E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP751　　**文献标识码**: A　　**文章编号**: 1006-8961(2026)03-0826-14

论文引用格式:Tian Y, Qian Y X, Huang Q B, Chen J Y, Zhong L and Wu X R. 2026. Image-text decoupling method for compositional zero-shot learning. Journal of Image and Graphics, 31(3):0826-0839(田弋, 钱毅鑫, 黄清宝, 陈佳岳, 钟磊, 伍贤瑞. 2026. 面向组合零样本识别的图文解耦合方法. 中国图象图形学报, 31(3):0826-0839)[DOI:10.11834/jig.250189]

面向组合零样本识别的图文解耦合方法

田弋[1],钱毅鑫[1],黄清宝[1],陈佳岳[1],钟磊[2*],伍贤瑞[2]

1. 广西大学电气工程学院,南宁　530004; 2. 中国移动通信集团广西有限公司,南宁　530000

摘　要:**目的**　组合零样本识别是计算机视觉领域零样本学习任务的子任务,旨在从已经见过的组合图像中学习属性和物体概念,并将其迁移到未见过的组合上。现有方法对组合图像中属性和物体的解耦合能力不足,并且未能充分发挥文本标签对于属性和物体信息的解耦合作用。**方法**　为解决组合图像中属性与物体信息纠缠的问题,针对文本与视觉模态的差异,提出双模态解耦机制:在文本端构建图神经网络以建模属性与物体间的语义关系,在视觉端引入交叉注意力机制增强对属性和物体特征的分离能力。该方法集成于语言图像预训练框架中,从语言与视觉两个层面提升模型对属性与物体概念的建模能力,从而增强未见组合的识别效果。**结果**　在3个主流的组合零样本识别基准数据集MIT-States、UT-Zappos和C-GQA(compositional GQA)上对所提方法进行了系统评估,结果表明模型性能显著提升。以MIT-States数据集为例,在封闭世界设置下,相较于性能排名第2的模型,本文方法的AUC(area under curve)提升3.3%,HM(Harmonic mean)提升2.4%,已见组合的识别准确率提升5.3%,未见组合提升1.0%;在开放世界设置下,本文方法的AUC提升0.9%,HM提升0.7%,已见组合与未见组合准确率分别提升3.2%和1.0%。此外,在MIT-States数据集上对提出的文本与视觉解耦模块及其上下文建模组件进行了消融实验,进一步验证了各子模块对整体性能的有效贡献。**结论**　所提出的图文双端解耦合模块提升了模型对于组合中属性和物体的学习能力,显著提升了模型在组合零样本识别任务上的表现。

关键词:零样本学习(ZSL);组合零样本识别(CZSL);解耦合;图卷积网络(GCN);交叉注意力;对比语言图像预训练(CLIP)

Image-text decoupling method for compositional zero-shot learning

Tian Yi[1], Qian Yixin[1], Huang Qingbao[1], Chen Jiayue[1], Zhong Lei[2*], Wu Xianrui[2]

1. School of Electrical Engineering, Guangxi University, Nanning　530004, China;

2. China Mobile Communications Group Guangxi Co., Ltd., Nanning　530000, China

Abstract: **Objective** Compositional zero-shot learning (CZSL) is a subtask of zero-shot learning (ZSL) in the field of computer vision, aiming to learn attribute and object concepts from seen composition images and transfer them to unseen compositions. For example, learning the attribute "cute" and the object "cat" from the composition images of "cute dog" and "nimble cat" to recognize the unseen composition of "cute cat". Existing methods are unable to decouple attributes and objects in compositional images and have not fully utilized the decoupling role of text labels in attribute and object information. In previous research on compositions, attributes and objects were often treated with the assumption that the object is the main subject, with attributes attached to the object, lacking recognition of the equal importance of attributes and

收稿日期:2025-05-08;**修回日期**:2025-07-31;**预印本日期**:2025-08-07

*通信作者:钟磊　15277166966@139.com

基金项目:广西自然科学基金重点项目(2025GXNSFDA069017);国家自然科学基金项目(62276072);广西八桂学者项目(2025)

Supported by: Guangxi Natural Science Foundation Key Project (2025GXNSFDA069017); National Natural Science Foundation of China (62276072)

objects. Although some studies recognize the importance of decoupling in the compositional zero-shot recognition task, they still focus on decoupling at the superficial level or fail to fully decouple attributes and objects. **Method** Decoupling the conceptual features of attributes and objects from images is both a prerequisite and a challenge for enabling the model to recognize unseen compositions. By simulating the interactions of attribute and object elements within compositions and between compositions and considering the differences in textual and visual modalities, we have designed a text decoupler based on graph neural networks and a visual decoupler based on cross-attention contrastive learning. These components are integrated into contrastive language-image pre-training (CLIP) to achieve the decoupling of attribute and object elements at the linguistic and visual levels. More specifically, we constructed a compositional graph network by using the text labels in the data and modeled the relationships between attributes, objects, and compositions within the graph by using graph convolutional networks. This approach allows the model to understand the corresponding relationships of the same labels in different composition texts through the compositional graph, further assisting the model in decoupling at the visual level. This representation at the text level helps the model understand the relationships between attributes and objects within the same composition, as well as the inherent connections between the same attributes and objects across different compositions. At the image level, for the composition image to be recognized, we search for an image in the dataset that shares the same attribute and another image that shares the same object. The image with the same attribute is then passed into the cross-attention module along with the composition image, allowing the model to learn the attribute through the attention layer. Applying the same operation on the object side, attention-level contrastive learning with visually similar images improves recognition of visual elements. In this way, we achieve the decoupling of attribute and object information in the image at the visual level. These two components are inserted into the final layers of CLIP's text encoder and image encoder, enhancing CLIP's decoupling ability while fully utilizing its powerful image-text representation and retrieval recognition capabilities. **Result** Evaluations on three well-known CZSL benchmark datasets, namely, MIT-States, UT-Zappos, and C-GQA, demonstrate that the proposed method significantly improves model performance. On the MIT-States dataset, compared with the second-best performing model, the proposed method achieved a 3.3% increase in area under curve (AUC), a 2.4% increase in HM, a 5.3% improvement in the recognition accuracy for seen compositions, and a 1.0% improvement for unseen compositions in the closed-world setting. In the open-world setting, AUC increased by 0.9%, HM improved by 0.7%, seen composition recognition accuracy improved by 3.2%, and unseen composition recognition accuracy improved by 1.0%. On the UT-Zappos dataset, in the closed-world setting, compared with the second-best performing model, the proposed method achieved a 0.5% increase in AUC, a 1.3% improvement in the recognition accuracy for seen compositions, a 0.1% increase for unseen compositions, and the second-best performance on HM. In the open-world setting, AUC increased by 2.5%, HM improved by 0.9%, seen composition recognition accuracy achieved the same performance as the state-of-the-art method, and unseen composition recognition accuracy improved by 2.6%. Our model achieved the second-best performance on the C-GQA dataset. Additionally, ablation studies on the text-visual decoupling module and its contextual components on the MIT-States dataset confirm the effectiveness of the proposed method. Ablation experiments on the MIT-States dataset show that removing the visual decoupling module resulted in a 9.8% decrease in AUC, an 8.9% decrease in HM, an 11.0% decrease in recognition accuracy for seen compositions, and a 9.0% decrease in recognition accuracy for unseen compositions. Removing the text decoupling module resulted in a 2.5% decrease in AUC, a 0.9% decrease in HM, a 3.7% decrease in recognition accuracy for seen compositions, and a 0.5% increase in recognition accuracy for unseen compositions. This paper further conducts ablation experiments on the contextual relationships within the image decoupling component and the text decoupling component. For the text decoupling component, we eliminate the contextual relationships by retaining only the trainable tokens related to the text prompt and removing the compositional graph information modeled by GCN. For the image decoupling component, we eliminate the contextual relationships by removing the images with matching attributes and objects, replacing them with the training images themselves. This paper conducts qualitative experimental analysis, including retrieving text through unseen images, retrieving images through text, and retrieving visual elements. **Conclusion** The proposed text-visual decoupling module enhances the model's ability to learn attributes and objects in compositions, significantly improving the model's performance in compositional zero-shot recognition tasks. Ablation experiments demonstrate that models for the compositional zero-shot learning task should fully utilize the contextual information from both images and texts in the training samples, providing a direction for future research.

