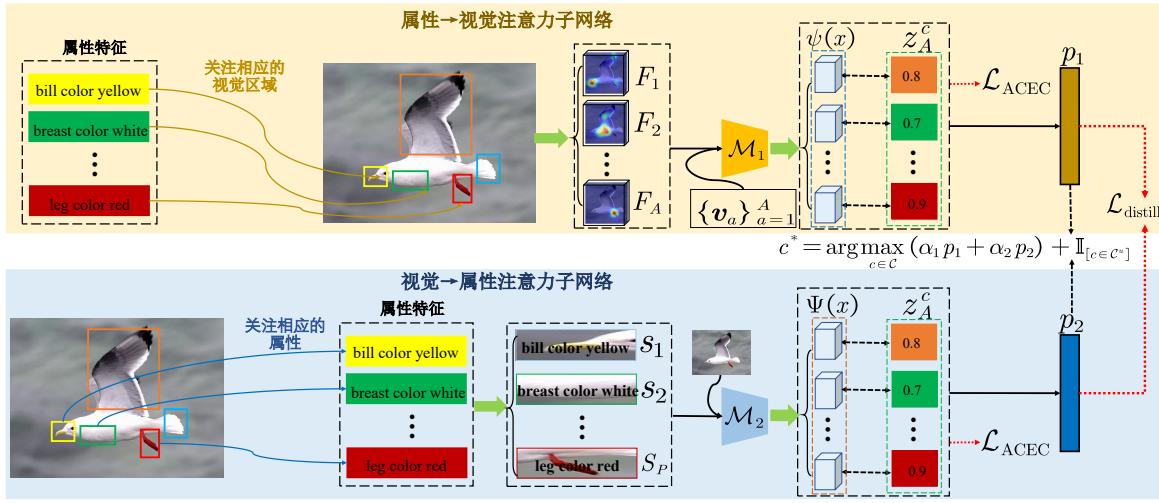


图 4.9 本章提出的 MSDN 模型结构示意图。

SUN 数据集<sup>[5]</sup> 和 AWA2 数据集<sup>[6]</sup> 上的大量定量和定性实验结果表明了 MSDN 的优越性和巨大潜力，并为零样本图像分类领域提供了一种新的基准方法。

### 4.3.2 互语义蒸馏网络

首先指出，本章节使用的基本符号定义见章节3.2.2。MSDN 的模型结构如图4.9所示，其包含属性  $\rightarrow$  视觉注意力子网络 ( $A \rightarrow V$ ) 和视觉  $\rightarrow$  属性注意力子网络 ( $V \rightarrow A$ )。在具有自校准机制的属性交叉熵损失约束下， $A \rightarrow V$  注意力子网络学习基于属性的视觉特征，而  $V \rightarrow A$  注意力子网络学习基于视觉的属性特征。在语义蒸馏损失的指导下，两个子网在整个优化过程中协作学习和相互教导，使得两个网络学习的基于属性的视觉特征和基于视觉的属性特征具有更强的一致语义表示，从而有效挖掘属性-视觉特征之间的关键公共语义知识，并取得领先的零样本图像分类性能。此外，MSDN 在训练过程中没有使用真实的未知类视觉特征，因此它也是一种归纳式方法。接下来将对具体模型和优化过程进行详细介绍。

**属性  $\rightarrow$  视觉注意力子网络：** 学习具有属性定位的视觉特征对零样本图像分类的知识迁移非常有效<sup>[2,40,41,81]</sup>。为此，本文提出了一个属性  $\rightarrow$  视觉注意力子网络 ( $A \rightarrow V$ ) 用于对视觉中相关的图像区域进行属性定位，从而为图像中每个属性提取相应基于属性的视觉特征。 $A \rightarrow V$  基于两个输入：图像的一组区域视觉特征  $\mathcal{R}_P = \{r_p\}_{p=1}^P$ ，每个视觉特征对图像中的一个区域进行编码；一组属性特征  $\mathcal{V}_A = \{\mathbf{v}_a\}_{a=1}^A$ 。 $A \rightarrow V$  可以关注每个属性的图像区域，并将每个属性关注的视觉区域特征进行比较，以确定每

# 华 中 科 技 大 学 博 士 学 位 论 文

---

个属性在相应视觉区域的相关性。对于第  $a$  个属性， $A \rightarrow V$  首先将其与一张图像的第  $r$  区域的注意力权重定义为：

$$\beta_a^p = \frac{\exp(v_a^\top W_1 r_p)}{\sum_{a=1}^A \exp(v_a^\top W_1 r_p)}, \quad (4.19)$$

其中  $W_1$  是一个可学习的矩阵，用于计算每个区域的视觉特征与每个属性特征之间的相似性。因此， $A \rightarrow V$  学习到一组注意力权重  $\{\beta_a^p\}_{p=1}^P$ 。

然后， $A \rightarrow V$  根据注意权重为每个属性提取相应基于属性的视觉特征。例如， $A \rightarrow V$  可以得到基于第  $a$  属性的视觉特征  $F_a$ ，其公式如下：

$$F_a = \sum_{p=1}^P \beta_a^p r_p. \quad (4.20)$$

直观上看， $F_a$  捕获用于定位图像中相应语义属性的视觉证据。如果图像具有明显的属性  $v_a$ ，则模型将为第  $a$  个属性分配较高的注意力值，否则，模型将第  $a$  属性分配一个较低的注意力值。给定一组属性特征  $\mathcal{V}_A = \{v_a\}_{a=1}^A$ ， $A \rightarrow V$  学习到一组基于属性的视觉特征  $F = \{F_1, F_2, \dots, F_A\}$ 。

$A \rightarrow V$  提取到图像基于属性的视觉特征之后，其进一步使用一个视觉-语义映射器  $\mathcal{M}_1$  将其映射到类语义空间，并用于后续的最近邻匹配实现零样本分类。为了使映射更加准确， $A \rightarrow V$  利用属性特征  $\mathcal{V}_A = \{v_a\}_{a=1}^A$  作为辅助避免信息损失。具体而言， $\mathcal{M}_1$  将基于属性的视觉特征  $F_a$  与其对应的属性特征  $v_a$  进行匹配：

$$\psi_a = \mathcal{M}_1(F_a) = v_a^\top W_2 F_a, \quad (4.21)$$

其中， $W_2$  是一个嵌入矩阵，它将  $F$  嵌入类语义空间。本质上， $\psi_a$  类似于类语义向量  $(z_a^c)$  是一个属性值，表示图像中具有第  $a$  个属性的置信度。最后， $A \rightarrow V$  为每个图像提取相应的语义嵌入  $\psi(x) = \{\psi_1, \psi_2, \dots, \psi_A\}$ 。

**视觉  $\rightarrow$  属性注意力子网络：**类似地，MSDN 进一步设计了一个视觉  $\rightarrow$  属性注意力子网络 ( $V \rightarrow A$ ) 学习基于视觉的属性特征。 $V \rightarrow A$  与  $A \rightarrow V$  构成一个属性-视觉的双向注意力机制，它们学习的基于视觉的属性特征基于属性的视觉特征形成互补关系。它们通过协作学习实现相互指导和校准，并挖掘属性-视觉特征之间的潜在语义表示提高零样本图像分类效果。 $V \rightarrow A$  首先关注每个图像区域对应的语义属性。形式上，

$V \rightarrow A$  为每个区域视觉特征  $r_p$  和第  $a$  属性特征定义注意力权重:

$$\tau_p^a = \frac{\exp(r_p^\top W_3 v_a)}{\sum_{p=1}^P \exp(r_p^\top W_3 v_a)}, \quad (4.22)$$

其中,  $W_3$  是一个可学习矩阵, 用于度量属性特征和每个视觉区域特征之间的相似性。因此,  $V \rightarrow A$  可以得到一组注意力权重  $\{\tau_p^a\}_{a=1}^A$ , 用于提取基于视觉的属性特征:

$$S_p = \sum_{a=1}^A \tau_p^a v_a. \quad (4.23)$$

本质上,  $S_p$  是视觉的语义属性表示。 $V \rightarrow A$  进一步引入视觉-语义映射器  $\mathcal{M}_2$ , 将基于视觉的属性特征  $S_p$  映射到语义空间:

$$\bar{\Psi}_p = \mathcal{M}_2(S_p) = r_p^\top W_4 S_p, \quad (4.24)$$

其中,  $W_4$  是嵌入矩阵。给定一张图像的一组区域视觉特征  $\mathcal{R}_P = \{r_p\}_{p=1}^P$ ,  $V \rightarrow A$  学到这张图像的类语义嵌入  $\bar{\Psi}(x) = \{\bar{\Psi}_1, \bar{\Psi}_2, \dots, \bar{\Psi}_P\}$ 。为了使  $V \rightarrow A$  学习到的类语义嵌入  $\bar{\Psi}(x)$  (维度为  $P$ ) 与  $A \rightarrow V$  学习的类语义嵌入  $\psi(x)$  (维度为  $A$ ) 相匹配, MSDN 进一步将  $\bar{\Psi}(x)$  映射到维度为  $A$  的类语义空间中:

$$\Psi(x) = \bar{\Psi}(x) \times Att = \bar{\Psi}(x) \times (\mathcal{R}_P^\top W_{att} \mathcal{V}_A), \quad (4.25)$$

其中  $W_{att}$  为可学习的矩阵。为此,  $\psi(x)$  和  $\Psi(x)$  形成互补的类语义嵌入表示。

**模型优化:** 为了实现 MSDN 的模型优化, 每个注意力子网络均使用具有自校准机制的属性交叉熵损失<sup>[40,41,81]</sup>进行训练。同时, 本章节提出语义蒸馏损失指导两个注意子网进行协作学习和相互教导, 使得两个子网络学习的基于属性的视觉特征和基于视觉的属性特征具有更强的一致语义表示。

**具有自校准机制的属性交叉熵损失:** 由于当相关属性直观地出现在图像中时, 相关图像的视觉-语义嵌入被投影在其类语义向量  $z^c$  附近, 因此 MSDN 采用具有自校准机制的属性交叉熵损失 ( $\mathcal{L}_{ACEC}$ ) 对两个子网络进行优化。这使得图像的视觉特征与其对应的类语义向量实现更准确的最近邻匹配。给定一批  $n_b$  训练样本  $\{x_i\}_{i=1}^{n_b}$

# 华 中 科 技 大 学 博 士 学 位 论 文

---

及其相关的类语义向量  $z^c$ ,  $\mathcal{L}_{\text{ACEC}}$  定义为:

$$\begin{aligned} \mathcal{L}_{\text{ACEC}} = & -\frac{1}{n_b} \sum_{i=1}^{n_b} [\log \frac{\exp(f(x_i) \times z^c)}{\sum_{\hat{c} \in \mathcal{C}^s} \exp(f(x_i) \times z^{\hat{c}})} \\ & - \lambda_{\text{cal}} \sum_{c'=1}^{|C^u|} \log \frac{\exp(f(x_i) \times z^{c'} + \mathbb{I}_{[c' \in \mathcal{C}^u]})}{\sum_{\hat{c} \in \mathcal{C}} \exp(f(x_i) \times z^{\hat{c}} + \mathbb{I}_{[\hat{c} \in \mathcal{C}^u]})}], \end{aligned} \quad (4.26)$$

其中,  $\lambda_{\text{cal}}$  是一个损失权重,  $\mathbb{I}_{[c \in \mathcal{C}^u]}$  是一个指示函数 (当  $c \in \mathcal{C}^u$  时, 其值为 1, 否则为-1)。在优化 A→V 子网络时,  $f(x_i) = \psi(x_i)$ ; 在优化 V→A 子网络时,  $f(x_i) = \Psi(x_i)$ 。 $\mathcal{L}_{\text{ACEC}}$  鼓励在训练期间给未知类分配更大的非零概率, 当给定未知类测试样本时, MSDN 为相应真实的未知类样本分配更大的预测概率。

**语义蒸馏损失:** 为了使 A→V 和 V→A 这两个互补的子网能够在优化过程中协作学习并相互教导, 本章节提出了语义蒸馏损失 (Semantic Distillation Loss,  $\mathcal{L}_{\text{distill}}$ )。 $\mathcal{L}_{\text{distill}}$  由两个注意力子网络的概率预测 (例如,  $p_1 = \{\psi(x_i) \times z^1, \dots, \psi(x_i) \times z^C\}$ ,  $p_2 = \{\Psi(x_i) \times z^1, \dots, \Psi(x_i) \times z^C\}$ ) 之间的 Jensen-Shannon 散度 (Jensen-Shannon Divergence, JSD) 和的  $\ell_2$  距离组成, 表示为:

$$\begin{aligned} \mathcal{L}_{\text{distill}} = & \frac{1}{n_b} \sum_{i=1}^{n_b} [\underbrace{\frac{1}{2} (D_{KL}(p_1(x_i) \| p_2(x_i)) + D_{KL}(p_2(x_i) \| p_1(x_i)))}_{\text{JSD}} \\ & + \underbrace{\|p_1(x_i) - p_2(x_i)\|_2^2}_{\ell_2}], \end{aligned} \quad (4.27)$$

其中,

$$D_{KL}(p || q) = \sum_{c=1}^{|C^s|} p^c \log(\frac{p^c}{q^c}). \quad (4.28)$$

最后, MSDN 的总体优化函数为:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ACEC}} + \lambda_{\text{distill}} \mathcal{L}_{\text{distill}}, \quad (4.29)$$

其中,  $\lambda_{\text{distill}}$  用于从控制语义蒸馏损失项的权重。

**零样本图像分类:** MSDN 优化完成之后, 首先根据 A→V 和 V→A 两个子网络获得测试样本  $x_i$  的类语义嵌入, 分别为  $\psi(x)$  和  $\Psi(x)$ 。然后, MSDN 使用两个组合系数  $(\alpha_1, \alpha_2)$  对其两个子网络的语义嵌入预测融合起来进行最近邻匹配, 并通过自校准缓解已知类偏差问题<sup>[41,81]</sup>, 公式如下:

$$c^* = \arg \max_{c \in \mathcal{C}^u / \mathcal{C}} (\alpha_1 \psi(x_i) + \alpha_2 \Psi(x_i))^T \times z^c + \mathbb{I}_{[c \in \mathcal{C}^u]}. \quad (4.30)$$

$\mathcal{C}^u/\mathcal{C}$  分别表示 MSDN 在 CZSL/GZSL 设置下进行分类。

#### 4.3.3 实验结果与分析

为验证 MSDN 方法的有效性，本文在三个主流的零样本图像分类标准数据集上同时进行 CZSL 和 GZSL 的实验，包括两个细粒度数据集（CUB<sup>[4]</sup>, SUN<sup>[5]</sup>）和一个粗粒度数据集（AWA2<sup>[6]</sup>）。本文依据 Xian 等人<sup>[6]</sup>最新的数据集划分方式进行模型训练和测试，数据集的详细介绍见章节2.3.1。本文利用在 ImageNet 上预训练的 ResNet-101 作为 CNN Backbone 对图像样本（分辨率裁剪为  $448 \times 448$ ）提取相应的视觉特征。MSDN 使用带有超参数的 RMSProp 优化器（momentum = 0.9, weight decay = 0.0001）进行模型优化。模型的学习率和批量大小分别设置为 0.0001 和 50。根据验证实验结果，本文将损失权重  $\{\lambda_{\text{cal}}, \lambda_{\text{distill}}\}$  在 CUB 和 AWA2 数据集上均设置为 {0.1, 0.001}，在 SUN 数据集上设置为 {0.0, 0.01}。

本章节通过以下实验验证 MSDN 的有效性：

- 超参实验分析；
- 消融实验分析；
- 定性实验分析；
- MSDN 和其他经典且先进的零样本图像分类方法在 CZSL/GZSL 不同设置上的实验结果对比。

接下来将对不同的实验进行详细阐述和分析。本章节方法 MSDN 的源代码地址：  
<https://github.com/shiming-chen/MSDN>。

**超参实验分析：**模型 MSDN 的主要超参数有包括公式4.30中的组合系数  $(\alpha_1, \alpha_2)$ 、公式4.26和4.29中的损失权重  $\lambda_{\text{cal}}$  和  $\lambda_{\text{distill}}$ 。本文将对这四个超参数在 CUB 数据集和 SUN 数据集上做相应的实验分析，并进行超参数设置。

**两个子网络的组合系数  $(\alpha_1, \alpha_2)$ ：**本章节通过实验验证  $A \rightarrow V$  和  $V \rightarrow A$  两个注意力子网的组合系数  $(\alpha_1, \alpha_2)$  对 MSDN 的影响。如图所示，当  $\alpha_1, \alpha_2$  设置得太小或太大时，MSDN 的性能表现不佳，因为基于属性的视觉特征和基于视觉的属性特征均不能单独地对属性-视觉特征之间的关键语义知识进行充分挖掘。结果显示，当组合系数  $(\alpha_1, \alpha_2)$  在 CUB 数据集和 SUN 数据集上分别设置为 (0.9,0.1) 和 (0.7,0.3) 时，MSDN 获得较好的性能。

**损失权重  $\lambda_{\text{cal}}$  和  $\lambda_{\text{distill}}$ ：**本章节研究损失权重  $\lambda_{\text{cal}}$  和  $\lambda_{\text{distill}}$  对 MSDN 的影响，它们

