中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(XXXX)XX-0001-18

论文引用格式:Wang Jin, Du XinYu, Ding Xin. Few-shot image classification with text-visual semantic association and multi-grained semantic alignment[J/OL]. Journal of Image and Graphics, XXXX:1-18. DOI: 10.11834/jig.260186.(王进, 杜欣豫, 丁新. 图文语义关联与多粒度语义对齐的小样本图像分类[J/OL]. 中国图象图形学报,XXXX:1-18. DOI: 10.11834/jig.260186.)[DOI:10.11834/jig.260186]

图文语义关联与多粒度语义对齐的小样本图像分类

王进[1],杜欣豫[1],丁新[1,2*]

1. 南通大学人工智能与计算机学院,南通　226019; 2. 华中科技大学电子信息与通信学院,武汉　430074

摘　要: **目的**　小样本图像分类旨在利用极少量标注样本完成新类别识别。现有图文语义增强方法多依赖类别名称或简短提示,语义粒度较粗,难以充分刻画类别局部属性、外观细节及类间细微差异;同时,文本语义与视觉特征之间存在模态鸿沟,简单融合方式难以实现充分对齐。为此,提出一种面向多粒度语义对齐的图文语义多分支对齐网络(text-visual semantic multi-branch alignment network,TSMA-Net)。**方法**　设计语义信息挖掘模块,以类别名称为语义锚点,引导大语言模型离线生成包含外观特征、局部属性和差异线索的细粒度语义描述,并通过语义精炼压缩冗余信息与歧义表达。进一步通过语义适配模块,将类别名称语义与多角度细粒度语义表示进行加权融合,得到更加稳健且具有判别性的类别文本表示。在此基础上,提出多分支对齐模块,将融合语义投影到多个独立子空间,与视觉特征进行深度对齐和交互,并结合残差重校准结构抑制跨模态噪声,提升语义特征的表达能力。最终通过视觉原型与语义增强原型双路径协同完成查询样本分类。**结果**　在miniImageNet、tieredImageNet、CIFAR-FS和FC100上,TSMA-Net均取得稳定提升。与相同Visformer-Tiny骨干的SimpleFSL相比,在1-shot/5-shot任务上分别提升2.11%/0.47%、2.35%/0.49%、1.21%/0.28%和0.54%/0.10%。消融实验验证了各模块的有效性。**结论**　所提出的TSMA-Net能够在小样本场景下有效挖掘更丰富、更具判别性的类别语义信息,并实现语义与视觉特征的深度对齐,从而提升类别原型的表征能力与分类性能。该方法在多个标准基准数据集上表现出良好的有效性与泛化能力,尤其在1-shot任务中优势更为明显。

关键词: 小样本学习;小样本图像分类;原型学习;多模态学习;跨模态语义对齐;大语言模型

Few-shot image classification with text-visual semantic association and multi-grained semantic alignment

Wang Jin[1], Du XinYu[1], Ding Xin[1,2*]

1. School of Artificial Intelligence and Computer Science, Nantong University, Nantong 226019, China; 2. School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China

Abstract: **Objective** Few-shot image classification aims to recognize novel visual categories using only a few labeled samples per class. It is an important research topic for improving the generalization ability of visual recognition systems under low-resource conditions. Existing few-shot learning methods have achieved considerable progress through metric learning, meta-learning, data augmentation, and pretrained visual representation transfer. However, when the number of labeled support images is extremely limited, especially in the 1-shot setting, purely visual methods often suffer from

收稿日期:2026-04-09;**修回日期**:XXXX-XX-XX

* **通信作者**:丁新 xding@163.com

基金项目:国家自然科学基金项目(62501245);南通市自然科学基金面上项目(JC2025090)