Key words: zero-shot learning (ZSL); compositional zero-shot learning (CZSL); decouple; graph convolutional network (GCN); cross-attention; contrastive language image pre-training (CLIP)

0 引 言

当人们看到图像“黄色叶子”和“黄色船”时，可以轻松地识别出它们具有相同的属性“黄色”。人们也可以识别出物体“猫”，即使它们是不同的品种。当看到未见的组合“黄色的猫”时，能够识别出它具有属性“黄色”，并且可以看出它是一个叫做“猫”的物体。使模型获得这种“组合”能力是一个具有挑战性的任务。组合零样本学习（compositional zero-shot learning，CZSL）旨在通过利用视觉原语（属性和物体）识别新的属性—物体组合，近年来引起了大量研究者的兴趣（Misra等，2017）。这种识别组合的能力不仅可以增强深度学习模型处理少量甚至无标签样本的能力，而且可以推动多模态学习研究的发展（Chen等，2023）。

在早期的组合零样本识别研究中，研究者认为组合中的物体是识别的基点，将属性信息作为附加判断类处理。Nagarajan和Grauman（2018）将属性建模为“操作符”，即将属性视为能够转换物体表示的操作，具体而言，属性被编码为一种变换方式，当应用于物体编码时，能够适当地改变其外观。Nagarajan和Grauman（2018）的工作将属性视为一个独立的标签而忽略了属性与物体之间的内在关系。Li等人（2020）借鉴群论中的对称性原理，提出了SymNet（symmetry and group in attribute-object compositions）框架，将属性和物体组合建模为一个群的操作来学习属性物体组合的对称性特征，在识别时，他们提出使用相对移动距离（relative moving distance，RMD）为给定属性在潜在空间中对物体进行属性添加和移除操作。Purushwalkam等人（2019）则认为这种将属性视做物体的附加的做法忽视了属性与物体之间的复杂关系，他们通过提出任务驱动的模块化网络，将网络划分为多个小型模块，每个模块负责处理特定的子任务。通过任务描述符对模块进行门控，模型可以在测试时动态组合这些模块，以适应新的任务。并且在特征提取阶段，模型不仅考虑输入图像，还结合属性—对象对的信息，以提取与任务相关的特征。Mancini等人（2021）提出了余弦相似度来直接计算组合的向量表示与图像特征之间的相似性，并提出了开放世界组合零样本学习，他们利用属性与物体的联合嵌入表示计算两者之间的可行性分数，来解决开放世界组合零样本识别中搜索空间过大的问题，低于可行性分数的组合标签会被模型自动过滤。

近年来，零样本学习和小样本学习领域已经涌现出大量工作，从视觉—语义映射、度量学习和跨域泛化等角度提升模型识别未见类别的能力（韩阿友等，2023；陶鹏 等，2024；王梓祺 等，2024；文浪 等，2025；余悦 等，2025；张桂梅 等，2025）。在此基础上，研究者进一步关注组合泛化问题，提出了多种面向组合零样本学习及其开放世界设定的方法（Anwaar等，2022；Hu 和 Wang，2023；Li等，2023）。尽管这些研究为视觉语义理解奠定了重要基础，但现有方法仍主要关注类别级或属性级的泛化能力，对图像中属性与物体的细粒度组合关系建模不足，因此难以从根本上解决CZSL中属性—物体解耦与未见组合识别的问题。

然而，近年来研究者认识到，在组合零样本学习任务中，图像样本的属性信息和物体信息这些视觉元素总是以纠缠的状态出现，能解开属性和物体的耦合是提升组合零样本学习模型性能的关键。同时研究者们发现组合零样本学习数据集中的图像存在相同标签的属性和物体在不同图像中表现形式完全不一致的问题，如图1所示，相同的属性“空的（empty）”在描述提包和公交车时表现完全不一致，这使得模型在学习这些视觉元素时很难学习到某一标签的固定表现形式，也使得研究者转向解开图像中属性和物体信息的耦合状态的研究中。Atzmon等人（2020）通过因果建模和嵌入学习来解开属性与物体之间耦合状态，他们认为属性与物体之间存在因果联系，通过因果嵌入模型解耦属性和物体的信息，学习属性和物体的潜在空间表示。Yang等人（2022）将组合图像的视觉特征嵌入到一个孪生对比空间中，分别捕捉属性和物体的原型表征，然后通过增加训练组合的多样性，提升模型对未见组合的鲁棒性。Hao等人（2023）尝试将一种基于推土机距离（earth move distance，EMD）规约函数加入到对比学习方法中，以提升模型对属性信息和物体信息的区

分能力，达到更好的解耦合效果。但是这些研究都没有充分利用到文本标签对于视觉元素解耦合的辅助作用，仅在视觉层面进行解耦合研究。

本文探究了在文本和图像模态中对组合图像中的属性和物体视觉的解耦合方式。本文采用对比语言图像预训练（contrastive language image pretraining，CLIP）（Radford 等，2021）模型作为模型骨干的网络。由于图像和文本之间存在模态差异，在 CLIP 的文本端和图像端针对两个模态数据的特点分别添加了不同的解耦组件，以增强 CLIP 对组合图像的解耦能力。对于文本端，将组合、属性和物体一起作为组合图的节点输入到图神经网络（graph neural network，GNN）（Wu 等，2021）中，通过 GNN 的消息传递框架捕获组合特征以及独立属性和物体的特征，从而实现属性和物体在文本端的解耦；对于图像端，将具有相同属性和物体的两幅图像与组合图像一起输入到交叉注意力模块（cross-attention）（Vaswani 等，2017）中，让模型通过交叉注意力分别注意到两幅图像中的相似部分，更加有针对性地学习属性和物体元素的特征，从而在视觉层面实现属性和物体的解耦。本文将这两个模块分别添加到 CLIP 的文本编码器和图像编码器中，增强了 CLIP 对属性和物体的理解能力，同时保持了 CLIP 在图像—文本对的强大编码能力。

图 1　相同属性的不同表达

Fig. 1　Different expressions of the same attribute

1　方法介绍

1.1　任务描述

组合零样本学习的目标是通过利用已见组合中的属性和物体，识别未见的组合。给定一个组合，其中，$\boldsymbol{C}=(\boldsymbol{C}^S,\ \boldsymbol{C}^U)$，$\boldsymbol{C}^S$ 和 $\boldsymbol{C}^U$ 分别表示已见组合和未见组合，且 $\boldsymbol{C}^S \cap \boldsymbol{C}^U=\varnothing$（即已见和未见组合没有交集）。对于任何一个组合 ，由属性 $\boldsymbol{a}$ 和物体 $\boldsymbol{o}$ 组成。$\boldsymbol{C}^S$ 和 $\boldsymbol{C}^U$ 包含了用于训练和测试的所有属性和物体。在训练过程中，整个 $\boldsymbol{C}^S$ 中的组合将被提供给模型，但并不是 $\boldsymbol{C}^S$ 中所有的图像都将用于训练，部分 $\boldsymbol{C}^S$ 中的图像会被用于测试阶段衡量模型对已经见过的组合的识别能力。而在未见组合 $\boldsymbol{C}^U$ 中的组合 $\boldsymbol{c}$ 的所有图像不会提供给模型。在测试阶段，测试图像将来自于 $\boldsymbol{C}^S$ 和 $\boldsymbol{C}^U$ 的组合。这样可以在测试过程中确保模型已经学习过了测试集中的所有属性和物体元素，但由这些属性和物体组成的组合模型可能没有见过。组合零样本学习任务包括两种设置：封闭世界（closed-world）和开放世界（open-world）。

在封闭世界的组合零样本学习测试中，提供给模型的组合文本标签来自于 $\boldsymbol{C}^S \cup \boldsymbol{C}^U$。而在开放世界的组合零样本学习测试中，测试的组合可能没有对应的图像（例如，“方形的树”）。组合零样本学习模型的目标可以总结为学习一个模型 $\boldsymbol{f}: \boldsymbol{I} \rightarrow \boldsymbol{C}$，用于预测给定图像 $\boldsymbol{I}$ 的组合标签 $\boldsymbol{C}$。

给定输入图像 $\boldsymbol{i}$ 和组合 $\boldsymbol{c}$，组合零样本学习模型应该计算图像和组合文本之间的兼容性得分，即

$$\boldsymbol{f}\left(\boldsymbol{i},\left(\boldsymbol{c}_a,\boldsymbol{c}_o\right)\right)=\operatorname{argmax}\left(score\left(\boldsymbol{\phi}(\boldsymbol{i}),\boldsymbol{\varphi}\left(\boldsymbol{c}_a,\boldsymbol{c}_o\right)\right)\right) \tag{1}$$