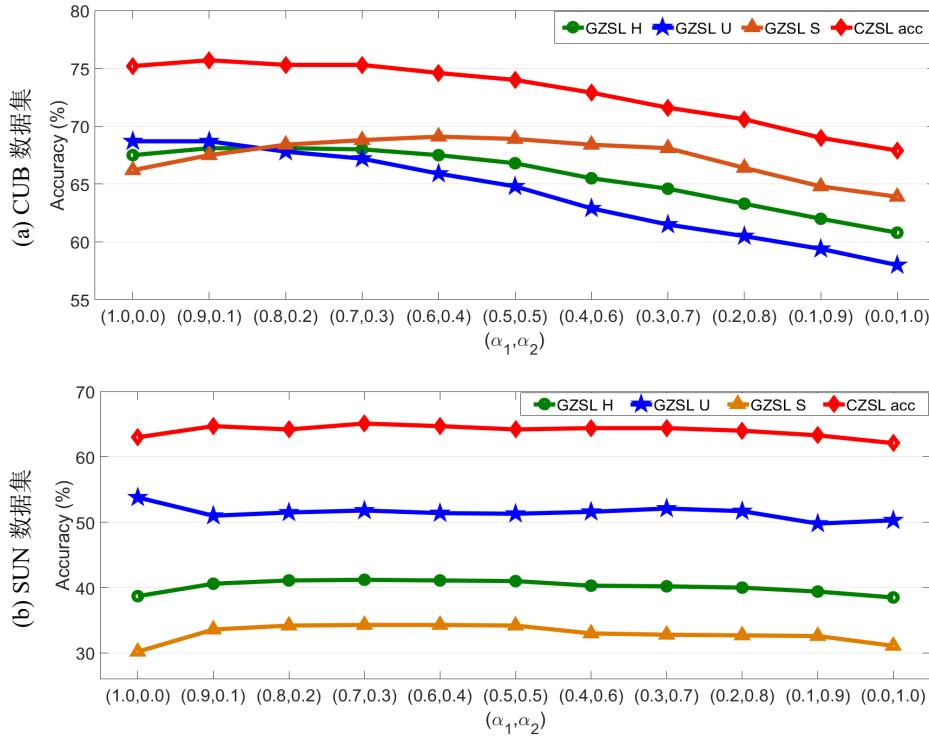
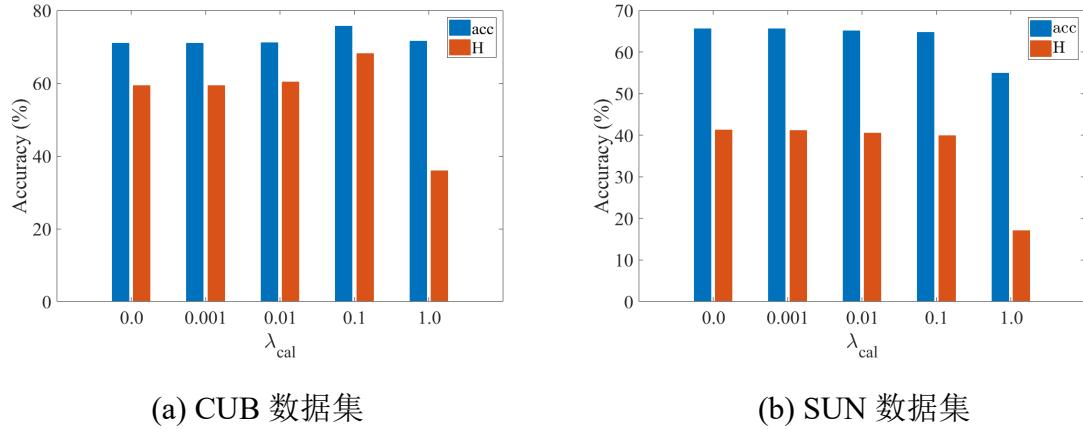
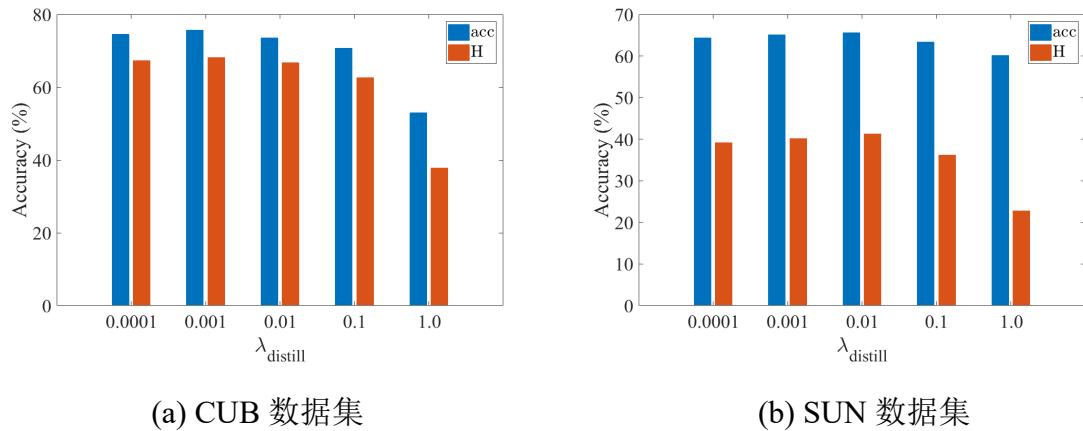


图 4.10 两个注意力子网络的组合系数  $(\alpha_1, \alpha_2)$  对 MSDN 的影响。

分别控制公式4.26中的自校准项和公式4.27中的语义蒸馏损失项。根据图4.11的实验结果，本文 MSDN 的损失权重  $\lambda_{cal}$  在 CUB/AWA2 数据集上设置为 0.1。由于 SUN 数据集的未知类数量远远大于已知类数量，使得零样本图像分类模型通常在 SUN 数据集上取得的未知类识别精度高于已知类精度。因此，对于 SUN 数据集，本章节将  $\lambda_{cal}$  设置为 0。根据图4.12的实验结果，MSDN 的损失权重  $\lambda_{distill}$  在 CUB/AWA2 数据和 SUN 数据集上分别设置为 0.001 和 0.01。

**消融实验分析：**为了进一步深入验证 MSDN 的有效性，本章节进行了消融实验分析，评估 MSDN 的  $V \rightarrow A$  注意力子网络（例如，MSDN( $V \rightarrow A$ ) 无  $\mathcal{L}_{distill}$ ）、 $A \rightarrow V$  注意力子网络（例如，MSDN( $A \rightarrow V$ ) 无  $\mathcal{L}_{distill}$ ），语义蒸馏损失（例如，MSDN( $V \rightarrow A$ ) 有  $\mathcal{L}_{distill}$ 、MSDN( $A \rightarrow V$ ) 有  $\mathcal{L}_{distill}$ ），基于 JSD 的语义蒸馏损失（例如，MSDN 有基于 JSD 的  $\mathcal{L}_{distill}$ ），基于  $\ell_2$  的语义蒸馏损失（MSDN 有基于  $\ell_2$  的  $\mathcal{L}_{distill}$ ）。本章节使用的基准方法表示其直接从 CNN Backbone 提取的全局视觉特征用于嵌入式零样本图像分类，模型流程如图4.13所示。实验结果如表4.3所示。与基准方法相比，MSDN 在没有语义蒸馏损失的约束时只使用单个注意子网学习属性-视觉的潜在语义知识用


 图 4.11 自校准权重 ( $\lambda_{SC}$ ) 对 MSDN 的影响。

 图 4.12 语义蒸馏损失权重 ( $\lambda_{distill}$ ) 对 MSDN 的影响。

于视觉-语义交互，促进从已知类到未知类的知识迁移，实现了显著的性能提升。例如，MSDN (V → A) 在 CUB 数据集和 SUN 数据集上分别获得了性能指标  $acc/H$  为 8.6%/6.3% 和 4.4%/3.3% 的提升，而 MSDN (A → V) 在 CUB 数据集和 SUN 数据集上的性能指标  $acc/H$  获得了 16.0%/16.3% 和 9.0%/8.0% 提升。使用语义蒸馏损失对 MSDN 进一步优化，使得 MSDN 的两个子注意力网络可以协作学习和相互教导，实现属性-视觉特征之间的语义知识蒸馏并促进有效的知识迁移，使零样本分类性能得到进一步提高。例如，MSDN(A→V) 有  $\mathcal{L}_{distill}$  情况下，在 CUB 数据集和 SUN 数据集上的性能指标  $H$  分别进一步提高了 5.4% 和 4.8%。当语义蒸馏损失仅使用一个距离进行表示时（即 JSD 或  $\ell_2$ ），MSDN 也能实现语义知识蒸馏，但效果不如两个

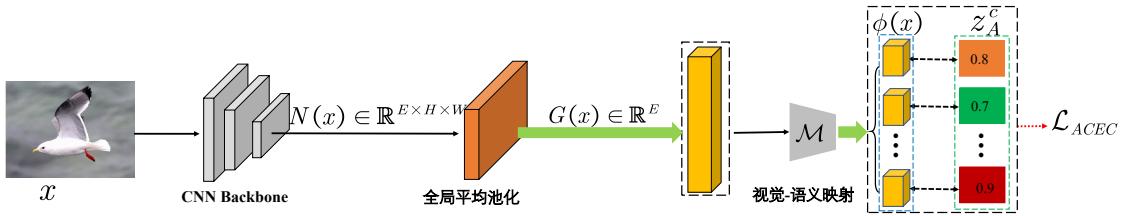


图 4.13 本章节使用的基准方法 (baseline) 模型结构示意图。

 表 4.3 在不同模型成分设置下，MSDN 在 CUB 数据集<sup>[4]</sup> 和 SUN 数据集<sup>[5]</sup> 上的实验结果。

不同设置下的 MSDN	CUB				SUN			
	acc	U	S	H	acc	U	S	H
基准方法	57.4	44.2	55.2	49.1	54.8	30.3	30.7	30.5
MSDN(V→A) 无 $\mathcal{L}_{\text{distill}}$	66.0	48.2	65.2	55.4	59.2	41.0	28.8	33.8
MSDN(A→V) 无 $\mathcal{L}_{\text{distill}}$	73.4	66.8	64.2	65.4	63.8	48.5	31.6	38.5
MSDN(V→A) 有 $\mathcal{L}_{\text{distill}}$	67.9	58.0	63.9	60.8	62.1	51.7	30.8	38.6
MSDN(A→V) 有 $\mathcal{L}_{\text{distill}}$	75.2	68.7	66.2	67.5	63.0	53.8	30.2	38.7
MSDN 有基于 JSD 的 $\mathcal{L}_{\text{distill}}$	74.3	66.6	67.5	67.0	64.7	51.0	32.2	39.4
MSDN 有基于 $\ell_2$ 的 $\mathcal{L}_{\text{distill}}$	74.4	68.2	67.1	67.6	64.9	54.0	32.8	40.8
MSDN	76.1	68.7	67.5	68.1	65.8	52.2	34.2	41.3

距离同时约束情况下的 MSDN。此外，由于 A→V 子网络学习的基于属性的视觉特征和 V→A 子网络学习的基于视觉的属性特征是能够互补，因此融合 MSDN 两个子网络的特征用于零样本图像分类可以有效提高分类性能，其在 CUB 数据集和 SUN 数据集上的性能指标  $acc/H$  分别比基准方法提高了 18.7%/19.0% 和 11.0%/10.8%。

**定性实验分析：**为直观地验证 MSDN 的有效性，本章节通过特征图可视化和视觉特征的 t-SNE 可视化<sup>[1]</sup> 进行定性实验分析。

**特征图可视化：**为了直观地展示 MSDN 提取属性-视觉特征之间关键公共语义知识的有效性，本章节将 MSDN 中两个注意力子网络（即 A→V 和 V→A）学习的特征图进行可视化。如图4.14所示，A→V 和 V→A 两个注意力子网络学习的基于属性的视觉特征和基于视觉特征属性特征均能有效表示属性-视觉之间的关键公共语义知识。一方面，两个子网络学习的大部分关键属性-视觉语义知识具有较高的一致性（例如，Elegant Tern 中的关键属性 “Bill Color Orange”、“Bill Shape Dagger”、“Leg Color Black” 等）。另一方面，两个子网络能各自挖掘其他潜在的重要属性-视觉关键语义知识（例如，Elegant Tern 中的关键属性 “Shape Chicken Like Marsh”，“Upper Tail

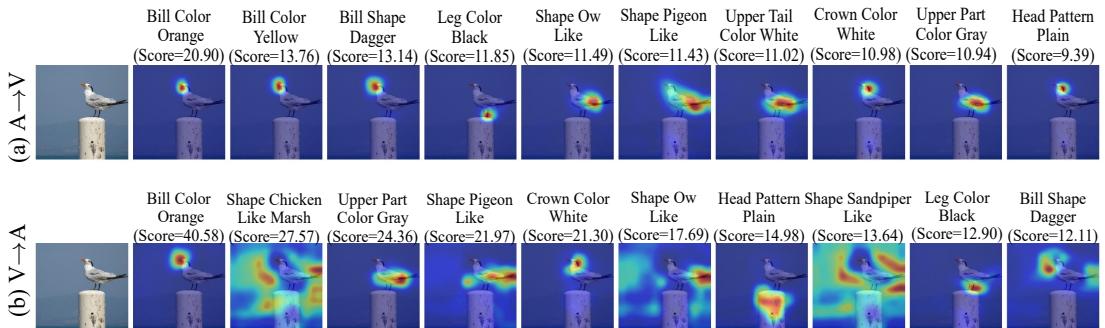
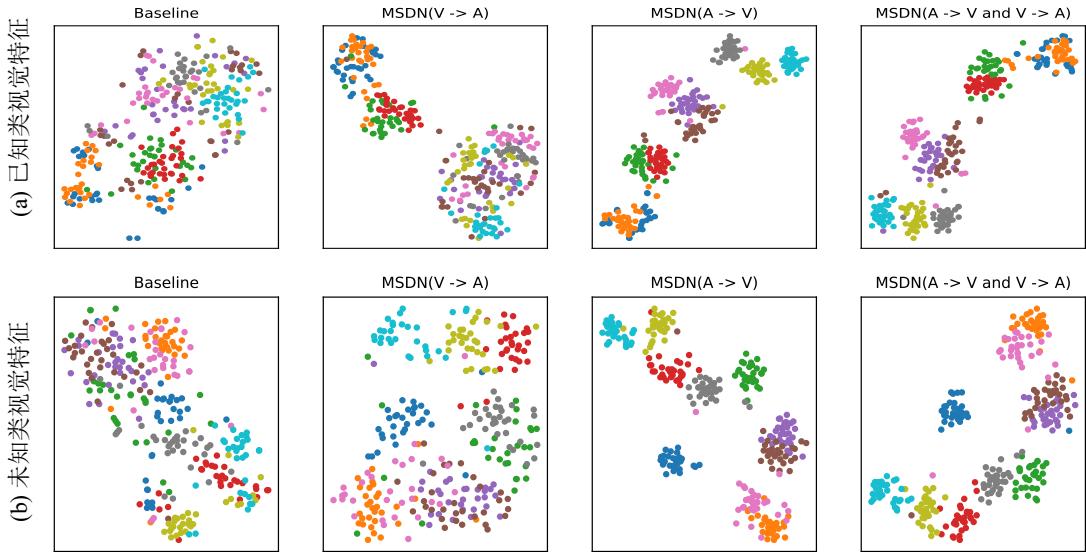


图 4.14 MSDN 的两个注意力子网络学习的特征图可视化。


 图 4.15 MSDN 的不同模型和基准方法在 CUB 数据集上学习的视觉特征 t-SNE<sup>[1]</sup> 可视化。

*Color White*”等), 形成互补的关键属性。MSDN 将两个子网络的挖掘的关键语义知识进行融合, 从而能更好地表示属性-视觉特征之间的关键公共语义知识, 进一步促进零样本图像分类的知识迁移。此外, 从结果中可以发现,  $V \rightarrow A$  子注意力网络能挖掘更高置信度的重要语义属性, 不同于当前大部分基于注意力机制方法<sup>[40,41,81,132]</sup> 使用  $A \rightarrow V$  学习局部特征。后期工作可以深入探索基于  $V \rightarrow A$  子注意力网络的方法。

视觉特征的 t-SNE 可视化: 如图4.15所示, 本章节展示了基准方法、MSDN 的  $V \rightarrow A$  子网络、MSDN 的  $A \rightarrow V$  子网络、MSDN 完整模型在 CUB 数据集上学习的已知类和未知类视觉特征的 t-SNE 可视化<sup>[1]</sup>。结果直观地显示, 相较于基准方法, MSDN 的不同模型显著的提高视觉特征的判别性, 有效促进零样本图像的知识迁移

表 4.4 MSDN 和其他先进的零样本图像分类方法在 CUB<sup>[4]</sup>、SUN<sup>[5]</sup>、AWA2<sup>[6]</sup> 数据集上的 CZSL/GZSL 实验结果对比。

当前先进的零样本图像分类方法	CUB 数据集				SUN 数据集				AWA2 数据集			
	CZSL		GZSL		CZSL		GZSL		CZSL		GZSL	
	acc	U	S	H	acc	U	S	H	acc	U	S	H
<b>生成式方法</b>												
f-CLSWGAN <sup>[89]</sup>	57.3	43.7	57.7	49.7	60.8	42.6	36.6	39.4	68.2	57.9	61.4	59.6
cycle-CLSWGAN <sup>[93]</sup>	58.4	45.7	61.0	52.3	60.0	49.4	33.6	40.0	66.3	56.9	64.0	60.2
f-VAEGAN <sup>[61]</sup>	61.0	48.4	60.1	53.6	64.7	45.1	38.0	41.3	71.1	57.6	70.6	63.5
LsrGAN <sup>[102]</sup>	60.3	48.1	59.1	53.0	62.5	44.8	37.7	40.9	—	53.1	68.8	60.0
OCD-CVAE <sup>[125]</sup>	—	44.8	59.9	51.3	—	44.8	42.9	43.8	—	59.5	73.4	65.7
GCM-CF <sup>[123]</sup>	—	61.0	59.7	60.3	—	47.9	37.8	42.2	—	60.4	75.1	67.0
<b>公共子空间式方法</b>												
DeViSE <sup>[68]</sup>	52.0	23.8	53.0	32.8	56.5	16.9	27.4	20.9	54.2	17.1	74.7	27.8
DCN <sup>[46]</sup>	56.2	28.4	60.7	38.7	61.8	25.5	37.0	30.2	65.2	25.5	84.2	39.1
CADA-VAE <sup>[3]</sup>	59.8	51.6	53.5	52.4	61.7	47.2	35.7	40.6	63.0	55.8	75.0	63.9
SGAL <sup>[105]</sup>	—	40.9	55.3	47.0	—	35.5	34.4	34.9	—	52.5	86.3	65.3
<b>嵌入式方法</b>												
SP-AEN <sup>[108]</sup>	55.4	34.7	70.6	46.6	59.2	24.9	38.6	30.3	58.5	23.3	90.9	37.1
SGMA* <sup>[40]</sup>	71.0	36.7	71.3	48.5	—	—	—	—	68.8	37.6	87.1	52.5
AREN* <sup>[2]</sup>	71.8	38.9	78.7	52.1	60.6	19.0	38.8	25.5	67.9	15.6	92.9	26.7
LFGAA* <sup>[74]</sup>	67.6	36.2	80.9	50.0	61.5	18.5	40.0	25.3	68.1	27.0	93.4	41.9
DAZLE* <sup>[81]</sup>	66.0	56.7	59.6	58.1	59.4	52.3	24.3	33.2	67.9	60.3	75.7	67.1
APN*(NeurIPS'20) <sup>[41]</sup>	72.0	65.3	69.3	67.2	61.6	41.9	34.0	37.6	68.4	57.1	72.4	63.9
<b>MSDN (本文方法)</b>	<b>76.1</b>	<b>68.7</b>	<b>67.5</b>	<b>68.1</b>	<b>65.8</b>	<b>52.2</b>	<b>34.2</b>	<b>41.3</b>	70.1	<b>62.0</b>	<b>74.5</b>	<b>67.7</b>

注：符号“—”表示相应结果缺失。符号“\*”表示基于注意力机制的零样本图像分类方法。

并取得显著的性能提升。这归功于 MSDN 的不同模型均能有效的挖掘属性-视觉特征之间的关键公共语义知识，并用于具有强表征能力的特征表示。

**MSDN 和其他先进的零样本图像分类方法的实验结果对比：**本章节将 MSDN 与其他先进的零样本图像分类方法在 CUB<sup>[4]</sup>、SUN<sup>[5]</sup>、AWA2<sup>[6]</sup> 数据集上的 CZSL/GZSL 实验结果进行对比，包括嵌入式方法（例如，SP-AEN<sup>[108]</sup>，SGMA<sup>[40]</sup>，AREN<sup>[2]</sup>，LFGAA<sup>[74]</sup>，DAZLE<sup>[81]</sup>，APN<sup>[41]</sup>）、生成式方法（例如，f-CLSWGAN<sup>[89]</sup>，cycle-CLSWGAN<sup>[93]</sup>，f-VAEGAN<sup>[61]</sup>，LsrGAN<sup>[102]</sup>，GCM-CF<sup>[123]</sup>，ZeroNAS<sup>[131]</sup>等）和公共子空间式方法（例如，DeViSE<sup>[68]</sup>，DCN<sup>[46]</sup>，CADA-VAE<sup>[3]</sup>，SGAL<sup>[105]</sup>等）。

本章节首先将 MSDN 与其他先进的零样本图像方法在 CZSL 设置下进行比较。表4.4展示了不同方法在 CUB<sup>[4]</sup>、SUN<sup>[5]</sup>、AWA2<sup>[6]</sup> 数据集上的实验结果。相相较于其他方法，MSDN 在 CUB 数据集和 SUN 数据集上均取得了领先的结果，识别精度 *acc* 分别为 76.1% 和 65.8%。这表明 MSDN 有效地挖掘了属性-视觉特征之间的关键公共语义知识用于表示判别性特征，从而区分细粒度未知类。如图4.1所示，