Supported by:National Natural Science Foundation of China(62501245)

unstable prototype estimation and insufficient discrimination between visually similar classes. Recently, image-text collaborative learning and semantic-enhanced few-shot learning have provided new opportunities for addressing this limitation. By introducing class names, attributes, textual prompts, or language model representations, semantic information can serve as an additional prior to compensate for insufficient visual evidence. Nevertheless, most existing semantic-enhanced methods still rely mainly on coarse-grained textual information, such as category names or short prompt templates. Such semantic information usually describes only the global concept of a class and is inadequate for capturing local attributes, appearance details, part-level cues, and subtle inter-class differences. Moreover, when fine-grained textual descriptions are directly generated or introduced, they may contain redundant expressions, ambiguous words, and weakly discriminative descriptions. Another important issue is the modality gap between textual semantics and visual features. Simple feature concatenation or additive fusion is insufficient for fully aligning multi-grained semantic information with visual representations. Therefore, how to construct high-quality class semantics and how to deeply align semantic representations with visual features remain key challenges for semantic-enhanced few-shot image classification. To address these issues, this paper proposes a few-shot image classification framework named text-visual semantic multi-branch alignment network (TSMA-Net), which aims to improve the robustness and discriminability of class prototypes by combining multi-angle semantic mining, semantic refinement, adaptive semantic fusion, and multi-branch cross-modal alignment. **Method** The proposed TSMA-Net consists of two main components: a text-visual semantic association module and a multi-branch alignment module. The semantic association module first mines fine-grained category descriptions from multiple complementary perspectives and refines them to reduce redundancy and ambiguity. These fine-grained semantics are then adaptively fused with class-name semantics to obtain robust and discriminative multi-grained textual representations. The multi-branch alignment module further projects the fused semantics into multiple subspaces and interacts them with visual features, enabling more sufficient cross-modal alignment than direct additive fusion. A residual recalibration structure is introduced to suppress noisy cross-modal responses, and visual prototypes and semantically enhanced prototypes are jointly used for query classification. **Result** Extensive experiments are conducted on four standard few-shot image classification benchmarks, including miniImageNet, tieredImageNet, CIFAR-FS, and FC100. The evaluation follows the widely used 5-way 1-shot and 5-way 5-shot settings. Compared with existing visual-only and semantic-enhanced few-shot learning methods, TSMA-Net achieves competitive or superior performance across different datasets. In particular, compared with SimpleFSL using the same Visformer-Tiny backbone, TSMA-Net improves the classification accuracy by 2.11% and 0.47% on miniImageNet under the 1-shot and 5-shot settings, respectively. On tieredImageNet, the corresponding improvements are 2.35% and 0.49%. On CIFAR-FS, TSMA-Net obtains gains of 1.21% and 0.28%, and on FC100, it improves the accuracy by 0.54% and 0.10%. These results demonstrate that the proposed method can effectively improve few-shot classification performance across datasets with different image distributions, category granularity, and semantic complexity. The performance gains are more pronounced in the 1-shot setting, indicating that the proposed semantic mining and multi-branch alignment strategy is particularly beneficial when visual evidence is extremely scarce. Ablation experiments further confirm the contribution of each component. The text-visual semantic association module improves the quality of class textual representations by introducing multi-angle fine-grained semantic descriptions, while the multi-branch alignment module enhances cross-modal interaction and improves prototype discriminability. The analysis of the fusion factor shows that an appropriate balance between class-name semantics and fine-grained semantics is important for stable semantic representation. The branch number analysis indicates that multiple semantic projection branches can improve performance by decomposing semantic information into different subspaces, whereas too many branches may introduce redundancy and over-fragment the semantic representation. The residual recalibration experiment also shows that recalibrating the fused cross-modal representation helps suppress noise and improve robustness. In addition, visualization results based on t-SNE show that TSMA-Net forms more compact intra-class clusters and clearer inter-class separation than the baseline method, which further verifies the effectiveness of the proposed semantic association and alignment mechanisms. **Conclusion** This paper proposes TSMA-Net for semantic-enhanced few-shot image classification. Unlike existing methods that mainly use category names or short prompts as textual priors, TSMA-Net explicitly mines multi-angle fine-grained semantic descriptions from a large language model and refines them into concise, relevant, and discriminative class semantics. By adaptively integrating class-name seman-