式中，$\boldsymbol{\phi}(\cdot)$ 表示视觉编码器，$\boldsymbol{\varphi}(\cdot)$ 表示文本编码器，$score(\cdot)$ 是相似性函数。模型 $\boldsymbol{f}$ 通过学习模型参数 $\boldsymbol{\theta}_{\phi}$ 和 $\boldsymbol{\theta}_{\varphi}$ 更新各项参数，优化目标是为输入图像 $\boldsymbol{i}$ 匹配到最相似的组合文本。

1.2　模型总览

本文为组合零样本学习任务提出了一种图形语义解耦合模型，用于有效地解开训练对中属性和物体的纠缠状态。基于文本和图像模态的不同特性，本文设计了不同的解耦方法模块，并将两种解耦模块与 CLIP 结合，赋予了 CLIP 模型更强的解耦合能力，提升了 CLIP 模型在组合零样本学习任务上的表现。如图 2 所示，将解耦组件插入到冻结的 CLIP 编码器中，每个组件的输入是 CLIP 一个计算单元的输出，组件的输出是 CLIP 下一个计算单元的输入。图像解耦组件由两个交叉注意力模块组成，通过对属性信息和物体信息分开进行交叉注意力计算，模型可以分别学习到属性和物体的高维向量特征表示。

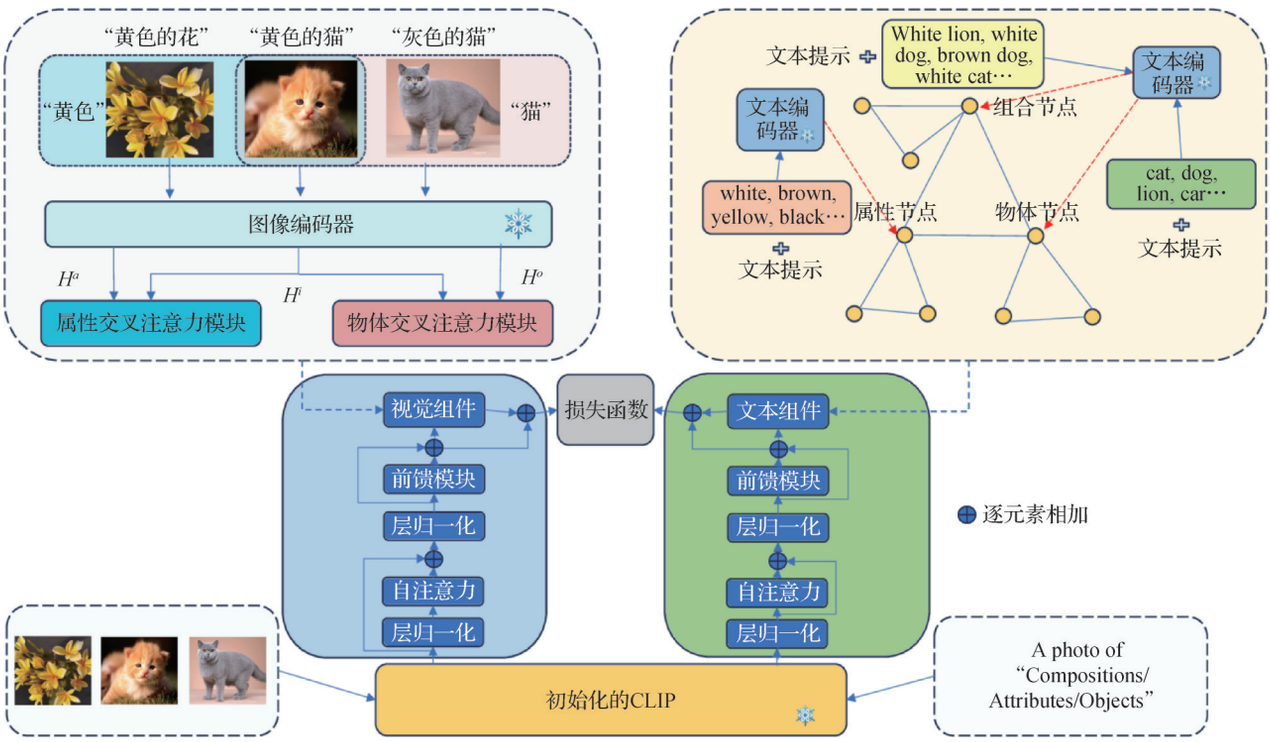

图2　模型总览

Fig. 2　Overview of proposed method

本文将属性、物体和组合文本联合起来建立了一个组合图，用于表示属性、物体和组合之间的全局关系，再使用一个图卷积神经网络（graph convolutional network，GCN）（Kipf和Welling，2017）进行特征层面的分离和聚合，从而实现对文本原始特征的上下文感知解耦和重组。

1.3　文本端解耦合组件

对于文本信息，本文将训练集中的每一个标签，包含组合、属性和物体，视为组合图中的3个节点，可以表示为

$$G = |\boldsymbol{A}| \cdot |\boldsymbol{O}| \cdot |\boldsymbol{C}| \cdot (\boldsymbol{A}) \tag{2}$$

式中，·表示节点之间有边连接，$|\boldsymbol{C}| \cdot (\boldsymbol{A})$表示该组合图与其他的组合图共用一个属性节点，意味着这两个训练文本的属性一致。测试集样本保留相同的设计，以确保模型能够推广到未见过的组合。更具体地说，对于所有文本样本$\boldsymbol{c}, \boldsymbol{a}, \boldsymbol{o}$，本文模型保持3个无向无权重的边$(\boldsymbol{a}, \boldsymbol{o})$，$(\boldsymbol{a}, \boldsymbol{c})$，$(\boldsymbol{o}, \boldsymbol{c})$，来构建组合图。这些图的属性和物体节点具有邻接结构。$\boldsymbol{A}, \boldsymbol{O}, \boldsymbol{C}$分别表示属性、物体和组合集合。初始化的节点特征是属性、物体和组合的语义表示。在本文中，它们由CLIP的文本编码器编码的文本提示联合嵌入表征。本文模型为属性、物体和组合标签分别添加文本提示（Nayak等，2023）：“a photo of［属性］”、“a photo of［物体］”和“a photo of［属性］［物体］”。本文模型将这3个提示输入同一个CLIP的文本编码器，得到3个隐藏状态$\boldsymbol{a}^H, \boldsymbol{o}^H, \boldsymbol{c}^H$，然后提取CLIP输出结果的最后一个令牌［EOT］，记做$\boldsymbol{a}^h, \boldsymbol{o}^h, \boldsymbol{c}^h$，作为图节点的输入特征。然后，利用多个GNN层在图节点之间传播特征。本文使用的GNN层是GCN模型，图的结构可以通过邻接矩阵$\boldsymbol{M}$表示，其中，如果节点$\boldsymbol{i}$和节点$\boldsymbol{j}$之间存在连接，则$\boldsymbol{M}_{i,j} = 1$，否则$\boldsymbol{M}_{i,j} = 0$。每个节点将添加自连接，即$\boldsymbol{M}_{i,i} = 1$，因此，每个查询通过相邻节点和自身聚合信息。例如，一个属性查询可以聚合邻接物体查询$\boldsymbol{\beta}(\boldsymbol{q}_o)$的信息以及自身$\boldsymbol{q}_a$的信息，即

$$\boldsymbol{q}^{\text{agree}} = \frac{\boldsymbol{q}_a + \sum_{\boldsymbol{q}_o \in N(\boldsymbol{q}_o)} \boldsymbol{q}_o}{\left|N(\boldsymbol{q}_o)\right| + 1} \tag{3}$$

式中，$\boldsymbol{q}^{\text{agree}}$表示聚合信息后的属性查询，为该属性查询添加权重$\boldsymbol{W}$和偏置$\boldsymbol{B}$，得到更新后的属性查询$\boldsymbol{q}'_a$，$N(\boldsymbol{q}_o)$表示节点$\boldsymbol{q}_o$的邻接节点集合。具体为

$$\boldsymbol{q}'_a = \theta\left(\boldsymbol{W} \cdot \boldsymbol{q}^{\text{agree}} + \boldsymbol{B}\right) \tag{4}$$