MSDN 可以从 Alifornia Gull、Parakeet Auklet、Pigeon Guillemot 等已知类上学习 “*bill color yellow*”、“*breast color white*” 和 “*leg color red*” 的关键语义属性，并迁移关键语义属性到未知类 Red Legged Kittiwake 实现未知类识别。对于粗粒度数据集（例如，AWA2），MSDN 仍然取得了有竞争力的性能表现，识别精度  $acc$  为 70.1%。和当前先进的嵌入式方法对比（例如，SP-AEN<sup>[108]</sup>，SGMA<sup>[40]</sup>，AREN<sup>[2]</sup>，LFGAA<sup>[74]</sup>，DAZLE<sup>[81]</sup>，APN<sup>[41]</sup>），MSDN 在 CUB、SUN、AWA2 数据集上的性能指标  $acc$  至少提升了 3.9%、4.2%、1.7%。

表4.4还展示了 GZSL 设置下不同方法的结果。结果显示，大多数先进的方法在 CUB 数据集和 AWA2 数据集上的已知类都取得了较好的识别精度，但在未知类上的结果却明显交叉，而本章节提出的 MSDN 同时在已知类和未知类取得较平衡的高识别精度。因此，MSDN 在 CUB 数据集和 AWA2 数据上取得的调和均值  $H$  分别达到了 68.1% 和 67.7% 的领先性能。这结果得益于 MSDN 有效挖掘属性-视觉特征之间的关键公共语义属性知识，促进语义知识从已知类迁移到未知类实现有效的视觉-语义交互。MSDN 和其他基于注意力的先进方法<sup>[?]</sup>相比，MSDN 挖掘的关键语义知识具有更强的判别性，有效区分细粒度类别，使其在 SUN 数据集上的调和均值  $H$  至少提升了 3.7%。这些结果表明了 MSDN 在零样本图像分类里具有明显的优势和潜力。

## 4.4 联合属性指导 Transformer 和互语义蒸馏网络的零样本图像分类

### 4.4.1 研究动机

为有效挖掘属性-视觉特征之间的关键公共语义知识，章节4.2提出了 TransZero 使用属性指导的 TransFormer 学习视觉的属性定位表示重要语义信息，章节4.3提出了 MSDN 使用两个注意力子网络和语义蒸馏损失进行语义协作学习实现潜在语义知识的充分探索。TransZero 基于单向的属性 → 视觉的 Transformer，而 MSDN 基于双向的简单注意力机制（例如，属性 → 视觉和视觉 → 属性注意力网络）。虽然这两种方法相较于现有基于注意力机制的方法（例如，AREN<sup>[2]</sup>）能更好的挖掘属性-视觉之间的语义知识，但他们挖掘的语义知识不具备更高置信度（即对类别的某些关键属性不能进行更突出的表示）或者挖掘的潜在语义知识不够充分，可能造成关键属性和普通属性不具备更明显的差异性表示<sup>[81,82]</sup>，如图所示4.16。特别对于细粒度类别分类，不同的细粒度类别的属性大部分相近，只能以少数关键语义属性对不同的

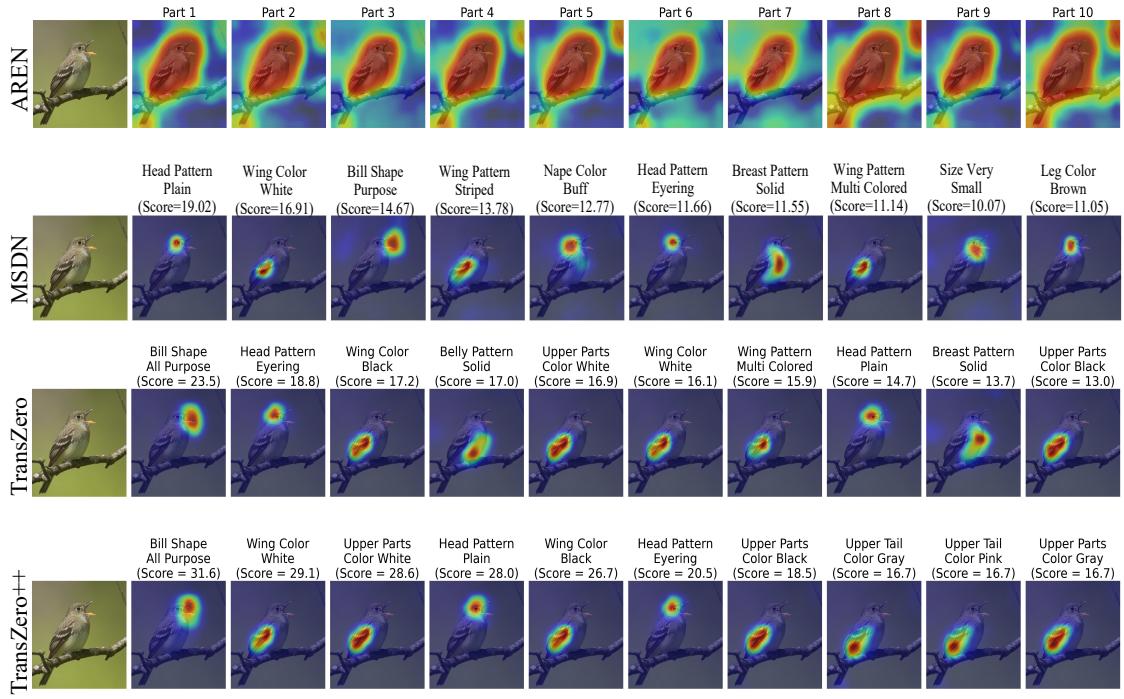


图 4.16 不同方法学习的特征图可视化对比。



图 4.17 不同细粒度类别的样例。

类别进行区分。例如，类别 Acadian Flycatcher 的属性-视觉特征之间的关键语义知识为“*Bill Shape All Purpose*”、“*Head Pattern Plain*”等，如果不能将这些关键特征更突出的表示出来用于知识迁移，它将和其他相似的类别无法区分，造成错误分类，如图4.17所示。

为此，本章节进一步使用互语义蒸馏网络对 TransZero 进行升级，构成一个能够更有效挖掘属性-视觉特征之间的关键语义知识的零样本图像分类模型，命名为 TransZero++。TransZero++ 由两个基于属性指导的 Transformer 子网络构成，包括属性 → 视觉 Transformer 子网络（Attribute→Visual Transformer, AVT）和视觉 → 属性

Transformer 子网络 (Visual→Attribute Transformer, AVT)，它们基于跨注意力机制分别学习基于属性的视觉特征和基于视觉的属性特征。随后，TransZero++ 使用两个视觉-语义映射器 ( $\mathcal{M}_1$  和  $\mathcal{M}_2$ ) 将学习的两种特征进一步映射到类语义空间，通过与真实的类语义向量进行最近邻匹配实现零样本图像分类。为促使两个子网络对属性-视觉特征之间的关键语义知识进行有效挖掘，TransZero 进一步利用互语义蒸馏网络实现语义协作学习 (Semantical Collaborative Learning)。和 MSDN 不同的是，TransZero++ 不仅进行预测层面的语义蒸馏，同时也进行特征层面的语义蒸馏。在三个主流标准数据集（例如，CUB 数据集<sup>[4]</sup>、SUN 数据集<sup>[5]</sup> 和 AWA2 数据集<sup>[6]</sup>）上的实验结果表明：相较于 TransZero 和 MSDN，TransZero++ 可以进一步对属性-视觉特征之间的关键公共语义知识进行挖掘，实现更有效的零样本图像分类。

## 4.4.2 基于属性指导 Transformer 的互语义蒸馏网络

虽然 TransZero 能较为准确的挖掘视觉和属性特征之间的公共语义知识，但是不能充分挖掘所有潜在的语义知识。MSDN 通过双向注意力网络较为充分地挖掘视觉和属性特征之间的公共语义知识，但对部分关键语义知识的表示不够准确。为此，本章节进一步将 TransZero 和 MSDN 统一到模型（命名为 TransZero++），对视觉和属性特征的公共语义知识进行准确地、充分地挖掘，有效提高视觉-语义特征的一致性。TransZero++ 的模型结构如图4.18所示。TransZero++ 由属性 → 视觉 Transformer 子网络 (Attribute→Visual Transformer, AVT) 和视觉 → 属性 Transformer 子网络 (Visual→Attribute Transformer, AVT) 组成。AVT 利用属性 → 视觉的解码器学习基于属性的视觉特征，并使用视觉-语义映射器  $\mathcal{M}_1$  将基于属性的视觉特征映射到类语义空间。由于其结构和章节4.2的 TransZero 模型一致，下文将不对其做进一步介绍。VAT 使用视觉 → 属性的解码器学习基于视觉的属性特征，并使用视觉-语义映射器  $\mathcal{M}_2$  将基于视觉的属性特征映射到类语义空间。类似于 MSDN，TransZero++ 将两个子网络的视觉-语义嵌入表示进行融合，并通过与真实的类语义向量进行最近邻匹配实现零样本图像分类。和 TransZero 类似，TransZero++ 使用属性回归损失、基于属性的交叉熵损失和自校准损失对每个子网络进行模型优化。为促使两个子网络实现语义协作学习进行语义蒸馏，TransZero 进一步使用语义蒸馏损失进行模型优化。不同于 MSDN 只使用预测层面的语义蒸馏损失，TransZero++ 同时还使用特征层面的语义蒸馏损失。

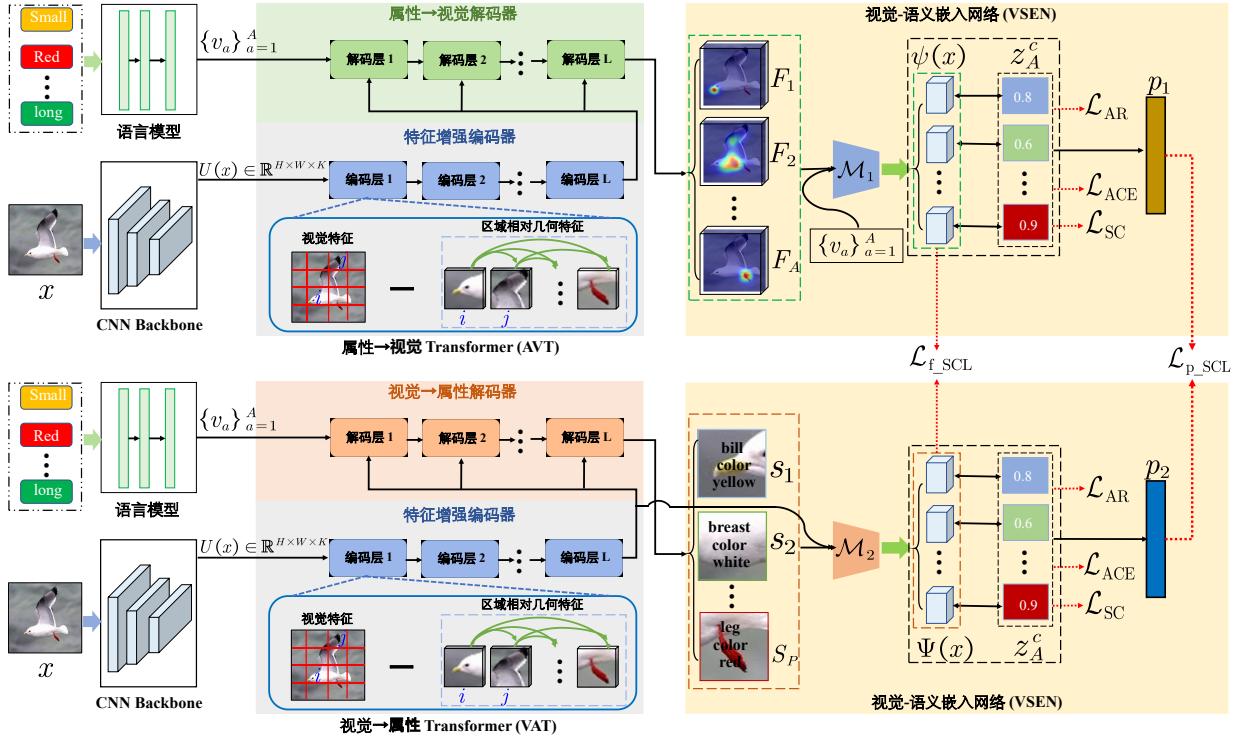


图 4.18 本章提出的 TransZero++ 模型结构示意图。

**视觉 → 属性 Transformer 子网络:** 视觉 → 属性 Transformer 子网络 (VAT) 由一个特征增强编码器和一个视觉 → 属性解码器构成。VAT 的特征增强编码器用于学习增强的视觉特征  $U^{v \rightarrow a}(x)$ , 模型结构和 TransZero 的特征增强编码器一致, 详细介绍请见章节4.2.2。VAT 学得  $U^{v \rightarrow a}(x)$  之后, 再使用一个视觉 → 属性解码器学习基于视觉的属性特征。具体而言, VAT 使用跨注意力机制关注视觉表示中的属性信息, 表示为:

$$Q_t^{v \rightarrow a} = U^{v \rightarrow a}(x)W_{qt}^{v \rightarrow a}, \quad (4.31)$$

$$K_t^{v \rightarrow a} = \mathcal{V}_A W_{kt}^{v \rightarrow a}, \quad (4.32)$$

$$V_t^{v \rightarrow a} = \mathcal{V}_A W_{vt}^{v \rightarrow a}, \quad (4.33)$$

$$\text{head}_t = \text{softmax} \left( \frac{Q_t^d K_t^{v \rightarrow a \top}}{\sqrt{\tau}} \right) V_t^{v \rightarrow a}, \quad (4.34)$$

$$\hat{S} = \|_{t=1}^T (\text{head}_t) W_o^{v \rightarrow a}, \quad (4.35)$$

其中  $W_{qt}^{v \rightarrow a}, W_{kt}^{v \rightarrow a}, W_{vt}^{v \rightarrow a}, W_o^{v \rightarrow a}$  均是可学习的权重, 而  $\|$  是串联操作。因此, VAT 学到一组基于视觉的属性特征  $\hat{S} = \{\hat{S}_1, \dots, \hat{S}_P\}$ 。本质上,  $\hat{S}$  是属性特征  $\mathcal{V}_A = \{v_a\}_{a=1}^A$

# 华 中 科 技 大 学 博 士 学 位 论 文

---

与图像中的  $P = H \times W$  个视觉区域相关的关键语义表示。随后, VAT 将具有两个线性变换的前馈神经网络 (FFN) 对基于视觉的属性特征  $\hat{S}$  做进一步的非线性变换:

$$S = ReLu \left( \hat{S}W_1^{v \rightarrow a} + b_1^{v \rightarrow a} \right) W_2^{v \rightarrow a} + b_2^{v \rightarrow a}, \quad (4.36)$$

其中,  $W_1^{v \rightarrow a}, W_2^{v \rightarrow a}, b_1^{v \rightarrow a}$  和  $b_2^{v \rightarrow a}$  分别是线性变换层的权重和偏差,  $S = \{S_1, \dots, S_P\}$  是 VAT 最终学习的基于视觉的属性特征。

随后, 类似于 MSDN 的视觉  $\rightarrow$  属性注意力子网络, 使用一个视觉-语义映射器将这些特征将嵌入到类语义空间得到相应的语义嵌入  $\Psi(x)$ 。

**模型优化:** 类似于 TransZero, TransZero++ 使用属性回归损失  $\mathcal{L}_{AR}$  (见公式4.14)、基于属性的交叉熵损失  $\mathcal{L}_{ACE}$  (见公式4.15) 和自校准损失  $\mathcal{L}_{SC}$  (见公式4.16) 对每个子网络进行模型优化:

$$\mathcal{L}_{AVT} = \mathcal{L}_{ACE}^{AVT} + \lambda_{AR}\mathcal{L}_{AR}^{AVT} + \lambda_{SC}\mathcal{L}_{SC}^{AVT}, \quad (4.37)$$

$$\mathcal{L}_{VAT} = \mathcal{L}_{ACE}^{VAT} + \lambda_{AR}\mathcal{L}_{AR}^{VAT} + \lambda_{SC}\mathcal{L}_{SC}^{VAT}, \quad (4.38)$$

其中  $\lambda_{AR}$  和  $\lambda_{SC}$  分别表示损失权重。

此外, TransZero++ 同时使用特征层面和预测层面的语义蒸馏损失指导两个子网络进行语义协作学习, 分别表示为:

$$\mathcal{L}_{f\_SCL} = \frac{1}{n_b} \sum_{i=1}^{n_b} \|\psi(x_i) - \Psi(x_i)\|_2^2, \quad (4.39)$$

$$\mathcal{L}_{p\_SCL} = \frac{1}{n_b} \sum_{i=1}^{n_b} \|p_1(x_i) - p_2(x_i)\|_2^2, \quad (4.40)$$

其中,  $p_1(x_i) = \{\psi(x_i) \times z_1^c, \dots, \psi(x_i) \times z_A^c\}$ ,  $p_2(x_i) = \{\Psi(x_i) \times z_1^c, \dots, \Psi(x_i) \times z_A^c\}$  为此, TransZero++ 的完整优化目标为:

$$\begin{aligned} \mathcal{L}_{total} = & \mathcal{L}_{AVT} + \lambda_{VAT}\mathcal{L}_{VAT} \\ & + \lambda_{f\_SCL}\mathcal{L}_{f\_SCL} + \lambda_{p\_SCL}\mathcal{L}_{p\_SCL}, \end{aligned} \quad (4.41)$$

其中,  $\lambda_{VAT}$ ,  $\lambda_{f\_SCL}$  和  $\lambda_{p\_SCL}$  分别表示为损失权重。

**零样本图像分类:** TransZero++ 和 MSDN 类似, 将两个子网络 (AVT 和 VAT) 的语义嵌入  $\psi(x_i)$  和  $\Psi(x_i)$  进行融合, 并与真实的类语义向量进行最近邻匹配, 实现零样

# 华 中 科 技 大 学 博 士 学 位 论 文

---

本图像分类：

$$c^* = \arg \max_{c \in \mathcal{C}^u / \mathcal{C}} (\alpha \psi(x_i) + (1 - \alpha) \Psi(x_i))^{\top} \times z^c + \mathbb{I}_{[c \in \mathcal{C}^u]}. \quad (4.42)$$

$\mathcal{C}^u / \mathcal{C}$  分别表示进行 CZSL 和 GZSL 设置下的零样本图像分类。

表 4.5 TransZero++ 和 TransZero (章节4.2)、MSDN (章节4.3) 以及其他先进的零样本图像分类方法在 CUB<sup>[4]</sup>、SUN<sup>[5]</sup>、AWA2<sup>[6]</sup> 数据集上的 CZSL/GZSL 实验结果对比。