tics and fine-grained descriptions, the method constructs a more robust multi-grained textual representation. Furthermore, the proposed multi-branch alignment module projects semantic information into multiple subspaces and performs deeper interaction with visual features, thereby alleviating the modality gap between text and vision. The residual recalibration structure further improves the stability of cross-modal fusion, and the dual-path prototype classification strategy enhances decision robustness. Experimental results on four benchmark datasets demonstrate that TSMA-Net can learn more discriminative class prototypes under limited supervision and consistently improve few-shot classification performance. The advantage is especially evident in the 1-shot scenario, where additional semantic priors are most needed. Overall, the proposed framework provides an effective way to exploit large language model-generated fine-grained semantics for few-shot visual recognition and offers a useful reference for future research on multimodal semantic alignment under data-scarce conditions.

Key words: few-shot learning; few-shot image classification; prototype learning; multimodal learning; cross-modal semantic alignment; large language model

论文引用格式:[DOI:10.11834/jig.260186]

0 引　言

人工智能的长期目标是开发出具备人类级别智能的学习能力。近年来,深度学习(LeCun等,2015)在计算机视觉(He等,2016;Dosovitskiy等,2020)和自然语言处理(Achiam等,2023)等领域取得了显著进展,主要得益于大规模标注数据集的可用性。然而,在许多实际场景中,高质量标注样本稀缺,严重限制了传统监督学习的适用性。与之相反,人类能够仅从少量样本中快速学习并识别新型类别。受此启发,提出了小样本学习(Lake等,2011),旨在使机器在仅有少量标注样本的情况下实现对新类别的有效泛化。

在小样本学习中,常采用N-way K-shot(Snell等,2017)设置,支持集为N个新型类别提供每个类别K个标注样本。模型的任务是对查询集中的样本进行分类,这些查询样本同样来自该N个类别。早期工作主要聚焦于基于视觉的学习方法,通过元学习、度量学习、自监督预训练和数据增强等策略来提高模型的泛化能力。近年来,视觉方法在原型建模与特征学习方面持续取得进展,例如PBML(Fu等,2024)通过引入原型的贝叶斯先验分布,对小样本条件下的不确定性进行显式建模,从而增强类别表征的稳健性;PRSN(Dong等,2025)则利用与当前任务语义相近的基础类局部特征对原型进行重构,并施加跨图像语义一致性约束,以提升原型空间的判别能力。然而,这类方法本质上主要依赖视觉监督,在标注样本极其有限时,尤其是在1-shot场景下,模型往往难以充分捕获类别的细粒度差异与稳定判别线索,因而性能提升仍受到明显制约。

为缓解纯视觉学习在极低样本条件下的表征不足,研究者开始将类别名称、属性描述等语义信息引入小样本学习,通过融合语言先验知识辅助新类别识别。特别是随着视觉—语言预训练模型的发展,语义增强小样本学习取得了显著进展。例如,SP-CLIP(Chen等,2023)通过语义提示将文本编码器提取的类别语义表示注入视觉特征提取过程,提升了小样本识别性能;SimpleFSL(Zhou等,2025)进一步利用可学习提示激活预训练语言模型的零样本泛化能力,并结合跨模态融合与自蒸馏策略构建更具判别性的语义增强类别表示。尽管如此,现有方法大多仍依赖类名或简短提示所提供的粗粒度语义信息,难以充分表达类别的局部属性、外观细节及类间差异;同时,简单的加性融合方式也难以有效缓解类名歧义并实现视觉—语义特征的深层对齐。因此,如何获取更丰富、更具判别性且能够与视觉表征有效协同的高质量语义信息,仍然是小样本图像分类中亟待解决的关键问题。