式中，θ是激活函数。此外，本文为GCN添加了一个跳跃连接，并且将GCN的深度设置为2，最后，基于GCN的关系推理模块的输出$\boldsymbol{q}_{a,o}^{\text{out}}$，可以表示为

$$\boldsymbol{q}_{a,o}^{\text{out}} = \boldsymbol{q}_{a,o} + GCN\left(\boldsymbol{q}_{a,o}\right) \tag{5}$$

式中，$\boldsymbol{q}_{a,o}$是属性和物体的查询集合。该文本端解耦合组件接入到CLIP的文本编码器之后，用以增强CLIP的文本端的解耦合能力。

1.4 图像端解耦合组件

本文利用多个原始共享输入的交叉注意力，实现了视觉原始特征的上下文感知解耦和重组。对于输入图像$\boldsymbol{i}$，模型会准备一幅具有和$\boldsymbol{i}$属性元素相同但是物体元素不同的图像$\boldsymbol{i}_a$和一幅与$\boldsymbol{i}$物体元素相同但是属性元素不同的图像$\boldsymbol{i}_o$。如图2所示，当模型识别组合“橘黄色的猫”时，模型从训练集中挑选一幅“黄色的花”的图像作为$\boldsymbol{i}_a$和一幅标签为“灰色的猫”的图像作为$\boldsymbol{i}_o$。这3幅图像将被输入冻结的CLIP图像编码器中提取特征后，得到隐藏状态$\boldsymbol{H}_a, \boldsymbol{H}_c, \boldsymbol{H}_o$来进行交叉注意力计算。由于需要在不同图像之间交换信息，当模型计算图像$\boldsymbol{i}$和$\boldsymbol{i}_a$之间的交叉注意力时，需要将$\boldsymbol{H}^a, \boldsymbol{H}^i, \boldsymbol{H}^i$分别设置为属性交叉注意力模块的查询向量、键向量和值向量输入，整体交叉注意力计算过程可以表示为

$$f_{\text{CrossAttention}}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right) = f_{\text{softmax}}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d'}}\right)\boldsymbol{V} \tag{6}$$

$$\boldsymbol{Q} = \boldsymbol{W}_Q \cdot \boldsymbol{H}^a, \boldsymbol{K} = \boldsymbol{W}_K \cdot \boldsymbol{H}^i, \boldsymbol{V} = \boldsymbol{W}_V \cdot \boldsymbol{H}^i \tag{7}$$

式中，$\boldsymbol{W}_Q$、$\boldsymbol{W}_K$、$\boldsymbol{W}_V$是3个线性变换矩阵，通常在自注意力中用于灵活计算d'为Q和K的特征维度。本文模型还在交叉注意力模块后添加了一个前馈层，以增强模型的特征表示能力。可以看出，每个输出令牌是值令牌的加权和，加权和的权重是通过查询向量和键向量的相似度计算得到的。以属性解耦合为例，通过将查询向量设置为$\boldsymbol{H}^a$，模型可以将$\boldsymbol{H}^i$细化，以保留图像$\boldsymbol{i}$的特定属性特征，并保留在不同组合中涉及图像$\boldsymbol{i}$的一般特征，还交换查询向量和值向量、交换查询向量和键向量，以细化$\boldsymbol{H}^i$的特征表示。因此，交叉注意力的输出可以表示为$\boldsymbol{H}_a^a$和$\boldsymbol{H}_a^i$。从配对的物体共享图像中提取的物体特定特征$\boldsymbol{H}_o^o$和$\boldsymbol{H}_o^i$。通过这种方式，本文模型利用属性共享和物体共享邻接图像来解耦输入图像的属性和物体特征。最后，模型将这两个解耦的原始特征与原始输入特征重新组合在一起，得到

$$\hat{\boldsymbol{H}}^i = \boldsymbol{H}_a^i + \boldsymbol{H}_o^i + \boldsymbol{H}^i \tag{8}$$

最终得到输入图像的特征输出，用于解耦合属性信息和物体信息的交叉注意力模块参数不共享，以便将属性元素和物体元素的原始特征投影到不同的特征子空间。

1.5 训练和推理

训练过程中，在文本编码器和图像编码器的最后一层注意力层中，本文模型通过跳跃连接获得文本组件和图像组件的输出。为了便于表达，将文本编码器的输出表示为$\hat{\boldsymbol{H}}^t$。模型提取组件的输出的最后一个[EOT]令牌的表示$\boldsymbol{h}^i$和$\boldsymbol{h}^t$用于衡量视觉特征和文本特征的兼容性。具体为

$$\boldsymbol{h}^i = \hat{\boldsymbol{H}}^i_{[\text{EOT}]} \tag{9}$$

对于输入的待学习图像$\boldsymbol{k}$和候选组$\boldsymbol{c}_k = \left(\boldsymbol{a}_k, \boldsymbol{o}_k\right)$，模型通过下式计算兼容性得分，具体为

$$score\left(\boldsymbol{k}, \boldsymbol{c}_k\right) = \alpha\left[\boldsymbol{H}_c^t \cdot \boldsymbol{H}^i\right] + \beta\left[\boldsymbol{H}_a^i \cdot \boldsymbol{H}^t\right] + \gamma\left[\boldsymbol{H}_o^i \cdot \boldsymbol{H}^t\right] \tag{10}$$

模型通过3个可学习的参数α、β和γ来平衡训练时对于图像中的属性元素和物体元素归纳假设时（判断图像中的属性元素和物体元素属于哪一类）的学习能力。在计算相似度时所有的向量表示都进行了归一化处理。最后，通过最小化训练集$\mathcal{D}_{\text{tr}}$上的交叉熵损失来优化模型对于这些可学习的令牌嵌入表示，训练集中的组合都来自$\boldsymbol{C}_s$，即

$$L = -\frac{1}{\left|\mathcal{D}_{\text{tr}}\right|}\sum_{\left(\boldsymbol{k},\boldsymbol{c}_k\right)\in\mathcal{D}_{\text{tr}}}\log\frac{\exp\left(\dfrac{s\left(\boldsymbol{k},\boldsymbol{c}_k=\left(\boldsymbol{a}_k,\boldsymbol{o}_k\right)\right)}{\tau}\right)}{\displaystyle\sum_{\boldsymbol{c}_k\in\boldsymbol{C}_s}\exp\left(\frac{s\left(\boldsymbol{k},\boldsymbol{c}_j=\left(\boldsymbol{a}_j,\boldsymbol{o}_j\right)\right)}{\tau}\right)} \tag{11}$$

式中，τ是以CLIP作为骨干模型的研究中广泛使用的温度参数，$s\left(\boldsymbol{k},\boldsymbol{c}_k=\left(\boldsymbol{a}_k,\boldsymbol{o}_k\right)\right)$的期望是真实标签。

在推理阶段，对于每个测试图像$\boldsymbol{i}_t$，模型会使用式(10)估算$\boldsymbol{i}_t$和来自组合集$\boldsymbol{C}^S \cup \boldsymbol{C}^U$中每个候选组合$\boldsymbol{c}$之间的兼容性得分，由于测试时模型不知道测试图像的属性和物体标签，模型会将测试图像本身作为测试时的属性信息共享图像和物体信息共享图像，用于计算$\boldsymbol{H}_a^i$和$\boldsymbol{H}_o^i$，测试时，参数α、β和γ不更新。

2 实验分析

2.1 数据集介绍

本文在3个广泛使用的组合零样本识别任务数

据集上进行实验验证，3个数据集分别是MIT-States、UT-Zappos以及C-GQA（compositional graph question answering）。数据集的基本信息如表1所示，其中，A表示数据集中属性的数量，O表示物体的数量，C_s表示数据集中的可见组合的数量，C_u表示数据集中不可见组合的数量，I表示数据集中图像的数量。本文使用Purushwalkam等人（2019）的数据集划分设置。

Isola等人（2015）通过早期的图像搜索引擎，构建了MIT-States数据集。他们构建了一个三元组（形容词，反义词，名词），然后通过搜索引擎收集图像，其中查询词是（反义词，名词）和（形容词，名词）组合。为了减少结果中的噪声，他们要求标注人员对数据进行了清理。在MIT-States中，数据集中的大多数属性是用来描述物体状态的形容词，例如“拼接的”和“腐烂的”，而其他数据集的属性信息则更多地描述物体的颜色、材质等。