当前先进的零样本图像分类方法	CUB 数据集				SUN 数据集				AWA2 数据集			
	CZSL		GZSL		CZSL		GZSL		CZSL		GZSL	
	acc	U	S	H	acc	U	S	H	acc	U	S	H
<b>端到端</b>												
QFSL <sup>[72]</sup>	58.8	33.3	48.1	39.4	56.2	30.9	18.5	23.1	63.5	52.1	72.8	60.7
LDF <sup>[73]</sup>	67.5	26.4	81.6	39.9	—	—	—	—	65.5	9.8	87.4	17.6
SGMA* <sup>[40]</sup>	71.0	36.7	71.3	48.5	—	—	—	—	68.8	37.6	87.1	52.5
AREN* <sup>[2]</sup>	71.8	38.9	78.7	52.1	60.6	19.0	38.8	25.5	67.9	15.6	92.9	26.7
LFGAA* <sup>[74]</sup>	67.6	36.2	<b>80.9</b>	50.0	61.5	18.5	<b>40.0</b>	25.3	68.1	27.0	<b>93.4</b>	41.9
<b>非端到端</b>												
<b>生成式方法</b>												
f-CLSWGAN <sup>[89]</sup>	57.3	43.7	57.7	49.7	60.8	42.6	36.6	39.4	68.2	57.9	61.4	59.6
cycle-CLSWGAN <sup>[93]</sup>	58.4	45.7	61.0	52.3	60.0	49.4	33.6	40.0	66.3	56.9	64.0	60.2
f-VAEGAN-D2 <sup>[61]</sup>	61.0	48.4	60.1	53.6	64.7	45.1	38.0	41.3	71.1	57.6	70.6	63.5
LsrGAN <sup>[102]</sup>	60.3	48.1	59.1	53.0	62.5	44.8	37.7	40.9	—	53.1	68.8	60.0
E-PGN <sup>[111]</sup>	72.4	52.0	61.1	56.2	—	—	—	—	73.4	52.6	83.5	64.6
Composer <sup>[132]</sup>	69.4	56.4	63.8	59.9	62.6	<b>55.1</b>	22.0	31.4	71.5	62.1	77.3	68.8
GCM-CF <sup>[123]</sup>	—	61.0	59.7	60.3	—	47.9	37.8	42.2	—	60.4	75.1	67.0
<b>嵌入式方法</b>												
SP-AEN <sup>[108]</sup>	55.4	34.7	70.6	46.6	59.2	24.9	38.6	30.3	58.5	23.3	90.9	37.1
PQZSL <sup>[124]</sup>	—	43.2	51.4	46.9	—	35.1	35.3	35.2	—	31.7	70.9	43.8
IIR <sup>[75]</sup>	63.8	30.4	65.8	41.2	63.5	22.0	34.1	26.7	67.9	17.6	87.0	28.9
TCN <sup>[76]</sup>	59.5	52.6	52.0	52.3	61.5	31.2	37.3	34.0	71.2	61.2	65.8	63.4
DVBE <sup>[107]</sup>	—	53.2	60.2	56.5	—	45.0	37.2	40.7	—	63.6	70.8	67.0
DAZLE* <sup>[81]</sup>	66.0	56.7	59.6	58.1	59.4	52.3	24.3	33.2	67.9	60.3	75.7	67.1
APN* <sup>[41]</sup>	72.0	65.3	69.3	67.2	61.6	41.9	34.0	37.6	68.4	57.1	72.4	63.9
<b>MSDN (本文方法)</b>	76.1	68.7	67.5	68.1	65.8	52.2	34.2	41.3	70.1	62.0	74.5	67.7
<b>TransZero (本文方法)</b>	76.8	<b>69.3</b>	68.3	68.8	65.6	<b>52.6</b>	33.4	40.8	70.1	61.3	82.3	70.2
<b>TransZero++ (本文方法)</b>	<b>78.3</b>	67.5	73.6	<b>70.4</b>	<b>67.6</b>	48.6	37.8	<b>42.5</b>	<b>72.6</b>	<b>64.6</b>	82.7	<b>72.5</b>

注：符号“—”表示相应结果缺失，符号“\*”表示基于注意力机制的方法。

### 4.4.3 实验结果与分析

类似于 TransZero，本章节在三个主流的零样本图像分类标准数据集上同时进行 CZSL 和 GZSL 实验验证 TransZero++ 的有效性，包括两个细粒度数据集 (CUB<sup>[4]</sup>, SUN<sup>[5]</sup>) 和一个粗粒度数据集 (AWA2<sup>[6]</sup>)。由于 TransZero++ 是 TransZero 和 MSDN

# 华中科技大学博士学位论文

---

的融合模型，为避免内容累赘，本章节主要展示 TransZero++ 与 MSDN、TransZero 以及当前先进的零样本图像方法方法的对比结果。

实验结果如表4.5所示。在 CZSL 设置下，TransZero++ 相较于 MSDN/TransZero 在 CUB、SUN、AWA2 数据集上分别取得了 2.2%/1.8%、1.8%/2.0%、2.5%/2.5% 的性能提升；在 GZSL 设置下，TransZero++ 相较于 MSDN/TransZero 在 CUB、SUN、AWA2 数据集上分别取得了 2.3%/1.6%、1.2%/1.7%、4.8%/2.3% 的调和均值  $\mathbf{H}$  的提升。这表明 TransZero++ 可以进一步准确且充分地挖掘属性-视觉特征之间的关键公共语义特征表示。如图4.17展示了三个方法学习的特征可视化，TransZero++ 比 MSDN 和 TransZero 能以更高的置信度表示视觉中的关键属性，直观地验证了 TransZero++ 的有效性。相较于当前先进的零样本图像分类方法，TransZero++ 在 CUB、SUN、AWA2 数据集上均取得领先的性能，进一步表明基于属性指导的 Transformer 和互语义蒸馏网络在零样本图像分类领域具有很大的潜力和优势。

## 4.5 本章小结

本章针对视觉-语义特征的表示不一致性问题，提出了基于属性-视觉关键公共语义知识的零样本图像分类方法。首先，TrasZero 通过使用视觉特征增强编码器消除区域几何先验以提高视觉特征的迁移性，并使用基于跨注意力机制的属性 → 视觉解码器进行视觉特征的属性定位，从而较为准确地学习属性-视觉特征之间重要的公共语义知识。随后，MSDN 基于属性 → 视觉的注意力子网络和视觉 → 属性注意力子网络分别学习基于属性的视觉特征和基于视觉的属性特征对属性-视觉之间的潜在语义知识进行表示，并利用语义蒸馏损失促使两个子网络进行协作学习和相互指导，充分地学习属性-视觉之间关键的语义知识。最后，本章将 TransZero 和 MSDN 融合到一个统一的模型，准确且充分地挖掘视觉-属性特征之间的公共语义知识，提高视觉和语义特征之间的语义一致性。通过大量的实验证明了所提出方法的有效性，并均取得当前零样本图像分类的领先性能。

## 5 基于层次语义-视觉适应的零样本图像分类

### 5.1 引言

公共子空间学习是实现零样本图像分类进行视觉-语义交互的一种经典方法，它将视觉域和语义域特征同时映射到公共子空间，从而使用最近邻匹配或者有监督分类器实现零样本图像分类。然而，现有公共子空间式零样本图像分类方法只通过单步适应（one-step adaptation）对视觉和语义域的特征分布进行对准<sup>[43–46]</sup>（例如，CADA-VAE<sup>[3]</sup>），忽略了视觉-语义特征的异构性同时存在特征分布差异和特征流形结构差异<sup>[47,48]</sup>，使得视觉和语义特征映射到不同的子流形空间上，未能实现视觉和语义特征在子空间中真正对准，如图5.1(a) 和5.1(c) 图所示。当采用欧几里德距离或流形距离<sup>[49,50]</sup>对不同类别之间的关系进行度量时，分类器不可避免地对一些样本进行错误分类，导致零样本图像分类性能根本上受到限制。

针对此问题，本章节提出一种基于层次语义-视觉适应的零样本图像分类方法（Hierarchical Semantic-Visual Adaptation for Zero-Shot Image Classification, HSVA）。不同于现有方法使用两个不同的编码器并采用分布对准约束只进行特征分布对准，HSVA 同时进行结构对准和分布对准以学习真正的公共子空间实现视觉特征和语义特征的对准。HSVA 将结构适应和分布适应集成在两个局部共享的变分自编码器里面。在结构适应模块中，HSVA 使用有监督的对抗差异性学习机制促使视觉和语义特征流形相互靠近，实现两种异构特征的流形结构对准。在分布适应模块中，HSVA 使用一个公共编码器将结构对准的视觉和语义特征映射到分布对准的公共子空间，该编码器通过最小化分布对准子空间中的视觉和语义特征之间多元高斯分布的 Wasserstein 距离<sup>[138]</sup> 实现分布对准。最终，HSVA 实现视觉和语义特征的真正对准，有效地提高零样本图像分类精度，并为公共子空间式零样本图像分类提供新的研究思路。

### 5.2 层次语义-视觉适应网络

首先指出，本章节使用的基本符号定义见章节3.2.2。HSVA 的模型结构如图5.2所示，其包含结构适应模块（Structure Adaptation, SA）和分布适应模块（Distribution

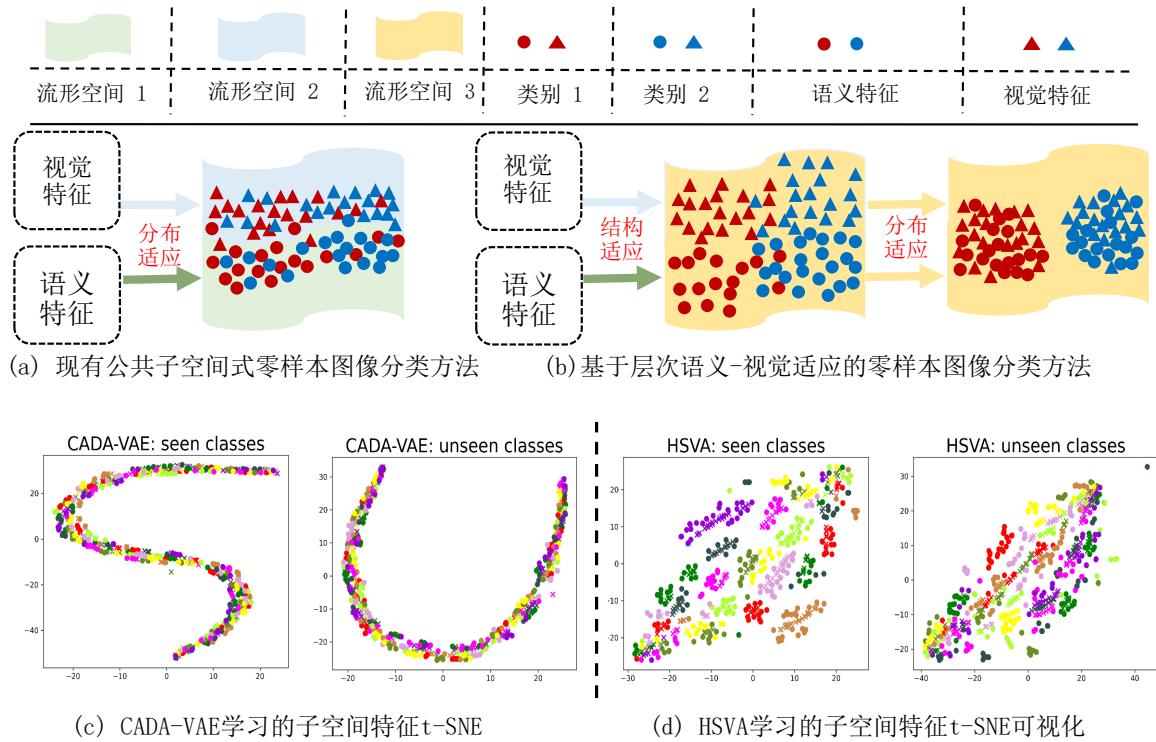


图 5.1 基于单步适应的和基于层次语义-视觉适应的公共子空间学习方法。

Adaptation, DA)，并且这两个模块被统一在两个局部共享的变分自编码器中 (VAE( $E^x, E^o, D^x$ ) 和 VAE( $E^z, E^o, D^z$ ))。DA 使用两个独立的特征编码器 ( $E^x, E^z$ ) 分别将视觉特征和语义特征映射到一个结构对准的子空间，并通过有监督的对抗差异性学习机制实现两种异构特征的流形结构对准。SA 使用一个共享的特征编码器  $E^o$  进一步将结构对准的视觉和语义特征映射到分布对准子空间，并通过约束分布对准子空间中视觉和语义特征之间多元高斯分布的 Wasserstein 距离实现两种特征的分布对准。

**局部共享的变分自编码器:** HSVA 的整体模型结构如图5.2所示。宏观上看，HSVVA 包含两个局部共享的变分自动编码器组成，分别为 VAE( $E^x, E^o, D^x$ ) 和 VAE( $E^z, E^o, D^z$ )。VAE( $E^x, E^o, D^x$ ) 包括两个编码器（分别为视觉特征编码器  $E^x$  和共享特征编码器  $E^o$ ）和一个解码器（视觉特征解码器  $D^x$ ）， $E^x$  和  $E^o$  分别学习结构对准和分布对准的视觉嵌入特征，它们与语义嵌入特征共享两个子空间（即结构对准子空间和分布对准子空间）， $D^x$  旨在将分布对准空间中的视觉和语义特征重构成语相应输入视觉特征一致的特征表示。类似地 VAE( $E^z, E^o, D^z$ ) 包含两个特征编码器（分别为语义特征编

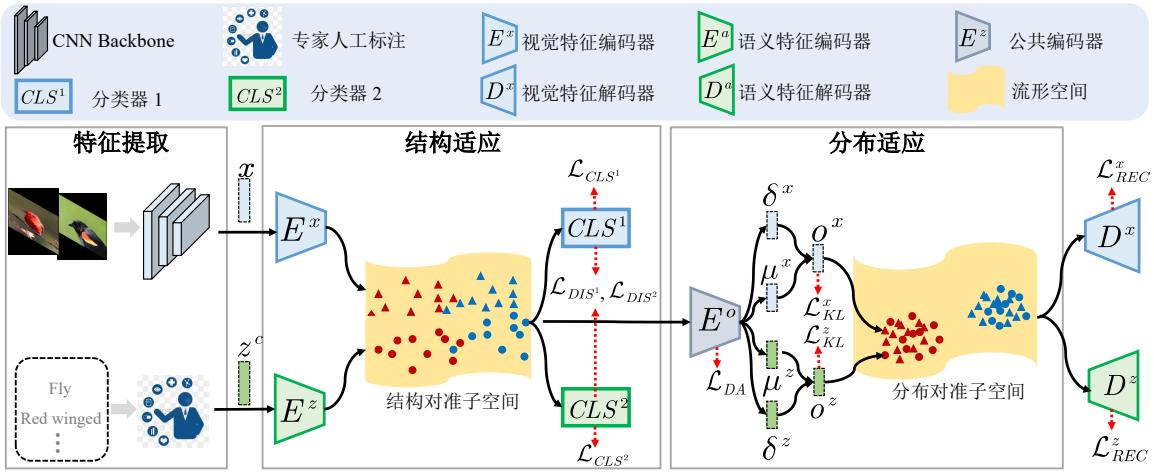


图 5.2 本章提出的 HSVA 模型结构示意图。

码器  $E^z$  和共享特征编码器  $E^o$ ）和一个语义特征解码器， $E^x$  和  $E^o$  分别学习结构对准和分布对准的语义嵌入特征， $D^z$  将分布对准空间中的视觉和语义特征重构成与相应输入语义特征一致的特征表示。这两个变分自编码器先使用自编码器损失进行优化：

$$\mathcal{L}_{VAE}^x(E^x, E^o) = \mathbb{E}[\log D^x(o^x)] - \gamma \text{KL}(E^o(E^x(x))) \| p(o^x|x), \quad (5.1)$$

$$\mathcal{L}_{VAE}^z(E^z, E^o) = \mathbb{E}[\log D^z(o^z)] - \gamma \text{KL}(E^o(E^z(z))) \| p(o^z|z), \quad (5.2)$$

其中， $o^x$  和  $o^z$  是分布对齐空间中的视觉嵌入特征和语义嵌入特征， $\gamma$  是 KL 散度的权重， $p(o^x|x)$  和  $p(o^z|z)$  均假定为  $\mathcal{N}(0, 1)$  的先验分布。为使得结构和分布对准空间中的视觉嵌入特征和语义嵌入特征更加一致，HSVA 进一步使用交叉重建损失来约束两个局部对准的变分自动编码器：

$$\mathcal{L}_{REC}(E^x, E^z, E^o, D^x, D^z) = \mathcal{L}_{REC}^x(E^x, E^o, D^x) + \mathcal{L}_{REC}^z(E^z, E^o, D^z), \quad (5.3)$$

其中，

$$\mathcal{L}_{REC}^x(E^x, E^o, D^x) = \|x - D^x(o^z)\|_1, \quad (5.4)$$

$$\mathcal{L}_{REC}^z(E^z, E^o, D^z) = \|z - D^z(o^x)\|_1. \quad (5.5)$$

随后，两个局部对准的变分自动编码器对视觉和语义特征进行结构适应和分布适应学习。

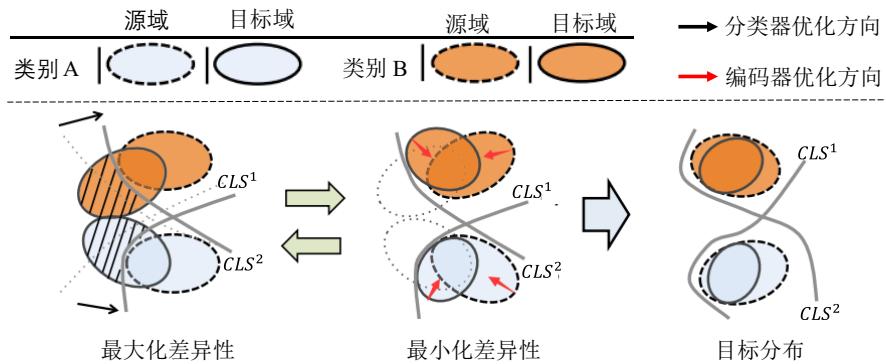


图 5.3 有监督的对抗差异性学习机制。

**结构适应:** 结构适应 (SA) 用于指导编码器  $E^x$  和  $E^z$  学习视觉和语义特征的结构对准子空间。现有公共子空间式零样本图像分类方法<sup>[3,45]</sup> 直接将视觉和语义特征使用分布距离约束 (例如, MMD) 进行特征分布对准, 这忽略了两个域的异构特征之间的流形结构差异。受基于特定任务的无监督领域自适应方法<sup>[139,140]</sup> 的启发, 本章节提出了一种有监督的对抗差异性学习机制 (Supervised Aversarial Dscrepancy Learning, SAD) 促使  $E^x$  和  $E^z$  学习的视觉和语义嵌入特征流形相互靠近, 实现两种异构特征的流形结构对准, 并学习具有类判别性的特征表示。其中差异性由两个特定任务分类器 (分类器  $CLS^1, CLS^2$ ) 输出进行度量。为此, SAD 类似于领域自适应的过程, 视觉特征和语义特征互为源域和目标域, 通过域适应实现两个域的特征对准。由于目标域样本 (例如, 语义嵌入特征  $E^z(z)$ ) 没有标签, 基于特定任务的无监督领域自适应方法<sup>[139,140]</sup> 只能通过检测目标域特征 (例如, 语义嵌入特征  $E^z(z)$ ) 是否远离于源域, 并通过最小化差异性将目标域推向于源域, 从而学习一个编码器 (生成器) 生成与源域靠近的嵌入特征。相反, SAD 同时检测源域样本特征 (例如视觉嵌入  $E^x(x)$  和目标样本 (例如语义嵌入  $E^z(z)$ ), 学习具有两个特定任务的编码器将视觉和语义特征实现结构对准, 优化过程如图5.3所示。具体而言, SAD 包括三个优化步骤:

- 两个特定任务分类器的分类学习, 对  $E^x, E^z, CLS^1$  和  $CLS^2$  进行优化;
- 最大化差异性性学习, 对  $CLS^1$  和  $CLS^2$  进行优化;
- 最小化差异性性学习, 对  $E^x$  和  $E^z$  进行优化。