Yu和Grauman（2014）构建了UT-Zappos50k数据集，该数据集包含各种鞋子，共有5万幅图像。鞋子根据材质和类型（形状）进行标注，其中鞋子材质被视为组合的属性，而鞋子的类型则作为组合中的物体。为了公平对比，本文进行的实验与其他工作一致，使用UT-Zappos子集进行模型的训练和推理工作。

Naeem等人（2021）从GQA（graph question answering）数据集中挑选并重新标注了C-GQA数据集，他们认为MIT-States数据集有一定的噪声干扰，而UT-Zappos数据集的属性和物体类不具有普遍性。C-GQA数据集的属性元素（例如，老旧、潮湿）和物体元素（例如，狗、公共汽车）更加常见，并且数据规模更大。

表1　数据集详情

Table 1　Details of datasets

数据集	组合信息		训练集		验证集		测试集	
	A	O	C_s	I	C_s/C_u	I	C_s/C_u	I
MIT-States	115	245	1 262	30 338	300/300	10 420	400/400	12 995
UT-Zappos	16	12	83	22 998	15/15	3 214	18/18	2 914
C-GQA	440	541	11 175	72 203	2 121/2 322	9 525	2 449/2 470	10 856

2.2　评估指标

本文按照广泛应用在组合零样本识别任务中的评价指标评估所提出的模型以及其他对比模型。在主要对比实验中，采用4个评估指标：曲线下区域面积（area under curve，AUC）、最优调和值（best harmonic mean，Best HM）、可见组合预测准确率（seen composition accuracy）和未见组合预测准确率（unseen composition accuracy）。由于组合零样本识别任务模型都基于见过的组合进行训练，在预测时模型具有预测偏见。具体而言，模型倾向于将组合图像辨认成见过的组合，在评估模型性能时应尽可能消除这种偏见的影响。Chao等人（2016）在广义零样本学习任务中提出了一种平衡可见类和未见类的评估指标，他们为未见类添加了一个校验偏置，让模型更加倾向于向未见类进行预测，通过调整校验偏置的值，可以很好地展现模型对抗偏置干扰的能力，体现模型的性能。Purushwalkam等人（2019）将这个指标引入到了组合零样本学习任务中，构造了一个可见—不可见类预测曲线（seen-unseen curve），如图3所示，图中曲线与两坐标轴的交点分别是最佳可见预测准确率（best seen）与最佳不可见预测准确率（best unseen），曲线下方与两坐标轴围出的区域面积就是AUC值，AUC值可以衡量模型预测可见组合与未见过组合的综合性能。通过计算可见准确率与不可见准确率最大值的调和平均数，可以获得Best HM这一指标（在实验中以HM代指）。为了表达方便，对于其他指标，本文用AUC指代曲线下区域面积这一指标，用Seen/Unseen指代可见组合预测准确率与未见组合预测准确率指标。

2.3　开放世界设置

本文在开放世界设置（Mancini等，2021）上评估了所提出的模型，这也是近期研究广泛评估的设置。与之相对的是封闭世界组合零样本识别设置，封闭世界的组合零样本识别设置中，模型需要考虑的文本标签只有具有相匹配图像的，而开放世界设置考虑了所有可能的组合，如，“长脚的蛇”这种现实中没有对应物体的组合标签会被用在开放世界组合零样本识别推理中。因此，开放世界设置的组合零样本

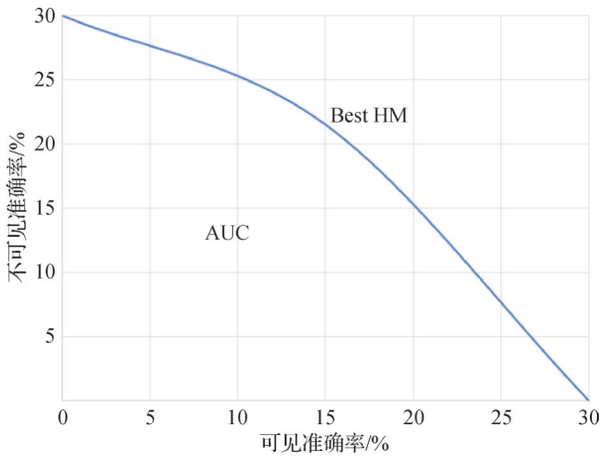

图 3　可见—不可见准确率曲线

Fig. 3　Seen-unseen accuracy curve

识别任务在推理过程中需要比封闭世界设置更大的测试空间。组合类别的数量显著增加，导致模型面临更严格的考察。在开放世界的 UT-Zappos 数据集中，可能的组合数量为 192，而在封闭世界中只有 36 个。需要注意的是，在测试期间，未见过的组合文本可能没有对应的图像。

2.4　超参数设置及实验环境

为了确保与其他基线模型进行公平对比，本文使用未经过微调的 CLIP-ViT-L/14 模型作为骨干模型，该模型包括 24 层的 ViT（vision Transformer）视觉编码器和 12 层的 Transformer 文本编码器，文本编码器和图像编码器中的隐藏状态分别为 768 维和 1 024 维。本文使用 GCN 实现 GNN 模块，设定 $K=2$，初始化的 α，β，γ 参数都设置为 1，与其他参数一起优化更新。使用 PyTorch 1.9.1 作为实验环境，优化器为 Adam 优化器，对于 MIT-States 和 C-GQA 数据集，模型的学习率设置为 5×10^{-5}，对于 UT-Zappos 数据集，模型的学习率设置为 1×10^{-5}，并且训练过程中学习率固定。所有实验在一张 NVIDIA 4090 显卡上进行，显存大小为 24 GB，每个数据集训练 40 轮，每轮每批次训练有 128 个数据样本。

2.5　实验结果与分析

本文的对比实验中，参与对比的基线模型包括非 CLIP 和 CLIP-based 两种，其中，非 CLIP 的模型有 SymNet（symmetry and group in attribute-object compositions）（Li 等，2020）、CGE（compositional cosine graph embeddings）（Naeem 等，2021）和 SCEN（siamese contrastive embedding network for compositional zero-shot learning）（Li 等，2022）；CLIP-based 的模型有 CLIP（Radford 等，2021）、CoOp（learning to prompt for vision-language models）（Zhou 等，2022）、CSP（learning to compose soft prompts for compositional zero-shot learning）（Nayak 等，2023）、HPL（hierarchical prompt learning for compositional zero-shot learning）（Wang 等，2023）、DFSP（decomposed soft prompt guided fusion enhancing for compositional zero-shot learning）（Lu 等，2023）、GIPCOL（graph-injected soft prompting for compositional zero-shot learning）（Xu 等，2024）和 CAILA（concept-aware intra-layer adapters for compositional zero-shot learning）（Zheng 等，2024）。

本文实验在封闭世界和开放世界两种实验设置下进行，其中，SCEN 未参与开放世界实验对比。实验结果如表 2 和表 3 所示。

本文模型在开放世界和封闭世界设置的 MIT-States 数据集上都取得了最优的表现；UT-Zappos 数据集在开放世界设置上取得最优性能，在封闭世界设置上除了 HM 也都达到了最优性能；在 C-GQA 数据集上，对于开放世界设置和封闭世界设置都取得了第 2 好的表现，在封闭世界设置上，在 AUC、HM、Seen 和 Unseen 指标上分别落后最优模型 2.2%、3.4%、2.7% 和 5.1%，对于开放世界设置的 C-GQA 数据集，本文模型在 AUC、HM、Seen 和 Unseen 这 4 个指标上分别落后最优模型 0.55%、0.6%、2.7% 和 0.3%。可能原因是本文提出模型对于属性标签的记忆域不够宽，并且对于相似属性的学习不够细粒度，导致模型对于 C-GQA 数据集大量的粗粒度属性标签识别能力不足。

2.6　消融实验

2.6.1　图文解耦合组件消融

为了验证提出的图文解耦合组件的有效性，本文通过系统地移除文本和图像组件评估它们各自对模型性能的贡献。具体而言，移除图像组件会使图像编码器退化为冻结图像编码器，生成每个输入图像的纠缠表示，这与大多数基于 CLIP 的基准方法保持一致；移除文本组件会使文本编码器表现为冻结文本编码器，缺乏跨组合图的特征传播能力，且不包含提示词中的可训练 token 嵌入。本文在封闭世界设置下对 MIT-States 数据集进行了消融实验，结果如表 4 所示。其中，完整模型用 Full-Model 表示，移除文本解耦合组件的模型用 w/o T-cpt 表示，移除视

表2　封闭世界实验结果

Table 2　Closed-world experimental results

/%

封闭世界		MIT-States				UT-Zappos				C-GQA			
类别	模型	Seen	Unseen	HM	AUC	Seen	Unseen	HM	AUC	Seen	Unseen	HM	AUC
非CLIP	SymNet	24.2	25.2	16.1	3.0	49.8	57.4	40.4	23.4	26.8	10.3	11.0	2.1
	CGE	32.8	28.0	21.4	6.5	64.5	71.5	**60.5**	33.5	33.5	15.5	16.0	4.2
	SCEN	29.9	25.2	18.4	5.3	63.5	63.1	47.8	32.0	28.9	25.4	17.5	5.5
CLIP-based	CLIP	30.2	46.0	26.1	11.0	15.8	49.1	15.6	5.0	7.5	25.0	8.6	1.4
	CoOp	34.4	47.6	29.8	13.5	52.1	49.3	34.6	18.8	20.5	26.8	17.1	4.4
	CSP	46.6	19.9	36.3	19.4	64.2	66.2	46.6	33.0	28.8	26.8	20.5	6.2
	HPL	47.5	50.6	37.3	20.2	63.0	68.8	48.2	35.0	30.8	28.4	22.4	7.2
	DFSP	46.9	52.0	37.3	20.6	66.7	71.7	47.2	36.0	38.2	32.0	27.1	10.5
	GIPCOL	48.5	49.6	36.6	19.9	65.0	68.5	48.8	36.2	31.9	28.4	22.5	7.1
	CAILA	51.0	53.9	39.9	23.4	67.8	74.0	57.0	44.1	**43.9**	**38.5**	**32.7**	**14.8**
	本文	**56.3**	**54.9**	**42.3**	**26.7**	**69.1**	**74.1**	58.1	**44.6**	41.2	33.4	29.3	12.6

注：加粗、下划线字体表示各列最优、次优结果。

表3　开放世界实验结果

Table 3　Open-world experimental results

/%

开放世界		MIT-States				UT-Zappos				C-GQA			
类别	模型	Seen	Unseen	HM	AUC	Seen	Unseen	HM	AUC	Seen	Unseen	HM	AUC
非CLIP	SymNet	21.4	7.0	5.8	0.8	53.3	44.6	34.5	18.5	26.7	2.2	3.3	0.43
	CGE	32.4	5.1	6.0	1.0	61.7	47.7	39.0	23.1	32.1	1.8	2.9	0.47
CLIP-based	CLIP	30.1	14.3	12.8	3.0	15.7	20.6	11.2	2.2	7.5	4.6	4.0	0.27
	CoOp	34.6	9.3	12.3	2.8	52.1	31.5	28.9	13.2	21.0	4.6	5.5	0.70
	CSP	46.3	15.7	17.4	5.7	64.1	44.1	38.9	22.7	28.7	5.2	6.9	1.20
	HPL	46.4	18.9	19.8	6.9	63.4	48.1	40.2	24.6	30.1	5.8	7.5	1.37
	DFSP	47.5	18.5	19.3	6.8	66.8	60.0	44.0	30.3	38.3	7.2	10.4	2.40
	GIPCOL	48.5	16.0	17.9	6.3	65.0	45.0	40.1	23.5	31.6	5.5	7.3	1.30
	CAILA	51.0	20.2	21.6	8.2	**67.8**	59.7	49.4	32.8	**43.9**	**8.0**	**11.5**	**3.08**
	本文	**54.2**	**21.2**	**22.3**	**9.1**	**67.8**	**62.3**	**50.3**	**35.3**	41.2	7.7	10.9	2.53

注：加粗、下划线字体表示各列最优、次优结果。

觉解耦合组件的模型用w/o V-cpt表示。实验结果表明，移除任一组件都会导致性能下降。对于AUC指标，移除文本和视觉组件分别导致2.5%和9.8%的性能下降；对于HM指标，移除文本和视觉组件分别导致0.9%和8.9%的性能下降。这证明了两个组件对模型都有积极贡献，且具有互补性。

进一步分析发现：1）移除图像组件造成的性能损失远大于移除文本组件；2）仅配备图像组件的模型性能已超越先前基于CLIP的基准模型，包括CAILA。这些发现验证了视觉原始特征相比文本原始特征具有更强的纠缠性。本文认为，尽管使用冻结的文本编码器，解耦的文本原始特征仍可通过独

立词汇有效捕获。然而，学习解耦的视觉原始特征对模型而言更为复杂，因此缺少对视觉信息的额外处理会导致更显著的性能损失。

表 4　图像/文本组件消融实验

Table 4　Image/text component ablation experiment

/%

模型	AUC	HM	Seen	Unseen
Full-Model	**26.7**	**42.3**	**56.3**	54.9
w/o T-cpt	24.2	41.4	52.6	**55.4**
w/o V-cpt	16.9	33.4	45.3	45.9

注：加粗字体表示各列最优结果。

2.6.2　上下文信息消融

本文进一步对图像解耦合组件和文本解耦合组件中的上下文关系建模进行消融分析。对于文本解耦合组件，通过仅保留与文本提示相关的可训练令牌，移除通过 GCN 建模的组合图信息来消除上下文关系；对于图像解耦合组件，通过将属性一致图像和物体一致图像替换为训练图像本身来消除上下文关系。在封闭世界设置的 MIT-States 数据集上的验证结果如表 5 所示。完整模型用 Full-Model 表示，移除文本解耦合组件中上下文关系的模型用 w/o T-con 表示，移除图像解耦合组件中上下文关系的模型用 w/o V-con 表示，移除所有上下文关系的模型用 w/o con 表示。结果显示，移除模型中的上下文关系表示模块均会导致性能下降，且视觉解耦合组件中的上下文关系建模对模型性能的影响更为显著。这进一步证明了上下文信息对于提升组合零样本识别性能的重要性。

表 5　上下文关系消融实验

Table 5　Contextual relationship ablation experiment

/%

模型	AUC	HM	Seen	Unseen
Full-Model	26.7	42.3	56.3	54.9
w/o T-con	25.7	42.2	54.1	55.2
w/o V-con	17.4	33.4	44.7	46.6
w/o con	16.9	34.0	44.3	46.4

2.7　定性实验分析

本文对文本到图像和图像到文本的检索进行了定性分析，以展示本文提出的模型在理解和组合图像概念方面的能力，同时设置了视觉元素检索实验来更加针对性地验证展示模型从图像中学习到的解耦后的属性和物体视觉元素。

2.7.1　图搜文实验效果

首先展示模型由图像检索文本的能力，本文展示了模型在 MIT-States 数据集上的检索实验结果。如图 4 所示，图中展示了 6 个不同的图像查询来检索文本的结果，模型为每个图像检索展示 5 个检索结果，图像的真实标签在上方展示，正确的预测结果用红色字体标出，模型为预测的结果进行排序，将模型认为最有可能正确的标签放在第 1 行，以此类推。可以看出，本文模型能够准确地找到查询图像的标签。图 4 中对“褶皱的棉布（ruffled cotton）”进行检索时，正确的结果被模型放在了第 3 位，但是该图像的物体属性不够标准，十分容易产生歧义，模型对于“褶皱”这一属性的认识基本是准确的，但是将“棉布（cotton）”错认为了是“包（bag）”，该错误并不足以对模型的检索识别能力产生质疑。

2.7.2　文搜图实验效果

图 5 展示了模型在 MIT-States 数据集上接收文本检索图像的能力。一共对 4 个文本查询进行检索，为每个查询寻找 5 幅匹配的图像。图 5 显示，本文模型对于文本查询“windblown tree（被风吹的树）”、“sliced pizza（切片的披萨）”、“tiny dog（小狗）”都能检索出完全正确的图像，证明了本文模型对于属性元素和物体元素进行了充分的理解和准确的判断，尤其是对于“windblown”这种抽象的属性概念，模型能够准确判断出这一概念应有的表现形式。但是对于查询“young cat”，模型检索出了 4 幅正确的图像和 1 幅错误的图像（已用红框标出），该图像的标签为“wet cat（湿的猫）”，图像中“wet”这一属性概念十分明显，但是被模型忽略。推测原因是“young”这一属性概念过于宽泛，对于猫这一物体概念来说，“young”这一属性概念并不是一个很浅显的特征，因此模型对这一概念的学习存在不足，并不能准确地推理判断。