**两个特定任务分类器的分类学习:** SAD 通过最小化 softmax 交叉熵损失优化  $E^x$ 、 $E^z$ 、 $CLS^1$  和  $CLS^2$ , 从而对结构对准子空间中的语义和视觉嵌入特征进行分类。这对于分类器和编码器 ( $E^x, E^z$ ) 获得判别性嵌入特征非常重要。视觉和语义嵌入特

征分类损失表示为：

$$\begin{aligned}\mathcal{L}_{CLS}(E^x, E^z, CLS^1, CLS^2) = & \mathcal{L}_{CLS^1}(E^x(x), y) + \mathcal{L}_{CLS^2}(E^x(x), y) \\ & + \mathcal{L}_{CLS^1}(E^z(z), y) + \mathcal{L}_{CLS^2}(E^z(z), y),\end{aligned}\quad (5.6)$$

其中，

$$\mathcal{L}_{CLS^1}(x', y') = \mathcal{L}_{CLS^2}(x', y') = -\mathbb{E} \sum_{k=1}^{C^s} \mathbb{I}_{[k=y']} \log p(y | x'). \quad (5.7)$$

$\mathbb{I}_{[k=y']}$  是一个指示函数（当  $k = y'$ , 取值为 1, 否则为 0),  $p(y|x')$  是分类器对  $x'$  的预测值概率。SAD 同时对视觉和语义嵌入特征进行分类, 实现对分类器的初始化, 用于后续的差异性学习。

**最大化差异性优化分类器  $CLS^1$  和  $CLS^2$ :** 在这一步, HSVA 固定特征编码器 ( $E^x, E^a$ ) 的参数, 通过最大化两个分类器在视觉和语义嵌入特征上的输出差异优化分类器 ( $CLS^1, CLS^2$ ), 使得在学习过程中两个分类器始终具有差异性表示能力。因此, 分类器识别出分类边界之外的源特征样本和目标特征样本, 学习更准确的分类器决策边界。由于切片 Wasserstein 差异性<sup>[140]</sup> (Sliced Wasserstein Discrepancy, SWD) 为刻画不同目标的差异性提供了几何意义指导, SAD 使用它来度量视觉嵌入特征在两个分类器预测之间的差异 ( $p_1(y|E^x(x)), p_2(y|E^x(x))$ ) 和语义嵌入特征在两个分类器之间的差异 ( $p_1(y|E^z(z)), p_2(y|E^z(z))$ ):

$$\mathcal{L}_{DIS^1}(CLS^1, CLS^2) = -\mathcal{L}_{SWD}(E^x(x)) - \mathcal{L}_{SWD}(E^z(z)), \quad (5.8)$$

$$\mathcal{L}_{SWD}(\mathbf{x}) = \sum_m (\mathcal{R}_{\theta_m} p_1(y|\mathbf{x}), \mathcal{R}_{\theta_m} p_2(y|\mathbf{x})), \quad (5.9)$$

其中  $\mathcal{R}_{\theta_m}$  是第  $m$  次一维线性投影运算, 而  $\theta$  是  $\mathbb{R}^d$  中的单位球面  $S^{(d-1)}$  的统一度量。切片 Wasserstein 差异性  $\mathcal{L}_{SWD}$  的详细介绍, 请见 Lee 等人的工作<sup>[140]</sup>。为此,  $\mathcal{L}_{DIS}$  越大, 差异性就越大。

**最小化差异性优化编码器  $E^x$  和  $E^z$ :** 为鼓励视觉嵌入特征 ( $E^x(x)$ ) 和语义嵌入特征 ( $E^z(z)$ ) 在结构对准子空间空间中对准, HSVA 固定两个分类器的参数并优化两个编码器 ( $E^x, E^z$ ), 以最大程度地减少两个分类器在  $E^x(x^s)$  和  $E^z(z^s)$  上的输出差异:

$$\mathcal{L}_{DIS^2}(E^x, E^z) = \mathcal{L}_{SWD}(E^x(x)) + \mathcal{L}_{SWD}(E^z(z)). \quad (5.10)$$

# 华 中 科 技 大 学 博 士 学 位 论 文

---

这将促使视觉特征编码器和语义特征编码器学习的视觉和语义特征嵌入特征的流形结构相互靠近，实现结构对准。

**分布适应：** 经过结构适应优化，HSVA 可以学习一个结构对准的公共空间，消除视觉嵌入特征和语义嵌入特征之间的流形结构差异。然而，结构对准的子空间仍然受到特征分布变化的影响。为了进一步学习视觉和语义特征的分布对准子空间，HSVA 进一步使用分布适应对模型做进一步优化。现有的公共空间式零样本图像分类方法<sup>[3,43–46]</sup> 使用两种不同的编码器（映射函数）将视觉和语义特征进行分布对准，以学习分布对准的嵌入特征。相比之下，HSVA 使用一个共享编码器  $E^o$  将结构对准子空间中的视觉和语义嵌入特征映射到分布对准子空间实现特征分布对准，从而保留两种异构特征在分布对准子空间中的结构一致性。HSVA 首先在分布适应模块通过最小化语义-视觉特征之间多元高斯分布的 Wasserstein 距离进行优化：

$$\mathcal{L}_{DA}(E^x, E^z, E^o) = \left( \|\mu^x - \mu^z\|_2^2 + \left\| (\delta^x)^{\frac{1}{2}} - (\delta^z)^{\frac{1}{2}} \right\|_F^2 \right)^{\frac{1}{2}}, \quad (5.11)$$

其中， $\|\cdot\|_F^2$  是 Frobenius 范数。

因为 HSVA 在利用分布对准子空间中的对准特征学习一个有监督的分类器实现零样本图像分类，所以 HSVA 也是一种生成式模型。为此，HSVA 也会面临已知类偏差问题<sup>[106]</sup>，即利用未知类的语义特征生成分布对准子空间中的未知类样本 ( $E^o(E^z(z^u))$ ) 容易和分布对准子空间中的已知类样本 ( $E^o(E^o(x^s))$ ) 混淆。HSVA 使用反相互关系对准<sup>[141]</sup> (Inverse CORrelation ALignment, iCORAL) 使  $E^o(E^z(z^u))$  和  $E^o(E^o(x^s))$  尽可能的分开，从而缓解已知类偏差问题，公式表示为：

$$\mathcal{L}_{iCORAL}(E^x, E^z, E^o) = -CORAL(E^o(E^x(x^s)), E^o(E^z(z^u))), \quad (5.12)$$

CORAL 的度量细节详见论文<sup>[141]</sup>。

**模型的优化：** HSVA 通过对两个局部共享的变分自编码器、结构适应模块、分布适应模块同时进行优化，总体优化目标为：

$$\begin{aligned} \mathcal{L}_{HSVA}(E^x, E^z, CLS^1, CLS^2, E^o, D^x, D^z) &= \mathcal{L}_{VAE}^x + \mathcal{L}_{VAE}^z + \lambda_1 \mathcal{L}_{REC} + \mathcal{L}_{CLS} \\ &\quad + \lambda_2 (\mathcal{L}_{DIS^1} + \mathcal{L}_{DIS^2}) + \lambda_3 (\mathcal{L}_{iCORAL} + \mathcal{L}_{DA}), \end{aligned} \quad (5.13)$$

其中， $\lambda_1, \lambda_2, \lambda_3$  分别控制相应损失项的权重。类似于 GANs<sup>[84]</sup> 模型的训练，HSVA 在结构适应模块中交替地优化  $E^x, E^z, CLS^1, CLS^2$ 。虽然 HSVA 总体优化目标由

三个组成部分，但 HSVA 在多个数据集上使用同样的权重超参设置均能取得优异的性能，表明 HSVA 是容易优化的。

**零样本图像分类：**HSVA 优化完成后，视觉和语义特征将被编码为分布一对准子空间中的嵌入特征表示，同时具有结构对准和分布对准特性。使用重参数化<sup>[83]</sup>，HSVA 使用  $E^x$  和  $E^o$  将训练集/测试集中真实的已知类样本 ( $x_{tr}^s$  和  $x_{te}^s$ ) 和测试集中真实的未知类样本 ( $x_{te}^u$ ) 的视觉特征编码为结构和分布对对准的嵌入特征表示，即  $o^{x_{tr}^s} = E^o(E^x(x_{tr}^s))$ 、 $o^{x_{te}^s} = E^o(E^x(x_{te}^s))$  和  $o^{x_{te}^u} = E^o(E^x(x_{te}^u))$ 。类似地，HSVA 使用  $E^a$  和  $E^o$  根据未知类的语义特征  $z^u$  生成未知类伪特征样本  $o^{z^{syn}} = E^o(E^z(z^u))$ 。随后，HSVA 使用  $o^{x_{tr}^s}$  和  $o^{z^{syn}}$  训练一个有监督分类器（例如，softmax），并利用训练好的分类器对测试样本的嵌入特征  $o^{x_{te}^s}$  和  $o^{x_{te}^u}$  进行测试。

## 5.3 实验设置

为验 HSVA 的有效性，本文在四个零样本图像分类数据集上进行实验，包括两个细粒度数据集（CUB<sup>[4]</sup>，SUN<sup>[5]</sup>）和两个粗粒度数据集（AWA1<sup>[9]</sup>，AWA2<sup>[6]</sup>）。本文依据 Xian 等人<sup>[6]</sup>最新的数据集划分方式进行模型训练和测试，数据集的具体介绍见章节2.3.1。在 HSVA 中，编码器 ( $E^x$ ,  $E^z$ ,  $E^o$ )、解码器 ( $D^x$ ,  $D^o$ ) 和分类器 ( $CLS^1$ ,  $CLS^2$ ) 均是多层感知机结构，它们的具体结构设置如表5.1所示。本文使用 Adam 优化器<sup>[130]</sup> ( $\beta_1=0.5$ ,  $\beta_2=0.999$ ) 对模型进行优化。类似于现有先进的公共子空间式零样本图像分类方法 CADA-VAE<sup>[3]</sup>，本章节使用模拟退火方法<sup>[142]</sup> (Annealing Scheme) 对所有数据集的超参数权重  $\gamma$ ,  $\lambda_1$ ,  $\lambda_2$ ,  $\lambda_3$  进行更新。具体而言， $\gamma$  以 0.0026/每步的速度从第 0-90 步不断增加， $\lambda_1$  从以 0.044/每步的速度第 21 步增加到 75， $\lambda_2$  和  $\lambda_3$  以 0.54/每步的速度从 0-22 步持续增加。在 CZSL 设置情况下，分别为 AWA1、CUB 和 SUN 数据集中的每个未知类生成 800、400 和 200 个伪特征样本用于训练分类器。在 GZSL 设置中，为所有数据集上的每个已知类和未知类分别生成 200 和 400 个伪特征样本用于训练分类器。在 CUB、AWA1、AWA2 数据集上，HSVA 的结构对准子空间和分布对准子空间的维度分别设置为 2048 和 64。

本章节通过以下实验验证 HSVA 的有效性：

- 超参实验分析；
- 消融实验分析；

表 5.1 HSVA 中的网络结构设置细节。

视觉特征编码器 ( $E^x$ )	输入层: $x$ , size=2048; 隐藏层: Fully connected, neurons=4096; LeakyReLU; 输出层: Fully connected, neurons=2048;
语义特征编码器 ( $E^z$ )	输入层: $x$ , size= $A$ ; 隐藏层: Fully connected, neurons=4096; LeakyReLU; 输出层: Fully connected, neurons=2048;
分类器 ( $CLS^1/CLS^2$ )	输入层: $E^x(x)$ or $E^z(z)$ , size =2048; 隐藏层: Fully connected, neurons=512; BachNorm;LeakyReLU; 输出层: Fully connected, neurons= $C^s$ ;
共享特征编码器 ( $E^o$ )	输入层: $E^x(x)$ or $E^z(z)$ , size=2048; 隐藏层: Fully connected, neurons=2048; LeakyReLU; 隐藏层: Fully connected, neurons=Dim_D*2; LeakyReLU; 编码层: $\mu^x$ =Dim_D and $\delta^x$ =Dim_D, or $\mu^z$ =Dim_D and $\delta^z$ =Dim_D; 输出层: Reparametrization, $o^x$ or $o^z$ , neurons=Dim_D;
视觉特征解码器 ( $D^x$ )	输入层: $o^x$ , size=Dim_D; 隐藏层: Fully connected, neurons=4096; LeakyReLU; 输出层: Fully connected, neurons=2048;
语义特征解码器 ( $D^z$ )	输入层: $o^z$ , size=Dim_D; 隐藏层: Fully connected, neurons=4096; LeakyReLU; 输出层: Fully connected, neurons= $A$ ;

- 定性实验分析;
- HSVA 和其他经典且先进的零样本图像分类方法在 CZSL/GZSL 不同设置上的实验结果对比。

接下来将对不同的实验进行详细阐述和分析。本章节方法 ViFR 的源代码地址:  
<https://github.com/shiming-chen/HSVA>。

## 5.4 超参实验分析

本章节主要分析每个已知类和未知类生成的伪特征样本个数 ( $N_s$  和  $N_u$ )、结构对准子空间和分布对准子空间的特征维度 (Dim\_S 和 Dim\_D) 对 HSVA 的影响。

CZSL 设置下每个未知类生成的伪特征样本个数  $N_u$ : 本章节评估了 CZSL 设置下每个未知类生成的伪特征样本个数  $N_u$  对 HSVA 的影响。本章节在 CUB、SUN 和 AWA1 数据集将  $N_u$  取值为 {200, 400, 800, 1200, 1600, 2000} 进行评估, 如图5.4所示。

总体而言，HSVA 的性能对 CZSL 设置下的  $N_u$  不敏感。根据验证试验结果，HSVA 在 CUB、SUN 和 AWA1 数据集上把  $N_u$  分别设置为 400、200 和 800。

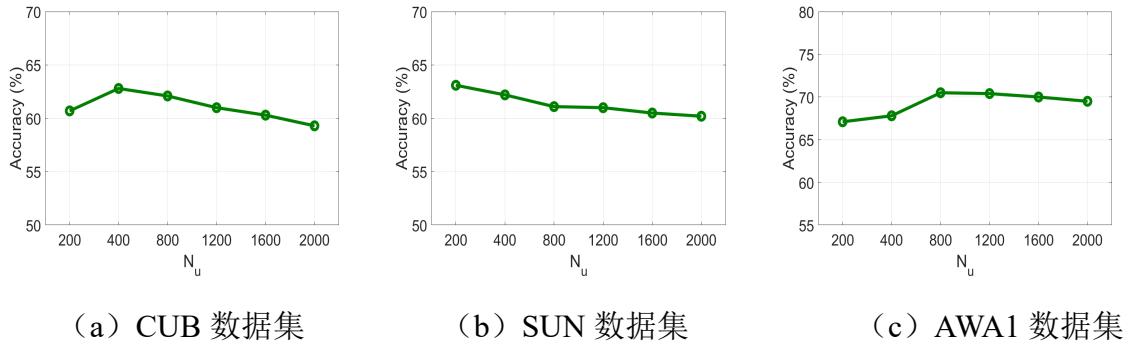


图 5.4 CZSL 设置下每个未知类生成的伪特征样本个数  $N_u$  对 HSVA 的影响。

GZSL 设置下每个已知类和未知类生成的伪特征样本个数 ( $N_s$  和  $N_u$ )：本章节分析了 GZSL 设置下每个已知类和未知类生成的伪特征样本个数 ( $N_s$  和  $N_u$ ) 对 HSVA 的影响。本章节在 CUB 和 AWA1 数据集上对  $N_s$  和  $N_u$  进行一定范围的取值 (即  $N_s = \{100, 200, 400\}$  和  $N_u = \{100, 200, 400, 800, 1200\}$ )，共有 15 组 ( $N_s, N_u$ ) 对 HSVA 进行评估，如图5.5所示。由于视觉特征的判别性信息跟个丰富，所以应该将使用视觉特征生成的每个已知类伪特征样本数量  $N_s$  小于使用语义特征生成的每个未知类伪样本数量  $N_u$ ，实现分类器的平衡学习。相较于 HSVA( $N_s/N_u = 1/1$ )，HSVA( $N_s/N_u = 1/2$ ) 取得了更高的分类精度，在 CUB 数据集和 SUN 数据集上的性能指标  $\mathbf{U}/\mathbf{H}$  至少提高了 17.5%/9.5% 和 36.3%/30.5%。然而，当  $N_s/N_u > 1/2$  时，HSVA 在在已知类上取得更好的识别性能。为更好的平衡 HSVA 在已知类和未知类之间的识别性能，本章节将 ( $N_s, N_u$ ) 在所有数据集上设置为 (200, 400)。

结构对准子空间和分布对准子空间的特征维度 (Dim\_S 和 Dim\_D)：本章节分析了结构对准子空间和分布对准子空间的特征维度 (Dim\_S 和 Dim\_D) 对 HSVA 的影响。如图5.6和图5.7所示，HSVA 在粗粒度数据集 (如 AWA1<sup>[9]</sup>) 受 Dim\_S 和 Dim\_D 影响较小，而在细粒度数据集如 (SUN<sup>[5]</sup>) 较为敏感。这是因为 HSVA 在子空间维度更高的情况下，其在 SUN 这种小数据集上更容易过拟合。为此，在 SUN 数据集上，当增加 Dim\_S 和 Dim\_D 时，HSVA 的已知类识别精度上升，未知类识别精度下降。当 (Dim\_S, Dim\_D) 设置为 (2048, 64)，HSVA 取得了较好的性能表现。因此 HSVA 在所有数据集上均将 (Dim\_S, Dim\_D) 设置为 (2048, 64)。

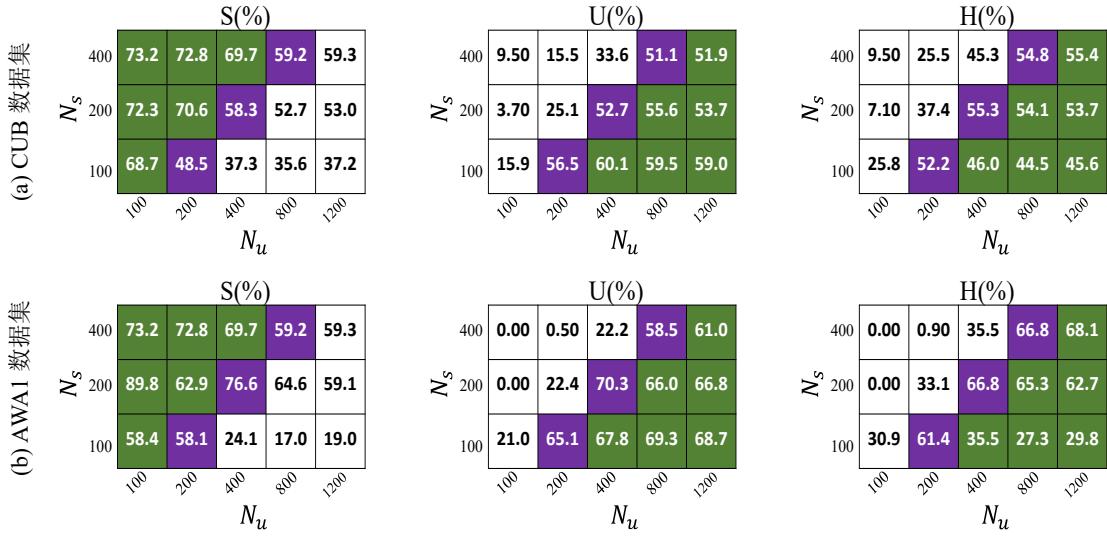


图 5.5 GZSL 设置下每个已知类和未知类生成的伪特征样本个数 ( $N_s$  和  $N_u$ ) 对 HSVA 的影响。

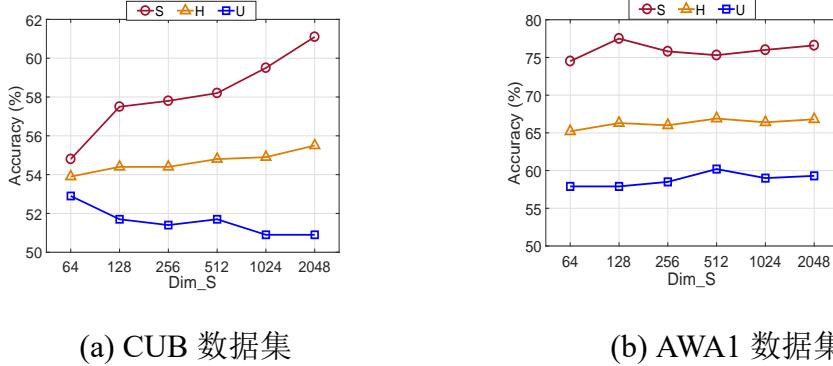


图 5.6 结构对准子空间的特征维度 Dim\_S 对 HSVA 的影响。

## 5.5 消融实验分析

为了进一步深入验证 HSVA 的有效性，本章节进行了消融实验分析，评估 HSVA 的结构适应学习（表示为  $\mathcal{L}_{SA}$ ）、分布适应学习（即  $\mathcal{L}_{DA} + \mathcal{L}_{iCORAL}$ ）以及反相互关系对准损失（ $\mathcal{L}_{iCORAL}$ ）。实验结果如表5.2所示。当结构适应模块中不适用 SAD 进行结构适应学习，HSVA 的性能比其完整模型性能显著下降，例如在 AWA1 和 CUB 数据集上的性能指标  $\mathbf{H}$  分别下降 4.5% 和 2.1%。如果在分布适应模块中不进行分布适应学习，HSVA 相较于其完整模型在细粒度数据集（例如，CUB 和 SUN）粗粒度数据集（AWA1 和 AWA2）的性能指标  $\mathbf{H}$  分别至少下降 23.4% 和 4.7%。这些结果

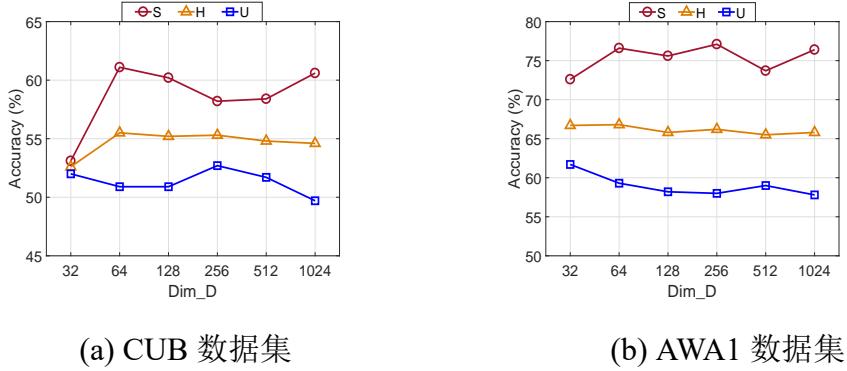


图 5.7 分布对准子空间的特征维度  $\text{Dim}_D$  对 HSVA 的影响。

表 5.2 在不同模型成分设置下，HSVA 在 AWA1<sup>[9]</sup>、AWA2<sup>[6]</sup>、CUB<sup>[4]</sup> 和 SUN<sup>[5]</sup> 数据集上的实验结果。

不同设置下的 HSVA	AWA1			AWA2			CUB			SUN		
	U	S	H	U	S	H	U	S	H	U	S	H
HSVA 无 $\mathcal{L}_{SA}$	55.2	71.5	62.3	53.9	77.2	63.5	49.3	57.9	53.2	49.7	37.2	42.6
HSVA 无 $\mathcal{L}_{DA}$ 和 $\mathcal{L}_{iCORAL}$	31.3	70.6	43.4	26.3	78.4	39.3	41.9	56.1	47.9	42.2	35.5	38.6
HSVA 无 $\mathcal{L}_{iCORAL}$	54.3	73.5	62.4	51.3	82.6	63.3	51.1	57.0	53.9	45.1	38.9	41.8
HSVA	59.3	76.6	66.8	56.7	79.8	66.3	52.7	58.3	55.3	48.6	39.0	43.3

表明，分布对准对齐是学习视觉-语义公共子空间的基本过程，而结构对准是学习视觉-语义公共子空间的必要操作。这通常被基于单步适应<sup>[3,43–46,105]</sup>的分布对齐方法所忽略。此外， $\mathcal{L}_{iCORAL}$  生成的未知类样本与已知类样本远离，有效解决已知类偏差问题<sup>[106]</sup>。HSVA 使用  $\mathcal{L}_{iCORAL}$  进行分布适应优化时，在 AWA1、AWA2、CUB 和 SUN 上的性能指标  $\mathbf{H}$  分别提升了 4.4%、3.0%、1.4% 和 1.5%。

## 5.6 定性实验分析

为直观地验证 HSVA 的有效性，本章节通过将 HSVA 和现有先进的公共子空间式零样本图像分类方法 CADA-VAE<sup>[3]</sup> 在 CUB<sup>[4]</sup> 和 AWA1<sup>[9]</sup> 数据集上学习的 10 个类别在分布对准的子空间中视觉嵌入特征和语义嵌入特征进行 t-SNE 可视化<sup>[1]</sup>。图中的“。”表示视觉嵌入特征，图中的“×”语义嵌入特征。

如图5.8所示，CADA-VAE 将语义特征和视觉特征映射到不同的子流形中，使得在二维 t-SNE 可视化空间中不同类别的视觉特征和语义特征相互混淆，导致一些样

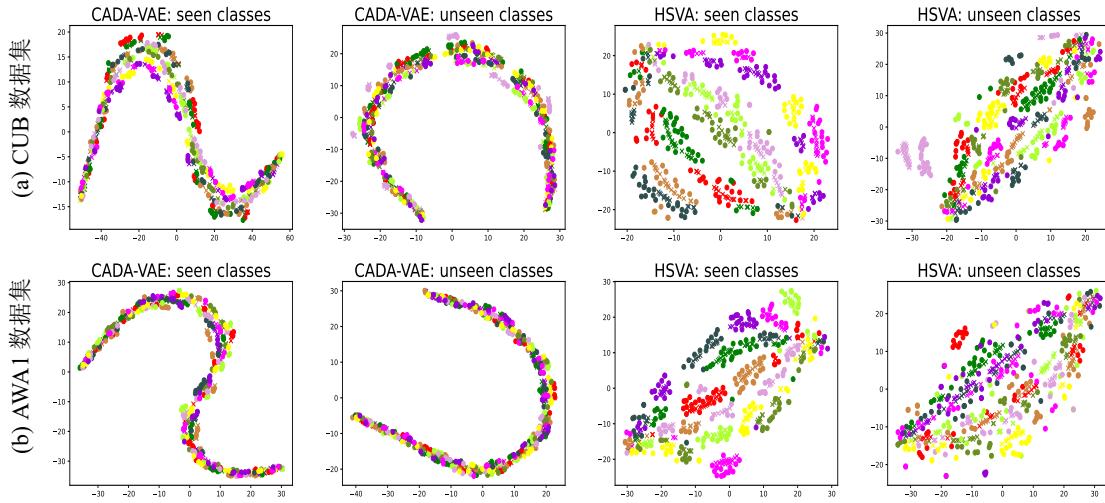


图 5.8 HSVA 和 CADA-VAE<sup>[3]</sup> 学习的子空间中视觉嵌入特征（图中的“○”）和语义嵌入特征（图中的“×”）的 t-SNE 可视化<sup>[1]</sup>。

本被错误分类。这直观地表明，现有基于单步适应的公共子空间式零样本图像分类方法忽略了视觉和语义域中异构特征表示的结构差异。相比之下，HSVA 学习了一个真正对准的公共子空间，该空间可以更好地实现视觉和语义异构特征的结构对准和分布对准，促进零样本图像分类进行有效的知识转移。因此，层次语义-视觉适应是实现公共子空间式零样本图像分类的有效解决方案。

表 5.3 在 CZSL 设置下，HSVA 和其他先进的公共子空间式零样本图像分类方法在 AWA1<sup>[9]</sup>、CUB<sup>[4]</sup>、SUN<sup>[5]</sup> 数据集上的实验结果对比。

当前先进的公共子空间式零样本图像分类方法	AWA1 数据集	CUB 数据集	SUN 数据集
	Acc	Acc	Acc
DeViSE <sup>[68]</sup>	54.2	52.0	56.5
DCN <sup>[46]</sup>	65.3	56.2	61.8
CADA-VAE <sup>[3]</sup>	63.0	59.8	61.7
<b>HSVA (本文方法)</b>	<b>70.6</b>	<b>62.8</b>	<b>63.8</b>

## 5.7 HSVA 和其他先进方法的对比

HSVA 是一种归纳式方法，本章节将 HSVA 与其他先进的归纳式零样本图像分类方法同时在 CZSL 和 GZSL 设置下进行比较，包括嵌入式方法（例如，CRnet<sup>[104]</sup>，PQZSL<sup>[124]</sup>，TCN<sup>[76]</sup>，DVBE<sup>[107]</sup>，DAZLE<sup>[81]</sup>，RGEN<sup>[39]</sup>，APN<sup>[41]</sup>）、生成式方法（例

表 5.4 在 GZSL 设置下，HSVA 和其他先进的零样本图像分类方法在 AWA1<sup>[9]</sup>、AWA2<sup>[6]</sup>、CUB<sup>[4]</sup>、SUN<sup>[5]</sup> 数据集上的实验结果对比。

当前先进的零样本图像分类方法	AWA1 数据集			AWA2 数据集			CUB 数据集			SUN 数据集		
	U	S	H	U	S	H	U	S	H	U	S	H
<b>嵌入式方法</b>												
CRnet <sup>[104]</sup>	58.1	74.7	65.4	-	-	-	45.5	56.8	50.5	34.1	36.5	35.3
AREN <sup>[2]</sup>	-	-	-	15.6	92.9	26.7	38.9	78.7	52.1	19.0	38.8	25.5
PQZSL <sup>[124]</sup>	-	-	-	31.7	70.9	43.8	43.2	51.4	46.9	35.1	35.3	35.2
TCN <sup>[76]</sup>	49.4	76.5	60.0	61.2	65.8	63.4	52.6	52.0	52.3	31.2	37.3	34.0
DVBE <sup>[107]</sup>	-	-	-	63.6	70.8	67.0	53.2	60.2	56.5	45.0	37.2	40.7
DAZLE <sup>[81]</sup>	-	-	-	60.3	75.7	67.1	59.6	56.7	58.1	52.3	24.3	33.2
RGEN <sup>[39]</sup>	-	-	-	<b>67.1</b>	76.5	<b>71.5</b>	60.0	<b>73.5</b>	66.1	44.0	31.7	36.8
APN <sup>[41]</sup>	-	-	-	56.5	78.0	65.5	<b>65.3</b>	69.3	<b>67.2</b>	41.9	34.0	37.6
<b>生成式方法</b>												
f-CLSWGAN <sup>[89]</sup>	57.9	61.4	59.6	-	-	-	43.7	57.7	49.7	42.6	36.6	39.4
f-VAEGAN <sup>[61]</sup>	-	-	-	57.6	70.6	63.5	48.4	60.1	53.6	45.1	38.0	41.3
LisGAN <sup>[95]</sup>	52.6	76.3	62.3	-	-	-	46.5	57.9	51.6	42.9	37.8	40.2
LsrGAN <sup>[102]</sup>	54.6	74.6	63.0	-	-	-	48.1	59.1	53.0	44.8	37.7	40.9
AGZSL <sup>[133]</sup>	-	-	-	65.1	78.9	71.3	41.4	49.7	45.2	29.9	<b>40.2</b>	34.3
<b>公共子空间式方法</b>												
DeViSE <sup>[68]</sup>	13.4	68.7	22.4	17.1	74.7	27.8	23.8	53.0	32.8	16.9	27.4	20.9
ReViSE <sup>[44]</sup>	46.1	37.1	41.1	46.4	39.7	42.8	37.6	28.3	32.3	24.3	20.1	22.0
DCN <sup>[46]</sup>	25.5	84.2	39.1	-	-	-	28.4	60.7	38.7	25.5	37.0	30.2
CADA-VAE <sup>[3]</sup>	57.3	72.8	64.1	55.8	75.0	63.9	51.6	53.5	52.4	47.2	35.7	40.6
SGAL <sup>[105]</sup>	52.7	74.0	61.5	52.5	<b>86.3</b>	65.3	40.9	55.3	47.0	35.5	34.4	34.9
<b>HSVA (本文方法)</b>	<b>59.3</b>	76.6	<b>66.8</b>	56.7	79.8	66.3	52.7	58.3	55.3	48.6	39.0	<b>43.3</b>

注：符号“-”表示相应结果缺失。

如，f-CLSWG<sup>[89]</sup>，f-VAEGAN<sup>[61]</sup>，LisGAN<sup>[95]</sup>，LsrGAN<sup>[102]</sup>，AGZSL<sup>[133]</sup> 等) 和公共子空间式方法（例如，DeViSE<sup>[68]</sup>，DCN<sup>[46]</sup>，CADA-VAE<sup>[3]</sup>，SGAL<sup>[105]</sup> 等）。

表5.3展示了 HSVA 和其他先进的公共子空间式零样本图像分类方法在 AWA1<sup>[9]</sup>、CUB<sup>[4]</sup>、SUN<sup>[5]</sup> 数据集上的 CZSL 实验结果。相较于其他先进的公共子空间式零样本图像分类方法（例如 DeViSE<sup>[68]</sup>，DCN<sup>[46]</sup>，CADA-VAE<sup>[3]</sup>），HSVA 在 AWA1、CUB 和 SUN 数据集上分别比至少提升了 5.3%、3.0% 和 2.0% 的识别精度。显著的性能提升表明，层次语义-视觉适应可以有效地学习真正的公共子空间将视觉和语义异构特征进行结构对准和分布对准，有效克服异构特征之间的分布和结构差异性，实现零样本图像分类有效的知识迁移。

表5.4展示了 HSVA 和其他先进的零样本图像分类方法（嵌入式方法<sup>[41,76,81,104,124]</sup>、生成式方法<sup>[61,89,95,102,133]</sup> 和公共子空间式方法<sup>[3,44,46,68,105]</sup>）在 AWA1<sup>[9]</sup>、AWA2<sup>[6]</sup>、CUB<sup>[4]</sup>、SUN<sup>[5]</sup> 数据集上的 GZSL 实验结果。相较于其他先进的公共子空间式零样

本图像分类方法（例如，DeViSE<sup>[68]</sup>，DCN<sup>[46]</sup>，CADA-VAE<sup>[3]</sup>，SGAL<sup>[105]</sup>等），本章节的HSVA在AWA1、AWA2、CUB和SUN等四个数据集上分别至少取得了2.7%、1.0%、2.9%和2.7%的性能( $\mathbf{H}$ )提升。特别指出，HSVA的性能明显的优于最新的公共子空间式方法(SGAL<sup>[105]</sup>)，在AWA1、AWA2、CUB和SUN等四个数据集上的性能指标 $\mathbf{H}$ 分别提高了5.3%、1.0%、8.3%和8.4%。由于HSVA也是一种生成式方法，因此本章节将其与其他先进的生成式方法进行比较。结果表明，HSVA在AWA1、CUB和SUN等三个数据集上的性能明显更优。相较于嵌入式方法，HSVA也取得了有竞争力的结果（例如，HSVA在SUN数据集上的性能指标调 $H$ 为43.3）。这些结果一致表明了基于层次语义-视觉适应的零样本图像分类方法的优越性和巨大潜力。

## 5.8 本章小结

本章研究了零样本图像分类中如何实现视觉-属性的异构特征对准的难题，从而实现公共子空间式零样本图像分类进行有效的视觉-语义交互。先前的方法使用单步适应的方式仅进行视觉和语义特征的分布对准，忽略了视觉-语义异构特征同时存在特征流形结构和特征分布的差异性，未能在公共子空间中实现视觉-语义特征的真正对准。本章节分析了异构特征之间同时存在特征流形结构和特征分布差异问题，并提出一种层次语义-视觉适应的零样本图像分类方法，通过层次适应的方式分别进行结构对准和分布对准，学习一个真正的公共子空间并有效促进零样本图像分类。本章节提出一种有监督的对抗差异性学习机制，实现视觉-语义异构特征的结构对准。在四个主流零样本标准数据集上的大量实验结果验证了提出方法的有效性。本章节工作也将有助于开发其他异构特征表示系统，包括视觉问答、图像字幕、基于自然语言的视觉推理等。

## 6 总结与展望

### 6.1 论文总结

本文针对基于深度表征的零样本图像分类面临的跨数据集偏差、视觉-语义的表示差异性、视觉-语义的异构性等三个重要问题展开工作：

- 针对基于深度表征的零样本图像分类任务中的跨数据集偏差问题，本文提出了基于视觉特征增强的零样本图像分类方法。基于嵌入式模型，本文第三章工作首先提出区域指导的图注意力网络通过挖掘局部特征之间的关系表示显式的全局视觉特征，并将其与局部视觉特征融合实现对视觉特征增强，从而促进视觉-语义的有效交互。基于生成式模型，本文第三章工作提出特征精细化学习机制对视觉特征进行增强，并促进生成器生成更真实的未知类伪视觉特征样本，从而训练一个有监督分类器实现更准确的零样本图像分类。实验结果表明，本文提出的这两种视觉增强方法能够有效地提高嵌入式和生成式方法中视觉特征的判别性和迁移性，从而促进视觉-语义的有效交互，进而取得领先的分类性能。
- 针对基于深度表征的零样本图像分类中视觉-语义特征的表示差异性问题，本文第四章提出了基于属性-视觉关键公共语义知识的零样本图像分类方法。首先，本文提出一种基于属性指导的 Transformer 方法，利用单向跨注意力机制学习具有属性定位的视觉特征，对属性-视觉特征之间的关键公共语义知识进行较为准确的表示。随后，本文提出一种基于互语义蒸馏网络的方法，通过属性 → 视觉和视觉 → 属性两个双向注意力子网络分别学习基于属性的视觉特征和基于视觉的属性特征，在互语义蒸馏学习机制的指导下，使得两个子网络更充分地挖掘视觉-属性潜在的语义知识。最后，本文也将这两种方法集成成为一个统一的模型，从而充分且准确地对关键公共语义知识进行表示，提高零样本图像分类从已知类到未知类的知识迁移能力，取得稳定的性能提升。
- 针对基于深度表征的零样本图像分类中视觉-语义特征的异构性问题，本文第五章提出一种层次语义-视觉适应的零样本图像分类方法。不同于现有方法使用单步适应只进行异构特征分布对准，该工作使用层次适应的学习方式同时进行异构特征的流形结构对准和分布对准，学习一个本真的公共子空间实现视觉和语

义特征对准，并使用子空间特征学习一个有监督分类器进行有效地零样本图像分类。实验结果表明了分布对准和结构对准分别是实现异构特征对准的基本任务和必要操作，本文提出的方法为公共子空间式零样本图像分类提供了新的研究思路。