2.7.3　视觉元素检索

Hao 等人（2023）提出了一个视觉概念检索实验，解决了之前的方法无法充分验证从图像中学习解耦视觉概念的问题。本文遵循他们的实验方法进行视觉元素检索实验，检索结果如图 4—图 6 所示。首先通过 CLIP 的视觉编码器从所有已见过的图像

图4　图搜文实验结果

Fig. 4　Image to text retrieval

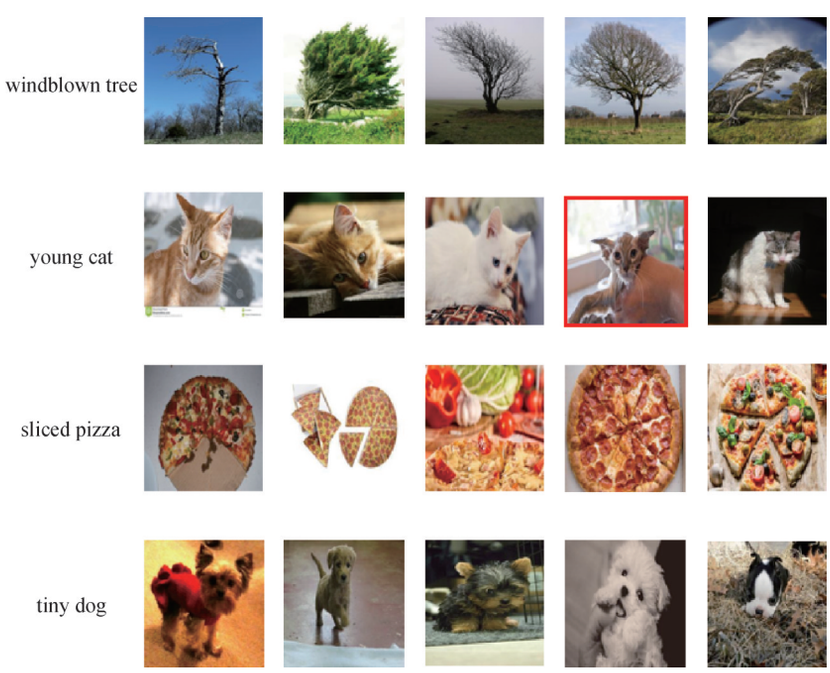

图5　文搜图检索结果

Fig. 5　Text to image retrieval

中提取属性特征和物体特征,然后给模型输入一幅未见图像,并通过测量与给定图像相似的属性特征和物体特征来分别检索前4个最接近的属性图像和物体图像。对于未见图像“有孔的盘子(pierced plate)”,图6在蓝线上方显示相同的“有孔的”物体,在虚线下方显示“盘子”图像。本文模型检索到了准确的“pierced + object”类图像:如“桶”、“柜子”、“杯子”等。而“属性 + 盘子”类图像则不包含任何“有孔的”元素。当输入“未成熟的橘子”图像时,本文模型仍然能正确地检索和分类具有相似属性和物体的图像,能够检索出“切片的橘子”这个破坏了原图像物体特征的图像。实验结果证明了本模型具备从组合图像中解耦属性和物体并完全理解它们的能力。

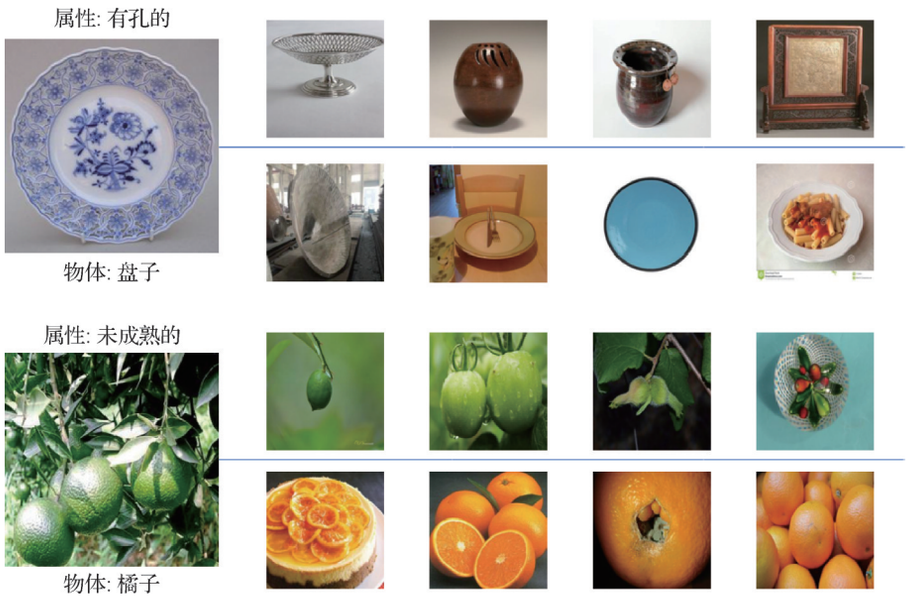

图6　视觉元素检索

Fig. 6　Visual concept retrieval

3　结　论

本文针对组合零样本识别任务提出了图像和文本信息双端解耦合组件，与CLIP模型的前馈网络模块进行拼接，旨在提升CLIP对于组合图像中属性和物体信息的解耦合能力，又能充分利用CLIP的图文信息交互能力。在3个广泛使用的组合零样本识别数据集上进行了充分的对比实验，同时在MIT-States数据集上进行了有针对性的消融实验和定性实验分析，证明了所提出模型的有效性。针对文本信息，本文提出了基于GCN的图分析模块，将文本标签中的属性、物体和组合作为节点构建成一个相互连接的组合图，其中，不同组合图之间相同的属性和物体信息之间通过加权边进行连接，然后利用GCN层在图节点之间传播特征。通过这种方式使模型理解文本数据中的邻接关系。本文在图像端设计了基于对比学习的交叉注意力解耦合组件。对于待学习图像样本，本文模型通过标签检索与之具有相同属性和相同物体的图像，将待学习图像与这两幅具有相同元素的图像分别放入到交叉注意力模块中，从更深层次的注意力级别对图像中的属性和物体元素进行学习，达到视觉端解耦合的目的。

本文对所提出的模型分别在封闭世界和开放世界实验设置下进行了广泛的实验评估。并且在MIT-States和UT-Zappos数据集上取得了优异的表现，在C-GQA数据集上取得了第2的表现，可能是因为C-GQA数据集中的样本属性元素不够清晰，并且本文提出的模型记忆域不够宽，导致对一些可能产生分歧的属性标签识别出了错误的结果。本文对所提出的组件在MIT-States数据集上进行了消融实验，证明所提出组件对于提升模型组合零样本识别能力都具有正向作用。本文还对组件中的上下文信息进行了消融，证明组合零样本识别这种识别标签的计算机视觉任务中，上下文信息依然关键。本文进行了3种定性实验分析，分别从图搜文、文搜图以及视觉元素检索3个层面对提出模型的能力进行了检验，虽然存在一些错误结果，但是对这些错误结果进行观察可以发现，这些错误并不能说明模型的效果存在问题。

通过本文研究，对于提升组合零样本识别任务模型的性能，构建一个属性标签足够明显的数据集具有现实意义，现今的模型在训练模型能力时十分受限于现有数据集中不够明确、容易产生歧义的属性标签，虽然这些属性在现实生活中广泛存在，但是在所选取的样本图像中并不能明确地表示出来，容易污染模型的知识库。同时，探究如何更加充分地在解耦合系统中实现图像信息和文本标签之间的信息交互十分重要，虽然本文对文本标签进行了利用，但是在与图像信息进行交互时，仍然依靠的是CLIP

模型本身的图文信息匹配能力，缺乏让模型明确将属性和物体进行区分的过程。如何让模型充分构建属性和物体元素与标签之间的对应关系，也即进行更加充分的、完全的属性物体解耦合，是进一步提升模型表现的主要方向。

参考文献（References）

Anwaar M U, Pan Z H and Kleinsteuber M. 2022. On leveraging variational graph embeddings for open world compositional zero-shot learning//Proceedings of the 30th ACM International Conference on Multimedia. Lisboa, Portugal: Association for Computing Machinery: 4645-4654 [DOI: 10.1145/3503161.3547798]

Atzmon Y, Kreuk F, Shalit U and Chechik G. 2020. A causal view of compositional zero-shot recognition//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #124