## 6.2 展望

对于未来工作，有如下规划和展望：

- 零样本图像分类在跨数据集偏差问题上仍有待进一步解决。本文工作虽然初步探索了使用不同视觉特征增强的方法缓解嵌入式和生成式零样本图像分类的跨数据偏差问题，但该工作需要对不同类型的零样图像分类方法设计不同的特征增强方案。未来工作可以研究一个统一有效的特征增强方法实现对任意零样本学习方法的视觉特征进行增强。
- 本文初次探讨了基于 Transformer 和语义蒸馏的零样本图像分类方法，用于挖掘视觉-属性特征之间关键公共语义知识，但仍有较大的探索空间。未来工作可以关注：如何利用 Transformer 实现对同类别不同视角的视觉特征和属性特征之间的语义知识表示；如何利用无监督对比学习<sup>[143]</sup>的思想通过视觉和属性特征表示之间的对比学习进一步挖掘视觉-属性特征之间的语义知识；如何借助现有的视觉-语言大模型（例如，CLIP<sup>[144]</sup>，Imagen<sup>[145]</sup>）实现视觉-属性特征之间有效的语义蒸馏等具有研究潜力的方向。
- 本文分析了视觉-语义特征的异构性同时存在流形结构和特征分布差异性，并提出层次适应的方式学习一个本真的子空间实现视觉-语义异构特征对准。图5.8显示，本文方法在子空间中虽然实现了较好的视觉和语义特征对准，但是这些特征的类别关系表示较为混乱，从而造成分类效果仍较为有限。为此，未来工作研究如何在层次适应结构中进行视觉-语义对准的同时有效地学习类别关系表示，从而进一步提高公共子空间中不同类别视觉和语义特征的可区分性。

## 致 谢

古有十年寒窗苦读而名扬天下，今有二十余载学习深造而初入科研的殿堂。一路走来，虽路途崎岖，但幸于恩师、朋友、家人的帮助，终达彼岸。此刻，心情舒畅的同时，更多的是心存感恩。

感谢我的恩师尤新革教授，是他以大师的风度、深邃的思维、宽阔的视野成就了我博士三年求学期间收获良多。恩师在科研上给我灌输了非常多的研究思维和研究方法，指导我做有意义、有价值的研究课题。恩师在我人生规划和人生价值上给予了非常多的建议，让我更清楚我努力的方向，更健全了我的人生观和价值观。恩师是我的良师益友，是我一生的恩师。

感谢实验室的袁魏老师、彭勤牧老师、邵远杰老师，在我日常科研中提供了很多的建议和帮助。特别感谢我的师弟也是挚友洪梓铭、王文杰，我们一起学术讨论、日常干饭的时刻令我终生难忘。我也要感谢实验室其他的同门杨传武、史玉峰、夏北浩、李珩、尹诗、张鹏、李博、袁佩佩、候文金、王聪浩等同学一起科研探讨和聚餐团建，祝你们前程似锦。

感谢特斯联首席科学家和国际总裁邵岭教授、南京理工大学的谢国森教授、南方科技大学的郑峰教授、军事科学院的赵健教授、南洋理工大学的杨文瀚老师、墨尔本大学的宫明明教授、悉尼大学的刘同亮教授、南通大学的丁卫平教授、南澳大学的曹泽宏教授等在学术上给与了很多的指导和建议。感谢阿里巴巴达摩院的孙佰贵和刘洋师兄、腾讯 AI Lab 的宋奕兵师兄在我实习期间给与的指导和帮助。

感谢我的父母从小对我的培养和无私的支持与鼓励，让我毫无顾虑的一直追求自己的理想。感谢一直关心我、支持我的所有亲人。

感谢的话永远也说不完，感恩的人也太多太多，那就铭记于心吧。我将带着感恩的心，朝着目标方向继续前行。

2022 年 7 月  
于国家防伪工程技术研究中心，华中科技大学

## 参考文献

- [1] L. V. D. Maaten, G. E. Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 2008, 9:2579-2605
- [2] G.-S. Xie, L. Liu, X. Jin, F. Zhu, Z. Zhang, J. Qin, et al. Attentive Region Embedding Network for Zero-Shot Learning. in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019:9376-9385
- [3] E. Schönfeld, S. Ebrahimi, S. Sinha, T. Darrell, Z. Akata. Generalized Zero- and Few-Shot Learning via Aligned Variational Autoencoders. in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019:8239-8247
- [4] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. J. Belongie, et al. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, Caltech,, 2010
- [5] G. Patterson, J. Hays. SUN attribute database: Discovering, annotating, and recognizing scene attributes. in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012:2751-2758
- [6] Y. Xian, C. H. Lampert, B. Schiele, Z. Akata. Zero-Shot Learning—A Comprehensive Evaluation of the Good, the Bad and the Ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, 41:2251-2265
- [7] S. Narayan, A. Gupta, F. Khan, C. G. M. Snoek, L. Shao. Latent Embedding Feedback and Discriminative Features for Zero-Shot Classification. in: *European Conference on Computer Vision (ECCV)*, 2020
- [8] S. Chen, G.-S. Xie, Y. Yang Liu, Q. Peng, B. Sun, H. Li, et al. HSVA: Hierarchical Semantic-Visual Adaptation for Zero-Shot Learning. in: *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2021
- [9] C. H. Lampert, H. Nickisch, S. Harmeling. Attribute-Based Classification for Zero-Shot Visual Object Categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, 36:453-465
- [10] V. N. Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 1999, 10(5):988-99
- [11] A. Krizhevsky, I. Sutskever, G. E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 2012, 60:84 - 90
- [12] M. Tan, Q. V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. in: *International Conference on Machine Learning (ICML)*, 2019

# 华 中 科 技 大 学 博 士 学 位 论 文

---

- [13] X. Zhu, A. B. Goldberg. Introduction to Semi-Supervised Learning. MIT Press, 2009
- [14] S. C. Fralick. Learning to recognize patterns without a teacher. IEEE Transaction Information Theory, 1967, 13:57-64
- [15] Z. Wu, Y. Xiong, S. X. Yu, D. Lin. Unsupervised Feature Learning via Non-parametric Instance Discrimination. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018:3733-3742
- [16] K. He, H. Fan, Y. Wu, S. Xie, R. B. Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020:9726-9735
- [17] A. Dosovitskiy, J. T. Springenberg, M. A. Riedmiller, T. Brox. Discriminative Unsupervised Feature Learning with Convolutional Neural Networks. in: Annual Conference on Neural Information Processing Systems (NeurIPS), 2014
- [18] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, et al. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, 2015, 115:211-252
- [19] I. Biederman. Recognition-by-components: a theory of human image understanding. Psychological review, 1987, 94:115-147
- [20] H. Larochelle, D. Erhan, Y. Bengio. Zero-data Learning of New Tasks. in: AAAI Conference on Artificial Intelligence (AAAI), 2008:646-651
- [21] M. Palatucci, D. Pomerleau, G. E. Hinton, T. M. Mitchell. Zero-shot Learning with Semantic Output Codes. in: Annual Conference on Neural Information Processing Systems (NeurIPS), 2009:1410-1418
- [22] C. H. Lampert, H. Nickisch, S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009:951-958
- [23] 冯耀功, 于剑, 桑基韬, 杨朋波. 基于知识的零样本视觉识别综述. 软件学报, 2021, 32(02):370-405
- [24] 宋杰. 基于深度模型的零样本迁移学习: [PhD Dissertation]. 浙江大学, 2020
- [25] 于云龙. Research on Zero-Shot Image Learning: [PhD Dissertation]. 天津大学, 2019
- [26] 徐晓峰. 基于属性的零样本学习方法研究: [PhD Dissertation]. 南京理工大学, 2020
- [27] Y. Fu, T. Xiang, Y.-G. Jiang, X. Xue, L. Sigal, S. Gong. Recent Advances in Zero-Shot Recognition: Toward Data-Efficient Understanding of Visual Content. IEEE Signal Processing Magazine, 2018, 35:112-125

# 华 中 科 技 大 学 博 士 学 位 论 文

---

- [28] 刘友发. 领域自适应和零样本学习的知识迁移方法研习: [PhD Dissertation]. 武汉大学, 2020
- [29] 张鲁宁, 左信, 刘建伟. 基于典型相关分析和距离度量学习的零样本学习. 自动化学报, 2020, 46(01):1-23
- [30] 冀中, 汪浩然, 于云龙, 庞彦伟. 零样本图像分类综述: 十年进展. 中国科学: 信息科学, 2019, 49(10):1299-1320
- [31] K. He, X. Zhang, S. Ren, J. Sun. Deep Residual Learning for Image Recognition. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016:770-778
- [32] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al. Attention is All you Need. in: Annual Conference on Neural Information Processing Systems (NeurIPS), 2017
- [33] A. Torralba, A. A. Efros. Unbiased look at dataset bias. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011
- [34] X. Li, Y. Grandvalet, F. Davoine. Explicit Inductive Bias for Transfer Learning with Convolutional Networks. in: International Conference on Machine Learning (ICML), 2018
- [35] D. Hendrycks, K. Lee, M. Mazeika. Using Pre-Training Can Improve Model Robustness and Uncertainty. in: International Conference on Machine Learning (ICML), 2019
- [36] M. Raghu, C. Zhang, J. Kleinberg, S. Bengio. Transfusion: Understanding Transfer Learning for Medical Imaging. in: Annual Conference on Neural Information Processing Systems (NeurIPS), 2019
- [37] Y. Guo, H. Shi, A. Kumar, K. Grauman, T. Simunic, R. S. Feris. SpotTune: Transfer Learning Through Adaptive Fine-Tuning. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019:4800-4809
- [38] Y. Guo, Y. Li, L. Wang, T. Simunic. AdaFilter: Adaptive Filter Fine-tuning for Deep Transfer Learning. in: AAAI Conference on Artificial Intelligence (AAAI), 2020
- [39] G.-S. Xie, L. Liu, X. Jin, F. Zhu, Z. Zhang, Y. Yao, et al. Region Graph Embedding Network for Zero-Shot Learning. in: European Conference on Computer Vision (ECCV), 2020
- [40] Y. Zhu, J. Xie, Z. Tang, X. Peng, A. Elgammal. Semantic-Guided Multi-Attention Localization for Zero-Shot Learning. in: Annual Conference on Neural Information Processing Systems (NeurIPS), 2019

# 华中科技大学博士学位论文

---

- [41] W. Xu, Y. Xian, J. Wang, B. Schiele, Z. Akata. Attribute Prototype Network for Zero-Shot Learning. in: Annual Conference on Neural Information Processing Systems (NeurIPS), 2020
- [42] Y. Liu, L. Zhou, X. Bai, Y. Huang, L. Gu, J. Zhou, et al. Goal-Oriented Gaze Estimation for Zero-Shot Learning. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021
- [43] Z. Zhang, V. Saligrama. Zero-Shot Learning via Joint Latent Similarity Embedding. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016:6034-6042
- [44] Y.-H. H. Tsai, L.-K. Huang, R. Salakhutdinov. Learning Robust Visual-Semantic Embeddings. in: IEEE International Conference on Computer Vision (ICCV), 2017:3591-3600
- [45] Q. Wang, K. Chen. Zero-Shot Visual Recognition via Bidirectional Latent Embedding. International Journal of Computer Vision, 2017, 124:356-383
- [46] S. Liu, M. Long, J. Wang, M. I. Jordan. Generalized Zero-Shot Learning with Deep Calibration Network. in: Annual Conference on Neural Information Processing Systems (NeurIPS), 2018
- [47] H. Li, S. J. Pan, R. Wan, A. Kot. Heterogeneous Transfer Learning via Deep Matrix Completion with Adversarial Kernel Embedding. in: AAAI Conference on Artificial Intelligence (AAAI), 2019
- [48] Y. Luo, Y. Wen, T. Liu, D. Tao. Transferring Knowledge Fragments for Learning Distance Metric from a Heterogeneous Domain. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41:1013-1026
- [49] Z. Fu, T. Xiang, E. Kodirov, S. Gong. Zero-Shot Learning on Semantic Class Prototype Graph. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40:2009-2022
- [50] Z. Fu, E. Kodirov, T. Xiang, S. Gong. Zero-shot object recognition by semantic manifold distance. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015:2635-2644
- [51] G. LoweDavid. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision, 2004
- [52] H. Bay, T. Tuytelaars, L. V. Gool. SURF: Speeded Up Robust Features. in: European Conference on Computer Vision (ECCV), 2006
- [53] N. Dalal, B. Triggs. Histograms of oriented gradients for human detection. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005:886-89

# 华中科技大学博士学位论文

---

- [54] E. Shechtman, M. Irani. Matching Local Self-Similarities across Images and Videos. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007
- [55] K. Simonyan, A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. in: International Conference on Learning Representations (ICLR), 2015
- [56] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, et al. Going deeper with convolutions. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015:1-9
- [57] L.-J. Li, H. Su, E. P. Xing, L. Fei-Fei. Object Bank: A High-Level Image Representation for Scene Classification & Semantic Feature Sparsification. in: Annual Conference on Neural Information Processing Systems (NeurIPS), 2010
- [58] 张嘉睿. 基于属性挖掘的零样本图像分类: [PhD Dissertation]. 中国矿业大学, 2020
- [59] Y. Xian, B. Schiele, Z. Akata. Zero-Shot Learning —The Good, the Bad and the Ugly. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, : 3077-3086
- [60] J. Pennington, R. Socher, C. D. Manning. Glove: Global Vectors for Word Representation. in: EMNLP, 2014
- [61] Y. Xian, S. Sharma, B. Schiele, Z. Akata. F-VAEGAN-D2: A Feature Generating Framework for Any-Shot Learning. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019:10267-10276
- [62] 李亚南. 零样本学习关键技术研究: [PhD Dissertation]. 浙江大学, 2018
- [63] 冀中, 孙涛, 于云龙. 一种基于直推判别字典学习的零样本分类方法. 软件学报, 2017, 28(11):2961-2970
- [64] Y. Fu, T. M. Hospedales, T. Xiang, S. Gong. Learning Multimodal Latent Attributes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36:303-316
- [65] M. Rohrbach, M. Stark, B. Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011:1641-1648
- [66] Y. Fu, T. M. Hospedales, T. Xiang, S. Gong. Transductive Multi-View Zero-Shot Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37:2332-2345
- [67] Z. Akata, F. Perronnin, Z. Harchaoui, C. Schmid. Label-Embedding for Image Classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38:1425-1438

# 华 中 科 技 大 学 博 士 学 位 论 文

---

- [68] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, et al. DeViSE: A Deep Visual-Semantic Embedding Model. in: Annual Conference on Neural Information Processing Systems (NeurIPS), 2013
- [69] Z. Akata, S. Reed, D. Walter, H. Lee, B. Schiele. Evaluation of output embeddings for fine-grained image classification. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015:2927-2936
- [70] B. Romera-Paredes, P. H. S. Torr. An embarrassingly simple approach to zero-shot learning. in: International Conference on Machine Learning (ICML), 2015
- [71] 冀中, 谢于中, 庞彦伟. 基于典型相关分析和距离度量学习的零样本学习. 天津大学学报(自然科学与工程技术版), 2017, 50(08):813-820
- [72] J. Song, C. Shen, Y. Yang, Y. Liu, M. Song. Transductive Unbiased Embedding for Zero-Shot Learning. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018:1024-1033
- [73] Y. Li, J. Zhang, J. Zhang, K. Huang. Discriminative Learning of Latent Features for Zero-Shot Recognition. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018:7463-7471
- [74] Y. Liu, J. Guo, D. Cai, X. He. Attribute Attention for Semantic Disambiguation in Zero-Shot Learning. in: IEEE International Conference on Computer Vision (ICCV), 2019:6697-6706
- [75] Y. L. Cacheux, H. Borgne, M. Crucianu. Modeling Inter and Intra-Class Relations in the Triplet Loss for Zero-Shot Learning. in: IEEE International Conference on Computer Vision (ICCV), 2019:10332-10341
- [76] H. Jiang, R. Wang, S. Shan, X. Chen. Transferable Contrastive Network for Generalized Zero-Shot Learning. in: IEEE International Conference on Computer Vision (ICCV), 2019:9764-9773
- [77] C. Wang, X. Chen, S. Min, X. Sun, H. Li. Task-Independent Knowledge Makes for Transferable Representations for Generalized Zero-Shot Learning. in: AAAI Conference on Artificial Intelligence (AAAI), 2021
- [78] S. Jia, Z. Li, N. Chen, J. Zhang. Towards Visual Explainable Active Learning for Zero-Shot Classification. IEEE Transactions on Visualization and Computer Graphics, 2022, 28:791-801
- [79] R. Socher, M. Ganjoo, C. D. Manning, A. Ng. Zero-Shot Learning Through Cross-Modal Transfer. in: Annual Conference on Neural Information Processing Systems (NeurIPS), 2013

# 华中科技大学博士学位论文

---

- [80] C. Wang, S. Min, X. Chen, X. Sun, H. Li. Dual Progressive Prototype Network for Generalized Zero-Shot Learning. in: Annual Conference on Neural Information Processing Systems (NeurIPS), 2021
- [81] D. Huynh, E. Elhamifar. Fine-Grained Generalized Zero-Shot Learning via Dense Attribute-Based Attention. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020:4482-4492
- [82] Y. Yu, Z. Ji, Y. Fu, J. Guo, Y. Pang, Z. Zhang. Stacked Semantic-Guided Attention Model for Fine-Grained Zero-Shot Learning. in: Annual Conference on Neural Information Processing Systems (NeurIPS), 2018
- [83] D. P. Kingma, M. Welling. Auto-Encoding Variational Bayes. in: International Conference on Learning Representations (ICLR), 2014
- [84] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al. Generative Adversarial Nets. in: Annual Conference on Neural Information Processing Systems (NeurIPS), 2014
- [85] D. J. Rezende, S. Mohamed. Variational Inference with Normalizing Flows. in: International Conference on Machine Learning (ICML), 2015
- [86] L. Dinh, D. Krueger, Y. Bengio. NICE: Non-linear Independent Components Estimation. in: International Conference on Learning Representations Workshop, 2015
- [87] 刘欢, 郑庆华, 赵洪科, 肖阳, 吕彦章. 基于跨域对抗学习的零样本分类. 计算机研究与发展, 2019, 56(12):2521-2535
- [88] G. Arora, V. Verma, A. Mishra, P. Rai. Generalized Zero-Shot Learning via Synthesized Examples. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018:4281-4289
- [89] Y. Xian, T. Lorenz, B. Schiele, Z. Akata. Feature Generating Networks for Zero-Shot Learning. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018:5542-5551
- [90] M. Bucher, S. Herbin, F. Jurie. Generating Visual Representations for Zero-Shot Classification. in: IEEE International Conference on Computer Vision Workshop, 2017:2666-2673
- [91] X. Zhao, Y. Shen, S. Wang, H. Zhang. Boosting Generative Zero-Shot Learning by Synthesizing Diverse Features with Attribute Augmentation. in: AAAI Conference on Artificial Intelligence (AAAI), 2022
- [92] D. Chen, Y. Shen, H. Zhang, P. H. S. Torr. Zero-Shot Logit Adjustment. in: International Joint Conference on Artificial Intelligence (IJCAI), 2022

# 华中科技大学博士学位论文

---

- [93] R. Felix, B. V. Kumar, I. D. Reid, G. Carneiro. Multi-modal Cycle-consistent Generalized Zero-Shot Learning. in: European Conference on Computer Vision (ECCV), 2018
- [94] Z. Han, Z. Fu, J. Yang. Learning the Redundancy-Free Features for Generalized Zero-Shot Object Recognition. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020:12862-12871
- [95] J. Li, M. Jing, K. Lu, Z. Ding, L. Zhu, Z. Huang. Leveraging the Invariant Side of Generative Zero-Shot Learning. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019:7394-7403
- [96] Z. Chen, Y. Luo, R. Qiu, S. Wang, Z.-Y. Huang, J. Li, et al. Semantics Disentangling for Generalized Zero-Shot Learning. in: IEEE International Conference on Computer Vision (ICCV), 2021:8692-8700
- [97] Z. Zhang, V. Saligrama. Zero-Shot Learning via Semantic Similarity Embedding. in: International Conference on Computer Vision (ICCV), 2015:4166-4174
- [98] S. Changpinyo, W.-L. Chao, F. Sha. Predicting Visual Exemplars of Unseen Classes for Zero-Shot Learning. International Conference on Computer Vision (ICCV), 2017, : 3496-3505
- [99] 赵鹏, 汪纯燕, 张思颖, 刘政怡. 一种基于融合重构的子空间学习的零样本图像分类方法. 计算机学报, 2021, 44(02):409-421
- [100] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, B. Schiele. Latent Embeddings for Zero-Shot Classification. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016:69-77
- [101] P. Morgado, N. Vasconcelos. Semantically Consistent Regularization for Zero-Shot Recognition. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017:2037-2046
- [102] M. R. Vyas, H. Venkateswara, S. Panchanathan. Leveraging Seen and Unseen Semantic Relationships for Generative Zero-Shot Learning. in: European Conference on Computer Vision (ECCV), 2020
- [103] A. Mishra, M. K. Reddy, A. Mittal, H. Murthy. A Generative Model for Zero Shot Learning Using Conditional Variational Autoencoders. in: IEEE Conference on Computer Vision and Pattern Recognition Workshop, 2018:2269-2277
- [104] F. Zhang, G. Shi. Co-Representation Network for Generalized Zero-Shot Learning. in: International Conference on Machine Learning (ICML), 2019
- [105] H. Yu, B. Lee. Zero-shot Learning via Simultaneous Generating and Learning. in: Annual Conference on Neural Information Processing Systems (NeurIPS), 2019

# 华中科技大学博士学位论文

---

- [106] Y. Shen, J. Qin, L. Huang. Invertible Zero-Shot Recognition Flows. in: European Conference on Computer Vision (ECCV), 2020
- [107] S. Min, H. Yao, H. Xie, C. Wang, Z. Zha, Y. Zhang. Domain-Aware Visual Bias Eliminating for Generalized Zero-Shot Learning. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020:12661-12670
- [108] L. Chen, H. Zhang, J. Xiao, W. Liu, S. Chang. Zero-Shot Visual Recognition Using Semantics-Preserving Adversarial Embedding Networks. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018:1043-1052
- [109] K. Li, M. R. Min, Y. R. Fu. Rethinking Zero-Shot Learning: A Conditional Visual Classification Perspective. in: IEEE International Conference on Computer Vision (ICCV), 2019:3582-3591
- [110] Y. Liu, D. Xie, Q. Gao, J. Han, S. Wang, X. Gao. Graph and Autoencoder Based Feature Extraction for Zero-shot Learning. in: International Joint Conference on Artificial Intelligence (IJCAI), 2019
- [111] Y. Yu, Z. Ji, J. Han, Z. Zhang. Episode-Based Prototype Generating Network for Zero-Shot Learning. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020:14032-14041
- [112] N. Quadrianto, J. Petterson, A. Smola. Distribution Matching for Transduction. in: Annual Conference on Neural Information Processing Systems (NeurIPS), 2009
- [113] A. Gretton, B. K. Sriperumbudur, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, et al. Optimal kernel choice for large-scale two-sample tests. in: Annual Conference on Neural Information Processing Systems (NeurIPS), 2012
- [114] M. Defferrard, X. Bresson, P. Vandergheynst. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. in: Annual Conference on Neural Information Processing Systems (NeurIPS), 2016
- [115] T. Kipf, M. Welling. Semi-Supervised Classification with Graph Convolutional Networks. in: International Conference on Learning Representations (ICLR), 2016
- [116] W. L. Hamilton, Z. Ying, J. Leskovec. Inductive Representation Learning on Large Graphs. in: Annual Conference on Neural Information Processing Systems (NeurIPS), 2017
- [117] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio' , Y. Bengio. Graph Attention Networks. in: International Conference on Learning Representations (ICLR), 2018

# 华中科技大学博士学位论文

---

- [118] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. in: International Conference on Machine Learning (ICML), 2015
- [119] S. Yun, M. Jeong, R. Kim, J. Kang, H. J. Kim. Graph Transformer Networks. in: Annual Conference on Neural Information Processing Systems (NeurIPS), 2019
- [120] D. Cai, W. Lam. Graph Transformer for Graph-to-Sequence Learning. in: AAAI Conference on Artificial Intelligence (AAAI), 2020
- [121] Z. Hu, Y. Dong, K. Wang, Y. Sun. Heterogeneous Graph Transformer. in: WWW, 2020
- [122] G.-S. Xie, Z. Zhang, G.-S. Liu, F. Zhu, L. Liu, L. Shao, et al. Generalized Zero-Shot Learning With Multiple Graph Adaptive Generative Networks. IEEE transactions on neural networks and learning systems, 2021
- [123] Z. Yue, T. Wang, H. Zhang, Q. Sun, X. Hua. Counterfactual Zero-Shot and Open-Set Visual Recognition. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021
- [124] J. Li, X. Lan, Y. Liu, L. Wang, N. Zheng. Compressing Unknown Images With Product Quantizer for Efficient Zero-Shot Classification. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019:5458-5467
- [125] R. Keshari, R. Singh, M. Vatsa. Generalized Zero-Shot Learning via Over-Complete Distribution. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020:13297-13305
- [126] H. Huang, C.-D. Wang, P. S. Yu, C.-D. Wang. Generative Dual Adversarial Network for Generalized Zero-Shot Learning. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019:801-810
- [127] Y. Wen, K. Zhang, Z. Li, Y. Qiao. A Discriminative Feature Learning Approach for Deep Face Recognition. in: European Conference on Computer Vision (ECCV), 2016
- [128] F. Schroff, D. Kalenichenko, J. Philbin. FaceNet: A unified embedding for face recognition and clustering. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015:815-823
- [129] A. L. Maas. Rectifier Nonlinearities Improve Neural Network Acoustic Models. in: International Conference on Machine Learning (ICML), 2013
- [130] D. P. Kingma, J. Ba. Adam: A Method for Stochastic Optimization. in: International Conference on Learning Representations (ICLR), 2015

# 华中科技大学博士学位论文

---

- [131] C. Yan, X. Chang, Z. Li, Z. Ge, W. Guan, L. Zhu, et al. ZeroNAS: Differentiable Generative Adversarial Networks Search for Zero-Shot Learning. *IEEE transactions on pattern analysis and machine intelligence*, 2021
- [132] D. T. Huynh, E. Elhamifar. Compositional Zero-Shot Learning via Fine-Grained Dense Feature Composition. in: Annual Conference on Neural Information Processing Systems (NeurIPS), 2020
- [133] T.-L. L. Yu-Ying Chou. Adaptive and Generative Zero-Shot Learning. in: International Conference on Learning Representations (ICLR), 2021
- [134] Y. Atzmon, F. Kreuk, U. Shalit, G. Chechik. A causal view of compositional zero-shot recognition. in: Annual Conference on Neural Information Processing Systems (NeurIPS), 2020
- [135] T. Chen, T. Pu, Y. Xie, H. Wu, L. Liu, L. Lin. Cross-Domain Facial Expression Recognition: A Unified Evaluation Benchmark and Adversarial Graph Learning. *IEEE transactions on pattern analysis and machine intelligence*, 2021
- [136] S. Herdade, A. Kappeler, K. Boakye, J. Soares. Image Captioning: Transforming Objects into Words. in: Annual Conference on Neural Information Processing Systems (NeurIPS), 2019
- [137] X. Zhang, X. Sun, Y. Luo, J. Ji, Y. Zhou, Y. Wu, et al. RSTNet: Captioning with Adaptive Attention on Visual and Non-Visual Words. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021
- [138] C. R. Givens, R. M. Shortt. A class of Wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 1984, 31:231-240
- [139] K. Saito, K. Watanabe, Y. Ushiku, T. Harada. Maximum Classifier Discrepancy for Unsupervised Domain Adaptation. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018:3723-3732
- [140] C.-Y. Lee, T. Batra, M. H. Baig, D. Ulbricht. Sliced Wasserstein Discrepancy for Unsupervised Domain Adaptation. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019:10277-10287
- [141] B. Sun, J. Feng, K. Saenko. Return of Frustratingly Easy Domain Adaptation. in: AAAI Conference on Artificial Intelligence (AAAI), 2016
- [142] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Józefowicz, S. Bengio. Generating Sentences from a Continuous Space. in: CoNLL, 2016
- [143] T. Chen, S. Kornblith, M. Norouzi, G. E. Hinton. A Simple Framework for Contrastive Learning of Visual Representations. in: International Conference on Machine Learning (ICML), 2020

# 华 中 科 技 大 学 博 士 学 位 论 文

---

- [144] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, et al. Learning Transferable Visual Models From Natural Language Supervision. in: International Conference on Machine Learning (ICML), 2021
- [145] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. arXiv preprint arXiv:2205.11487, 2022

## 附录 1 答辩委员会决议

首行缩进两个字符，中文字体采用小四宋体，英文字体采用 Times New Roman，  
字体大小为小四，行间距为固定值 20 磅。

## 附录 2 攻读博士学位期间取得的研究成果

### 已发表论文

- [1] **Shiming Chen**, Ziming Hong, Guo-Sen Xie, Wenhan Yang, Qinmu Peng, Kai Wang, Jian Zhao and Xinge You. MSDN: Mutually Semantic Distillation Network for Zero-Shot Learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022. (CCF-A 类会议；署名单位: 华中科技大学)
- [2] **Shiming Chen**, Ziming Hong, Yang Liu, Baigui Sun, Guo-Sen Xie, Xinge You and Ke Lv. TransZero: Attribute-guided Transformer for Zero-Shot Learning. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2022. (CCF-A 类会议；署名单位: 华中科技大学)
- [3] **Shiming Chen**, Guo-Sen Xie, Qinmu Peng, Yang Liu, Baigui Sun, Hao Li, Xinge You, Ling Shao. HSVA: Hierarchical Semantic-Visual Adaptation for Zero-Shot Learning. In: Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS), 2021. (CCF-A 类会议；署名单位: 华中科技大学)
- [4] **Shiming Chen**, Wenjie Wang, Beihao Xia, Qinmu Peng, Xinge You, Feng Zheng, Ling Shao. FREE: Feature Refinement for Generalized Zero-shot Learning. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2021. (CCF-A 类会议；署名单位: 华中科技大学)
- [5] **Shiming Chen**, Ziming Hong, Guo-Sen Xie, Xinge You, Weiping Ding and Ling Shao. GNDAN: Graph Navigated Dual Attention Network for Zero-Shot Learning. IEEE Transactions on Neural Networks and Learning Systems, 2022. (SCI 源刊；中科院大类一区；署名单位: 华中科技大学)
- [6] **Shiming Chen**, Wenjie Wang, Beihao Xia, Xinge You, Qinmu Peng, Zehong Cao, Weiping Ding. CDE-GAN: Cooperative Dual Evolution Based Generative Adversarial Network. IEEE Transactions on Evolutionary Computation, 2021, 25(5): 986-1000. (SCI 源刊；中科院大类一区；署名单位: 华中科技大学)
- [7] **Shiming Chen**, Peng Zhang, Guo-Sen Xie, Qinmu Peng, Zehong Cao, Wei Yuan,

# 华中科技大学博士学位论文

---

- Xinge You. Kernelized Similarity Learning and Embedding for Dynamic Texture Synthesis. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2022. (SCI 源刊; 中科院大类一区; 署名单位: 华中科技大学)
- [8] Ziming Hong\*, **Shiming Chen\***<sup>#</sup>, Guo-Sen Xie, Yuanjie Shao, Qinmu Peng, Jian Zhao, Xinge You. Semantic Compression Embedding for Zero-Shot Learning. In: *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2022. (共同一作, 通讯作者, CCF A-类会议; 署名单位: 华中科技大学)
- [9] Wenjie Wang, Yufeng Shi, **Shiming Chen**, Qinmu Peng, Feng Zheng, Xinge You. Norm-guided Adaptive Visual Embedding for Zero-Shot Sketch-Based Image Retrieval. In: *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2021. (CCF-A 类会议; 署名单位: 华中科技大学)

## 待发表论文

- [10] **Shiming Chen**, Ziming Hong, Guo-Sen Xie, Jian Yang, Hao Li, Xinge You, Shuicheng Yan and Ling Shao. TransZero++: Cross Attribute-Guided Transformer for Zero-Shot Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. (SCI 源刊; 中科院大类一区; CCF-A 类期刊; 署名单位: 华中科技大学; Minor Revision)
- [11] **Shiming Chen**, Ziming Hong, Guo-Sen Xie, Yibing Song, Jian Yang, Xinge You, Shuicheng Yan and Ling Shao. ViFR: Visual Feature Refinement for Zero-Shot Learning. *International Journal of Computer Vision*, 2022. (SCI 源刊; 中科院大类二区; CCF-A 类期刊; 署名单位: 华中科技大学; Under Review)
- [12] **Shiming Chen**, Wenjin Hou, Ziming Hong, Yibing Song, Tongliang Liu, Xinge You and Kun Zhang. DSP: Dynamic Semantic Prototype for Zero-Shot Learning. *International Conference on Learning Representations (ICLR)*, 2023. (机器学习 3 大会之一; THU-A 类会议; 署名单位: 华中科技大学; Submitted)
- [13] Shuhuang Chen, **Shiming Chen**<sup>#</sup>, Debing Zhang, Haofan Wang, Xinge You. DSP: Rethinking Attribute Localization for Zero-Shot Learning. Submitted to *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. (通讯作者; CCF-A 类会议; 署名单位: 华中科技大学)

# 华 中 科 技 大 学 博 士 学 位 论 文

---

## 专 利

- [1] 彭勤牧, 尤新革, 陈使明. 一种基于核相似度嵌入的动态纹理生成方法及装置. 中国, 发明专利, CN112613537A

### 附录 3 公开发表的学术论文与博士学位论文的关系

序号	成果名称	成果形式	成果主要内容	与学位论文对应的关系
1	<b>Shiming Chen</b> , Guo-Sen Xie, Qinmu Peng, Yang Liu, Baigui Sun, Hao Li, Xinge You, Ling Shao. HSVA: Hierarchical Semantic-Visual Adaptation for Zero-Shot Learning. In: Proceedings of the Annual Conference on Neural Information Processing Systems ( <b>NeurIPS</b> ), 2021.	第一作者; CCF-A类会议	提出层次语义-视觉适应的零样本图像分类方法实现视觉-语义异构特征的对准	第五章
2	<b>Shiming Chen</b> , Wenjie Wang, Beihao Xia, Qinmu Peng, Xinge You, Feng Zheng, Ling Shao. FREE: Feature Refinement for Generalized Zero-shot Learning. In: Proceedings of the IEEE International Conference on Computer Vision ( <b>ICCV</b> ), 2021.	第一作者; CCF-A类会议	提出一种基于后置视觉特征增强的生成式零样本图像分类方法解决跨数据集偏差问题	第三章
3	<b>Shiming Chen</b> , Wenjie Wang, Beihao Xia, Xinge You, Qinmu Peng, Zehong Cao, Weiping Ding. CDE-GAN: Cooperative Dual Evolution Based Generative Adversarial Network. <i>IEEE Transactions on Evolutionary Computation</i> , 2021, 25(5): 986-1000.	第一作者; SCI 源刊; 中科院大 类一区	提出一种协同演化的生成对抗网络为生成式零样本图像分类提供理论基础	第三章
4	<b>Shiming Chen</b> , Ziming Hong, Guo-Sen Xie, Wenhan Yang, Qinmu Peng, Kai Wang, Jian Zhao and Xinge You. MSDN: Mutually Semantic Distillation Network for Zero-Shot Learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition ( <b>CVPR</b> ), 2022.	第一作者; CCF-A类会议	提出一种基于互语义蒸馏网络的零样本图像分类挖掘视觉-属性特征的关键公共语义知识	第四章

# 华 中 科 技 大 学 博 士 学 位 论 文

---

5	<b>Shiming Chen</b> , Ziming Hong, Yang Liu, Baigui Sun, Guo-Sen Xie, Xinge You and Ke Lv. TransZero: Attribute-guided Transformer for Zero-Shot Learning. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2022.	第一作者; CCF-A类会议	提出一种基于属性指导Transformer的零样本图像分类挖掘视觉-属性特征的关键公共语义知识	第四章
6	<b>Shiming Chen</b> , Ziming Hong, Guo-Sen Xie, Xinge You, Weiping Ding and Ling Shao. GNDAN: Graph Navigated Dual Attention Network for Zero-Shot Learning. IEEE Transactions on Neural Networks and Learning Systems, 2022.	第一作者; SCI源刊; 中科院大类一区	提出一种基于图指导双注意力网络的嵌入式零样本图像分类方法解决跨数据集偏差问题	第三章
7	Ziming Hong*, <b>Shiming Chen*</b> <sup>#</sup> , Guo-Sen Xie, Yuanjie Shao, Qinmu Peng, Jian Zhao, Xinge You. Semantic Compression Embedding for Zero-Shot Learning. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2022.	共同第一作者; 通讯作者; CCF-A类会议	提出一种基于前置视觉特征增强的生成式零样本图像分类方法解决跨数据集偏差问题	第三章
8	<b>Shiming Chen</b> , Ziming Hong, Guo-Sen Xie, Jian Yang, Hao Li, Xinge You, Shuicheng Yan and Ling Shao. TransZero++: Cross Attribute-Guided Transformer for Zero-Shot Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.	第一作者; SCI源刊; 中科院大类一区; CCF-A类期刊; Minor Revision	联合属性指导的Transformer和互语义蒸馏网络挖掘视觉-属性特征的关键公共语义知识	第四章
9	<b>Shiming Chen</b> , Ziming Hong, Guo-Sen Xie, Yibing Song, Jian Yang, Xinge You, Shuicheng Yan and Ling Shao. ViFR: Visual Feature Refinement for Zero-Shot Learning. International Journal of Computer Vision, 2022.	第一作者; SCI源刊; 中科院大类二区; CCF-A类期刊; Under Review	联合前置和后置视觉特征精细化模块学习增强的视觉特征以解决跨数据集偏差问题	第三章

# 华 中 科 技 大 学 博 士 学 位 论 文

---

10	<p><b>Shiming Chen</b>, Wenjin Hou, Ziming Hong, Yibing Song, Tongliang Liu, Xinge You and Kun Zhang. DSP: Dynamic Semantic Prototype for Zero-Shot Learning. International Conference on Learning Representations (<b>ICLR</b>), 2023.</p>	<p>第一作者: THU A类会议; Submitted</p>	<p>增强生成模型中视觉特征的语义信息解决跨数据集偏差问题</p>	<p>第三章</p>
----	---	--	-----------------------------------	------------

## 附录 4 攻读博士学位期间参与的科研项目

### 1. 国家重点研发计划

项目名称: 出入境证件智能化真伪识别装备研究

项目编号: No. 2022YFC3301004

起止时间: 2022 年 10 月至 2025 年 9 月

担任角色: 项目参与者, 负责基于零样本学习的跨国证件识别算法设计

### 2. 国家自然科学基金

项目名称: 基于社会活动场的多混合智能体轨迹预测研究

项目编号: No. 62172177

起止时间: 2022 年 1 月至 2025 年 12 月

担任角色: 项目参与者, 负责基于零样本学习的轨迹预测新场景认知算法设计

### 3. 湖北省重点研发计划

项目名称: 基于颈动脉易损斑块建设与脑梗风险预测

项目编号: No. 2021CFB332

起止时间: 2021 年 9 月至 2024 年 8 月

担任角色: 项目参与者, 负责基于知识迁移的颈动脉斑块识别算法设计