Chao W L, Changpinyo S, Gong B Q and Sha F. 2016. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild//Proceedings of the 14th European Conference on Computer Vision——ECCV 2016. Amsterdam, the Netherlands: Springer: 52-68 [DOI: 10.1007/978-3-319-46475-6_4]

Chen Z, Huang Y F, Chen J Y, Geng Y X, Zhang W, Fang Y, et al. 2023. DUET: cross-modal semantic grounding for contrastive zero-shot learning//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI Press: 405-413 [DOI: 10.1609/aaai.v37i1.25114]

Han A Y, Yang G, Liu X M and Liu Y. 2023. Visual-semantic dual-disentangling for generalized zero-shot learning. Journal of Image and Graphics, 28(9): 2913-2926 (韩阿友, 杨关, 刘小明, 刘阳. 2023. 视觉—语义双重解纠缠的广义零样本学习. 中国图象图形学报, 28(9): 2913-2926) [DOI: 10.11834/jig.220486]

Hao S Z, Han K and Wong K Y K. 2023. Learning attention as disentangler for compositional zero-shot learning//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 15315-15324 [DOI: 10.1109/CVPR52729.2023.01470]

Hu X M and Wang Z L. 2023. Leveraging sub-class discrimination for compositional zero-shot learning//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI Press: 890-898 [DOI: 10.1609/aaai.v37i1.25168]

Isola P, Lim J J and Adelson E H. 2015. Discovering states and transformations in image collections//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE: 1383-1391 [DOI: 10.1109/CVPR.2015.7298744]

Kipf T N and Welling M. 2017. Semi-supervised classification with graph convolutional networks//Proceedings of 2017 International Conference on Learning Representations. 1-10 [DOI: 10.48550/arXiv.1609.02907]

Li X Y, Yang X, Wei K, Deng C and Yang M L. 2022. Siamese contrastive embedding network for compositional zero-shot learning//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 9316-9325 [DOI: 10.1109/CVPR52688.2022.00911]

Li Y, Liu Z, Jha S and Yao L N. 2023. Distilled reverse attention network for open-world compositional zero-shot learning//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 1782-1791 [DOI: 10.1109/ICCV51070.2023.00171]

Li Y L, Xu Y, Mao X H and Lu C W. 2020. Symmetry and group in attribute-object compositions//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 11313-11322 [DOI: 10.1109/CVPR42600.2020.01133]

Lu X, Guo S, Liu Z and Guo J. 2023. Decomposed soft prompt guided fusion enhancing for compositional zero-shot learning//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE Computer Society: 23560-23569. [DOI: 10.1109/CVPR52729.2023.02256]

Mancini M, Naeem M F, Xian Y Q and Akata Z. 2021. Open world compositional zero-shot learning//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 5218-5226 [DOI: 10.1109/CVPR46437.2021.00518]

Misra I, Gupta A and Hebert M. 2017. From red wine to red tomato: Composition with context//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 1160-1169 [DOI: 10.1109/CVPR.2017.129]

Naeem M F, Xian Y Q, Tombari F and Akata Z. 2021. Learning graph embeddings for compositional zero-shot learning//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 953-962 [DOI: 10.1109/CVPR46437.2021.00101]

Nagarajan T and Grauman K. 2018. Attributes as operators: factorizing unseen attribute-object compositions//Proceedings of the 15th European Conference on Computer Vision——ECCV 2018. Munich, Germany: Springer: 172-190 [DOI: 10.1007/978-3-030-01246-5_11]

Nayak N V, Yu P and Bach S H. 2023. Learning to compose soft prompts for compositional zero-shot learning//Proceedings of 2023 International Conference on Learning Representations (ICLR). 1-15 [DOI: 10.48550/arXiv.2204.03574]

Purushwalkam S, Nickel M, Gupta A and Ranzato M. A. 2019. Task-driven modular networks for zero-shot compositional learning//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 3592-3601 [DOI: 10.1109/ICCV.2019.00369]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning.Virtual:PMLR:8748-8763

Tao P, Feng L, Du Y D, Gong X and Wang J. 2024. Meta-cosine loss for few-shot image classification. Journal of Image and Graphics, 29(2): 506-519 (陶鹏, 冯林, 杜彦东, 龚勋, 王俊. 2024. 面向元余弦损失的少样本图像分类. 中国图象图形学报, 29(2): 506-519) [DOI: 10.11834/jig.230127]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 5998-6008

Wang H N, Yang M L, Wei K and Deng C. 2023. Hierarchical prompt learning for compositional zero-shot recognition//Proceedings of the 32nd International Joint Conference on Artificial Intelligence. Macao, China: #163 [DOI: 10.24963/ijcai.2023/163]

Wang Z Q, Li Y, Zhang R, Wang J B, Li Y C and Chen Y. 2024. Few-shot SAR image classification: a survey. Journal of Image and Graphics, 29(7): 1902-1920 (王梓祺, 李阳, 张睿, 王家宝, 李允臣, 陈瑶. 2024. 小样本SAR图像分类方法综述. 中国图象图形学报, 29(7): 1902-1920) [DOI: 10.11834/jig.230359]

Wen L, Gou G L, Bai R F and Miao W Y. 2025. Bifurcated attention and feature interaction for few-shot fine-grained learning. Journal of Image and Graphics, 30(5): 1419-1432 (文浪, 苟光磊, 白瑞峰, 缪宛谕. 2025. 双分支注意和特征交互的小样本细粒度学习. 中国图象图形学报, 30(5): 1419-1432) [DOI: 10.11834/jig.240429]

Wu Z H, Pan S R, Chen F W, Long G D, Zhang C Q and Yu P S. 2021. A comprehensive survey on graph neural networks. IEEE Transactions on Neural Networks and Learning Systems, 32(1): 4-24 [DOI: 10.1109/TNNLS.2020.2978386]

Xu G Y, Chai J and Kordjamshidi P. 2024. GIPCOL: graph-injected soft prompting for compositional zero-shot learning//Proceedings of 2024 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 5774-5783 [DOI: 10.1109/WACV57701.2024.00567]

Yang F Y, Wang R P and Chen X L. 2022. SEGA: semantic guided attention on visual prototype for few-shot learning//Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE: 1056-1066 [DOI: 10.1109/WACV51458.2022.00165]

Yu A and Grauman K. 2014. Fine-grained visual comparisons with local learning//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 192-199 [DOI: 10.1109/CVPR.2014.32]

Yu Y, Chen N and Cheng K Y. 2025. Uncertainty domain awareness network for cross-domain few-shot image classification. Journal of Image and Graphics, 30(2): 518-532 (余悦, 陈楠, 成科扬. 2025. 不确定性域感知网络在少样本跨域图像分类中的研究. 中国图象图形学报, 30(2): 518-532) [DOI: 10.11834/jig.240142]

Zhang G M, Yan W S and Huang J Y. 2025. Generative zero-shot image recognition combined with double contrastive embedding learning. Journal of Image and Graphics, 30(5): 1389-1403 (张桂梅, 闫文尚, 黄军阳. 2025. 结合双重对比嵌入学习的生成式零样本图像识别. 中国图象图形学报, 30(5): 1389-1403) [DOI: 10.11834/jig.240526]

Zheng Z H, Zhu H D and Nevatia R. 2024. CAILA: concept-aware intra-layer adapters for compositional zero-shot learning//Proceedings of 2024 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 1710-1720 [DOI: 10.1109/WACV57701.2024.00174]

Zhou K Y, Yang J K, Loy C C and Liu Z W. 2022. Learning to prompt for vision-language models. International Journal of Computer Vision, 130(9): 2337-2348 [DOI: 10.1007/s11263-022-01653-1]

作者简介

田弋，男，硕士研究生，主要研究方向为多模态学习。
E-mail:2412391054@st.gxu.edu.cn
钟磊，通信作者，男，高级工程师，主要研究方向为视联网、人工智能和网络数据安全。E-mail:15277166966@139.com
钱毅鑫，男，硕士研究生，主要研究方向为计算机视觉、多模态信息处理和图像检索。E-mail:2212391035@st.gxu.edu.cn
黄清宝，男，教授，主要研究方向为自然语言处理、图像理解与多模态智能。E-mail: qbhuang@gxu.edu.cn
陈佳岳，男，本科生，主要研究方向为多模态学习。
E-mail:2741960394@qq.com
伍贤瑞，男，工程师，主要研究方向为视联网、人工智能和全业务通信网络。E-mail:13877187666@139.com