为解决上述问题,本文提出了一种新的小样本图像分类框架——图文语义多分支对齐网络(图文语义多分支对齐网络(text-visual semantic multi-branch alignment network,TSMA-Net)。该框架主要包含两个相互协同的组成部分:图文语义关联模块和多分支对齐模块。其中图文语义关联模块从视觉与文本两个模态协同建模:一方面,通过对视觉骨干网络进行预训练,获得稳定且具有判别性的视觉特征表示;另一方面,引入语义信息挖掘模块,以类别名称为语义锚点,利用提示词工程引导大语言模型离线生成与类别相关的细粒度语义描述,并通过语义精炼压缩冗余信息、消除歧义表达。随后,将生成

的细粒度语义文本输入预训练文本编码器，得到类别的多角度细粒度语义表示，并进一步与类别名称语义进行加权融合，从而形成更加稳健且更具区分能力的类别文本表示。考虑到视觉模态与语义模态在特征分布和表示结构上存在天然差异，若缺乏有效的对齐机制，两种模态信息难以在统一空间中实现充分融合。为此，本文进一步设计多分支对齐模块，将融合语义映射到多个独立子空间，与视觉特征进行深度交互和对齐，并结合残差重校准结构抑制跨模态噪声，从而提升类别原型的判别能力和模型对新类别的泛化性能。

总体来说，本文的贡献可以总结为以下几点：1）提出图文语义关联模块，在视觉特征预训练的基础上，从大语言模型中挖掘细粒度类别语义，并将其与稳定的粗粒度类名语义进行自适应融合，从而构建兼具语义稳定性与判别能力的多粒度类别文本表示。2）提出一种跨模态多分支对齐融合方法，将融合后的多粒度语义表示映射到多个独立子空间中，与视觉特征进行分解式交互与深度融合，并结合残差重校准结构抑制跨模态噪声，提升融合表示的稳健性和类别原型的判别能力。3）所提出的TSMA-Net在四个标准小样本学习基准数据集上取得了显著的性能提升，尤其在1-shot学习任务中表现出更明显的优势。

1 相关工作

1.1 基于图像的小样本学习方法

小样本学习的早期研究主要依赖于视觉模态信息，并围绕如何在极少标注样本条件下提升模型的泛化能力，发展出以元学习、度量学习、自监督预训练与数据增强为代表的一系列方法。

其中，元学习（Meta-Learning）的核心思想是“学会如何学习”，即通过在大量小样本任务上的元训练，使模型具备快速适应新任务的能力，从而缓解数据稀缺带来的训练困难。典型代表如MAML（Finn等，2017），通过学习一个具有良好迁移性的参数初始化，使模型仅需少量梯度更新即可快速适配新任务；Meta-AdaM（Sun等，2023）则从优化器层面进行元学习，学习具有动量特性的自适应优化策略，从而显著加快小样本场景下的模型收敛速度。此外，Set-Feat（Afrasiyabi等，2022）将浅层自注意力机制嵌入特征提取网络，并以特征集合的形式对样本进行建模，有效增强了小样本场景下表征的多样性与稳定性；PBML（Fu等，2024）则在贝叶斯推断框架下引入原型先验分布，通过对不确定性的显式建模提升了模型的泛化能力。

度量学习（Metric Learning）旨在学习一个具有良好判别性的嵌入空间，使不同类别的样本在该空间中能够通过距离度量进行有效区分。ProtoNet（Snell等，2017）采用每类支持样本特征的均值作为类别原型，并基于最近邻原则完成分类，奠定了小样本度量学习的基础框架。DeepEMD（Zhang等，2020）进一步利用Earth Mover's Distance在支持集与查询样本的局部特征之间进行最优匹配，从而捕获更细粒度的局部相似性关系，有效提升了复杂场景下的分类精度。PRSN（Dong等，2025）则引入与当前任务语义相近的基础类局部特征，结合查询样本对原型进行重构，并施加跨图像的语义一致性约束，从而构建了一个更加稳健且判别性更强的原型度量空间。

随着自监督学习的发展，自监督预训练（Self-Supervised Pre-training）也被广泛引入小样本学习，用以提升特征表示的通用性与泛化能力。Align（Afrasiyabi等，2020）通过特征空间对齐的自监督预训练策略，增强了模型对下游任务变化的鲁棒性；CORL（He等，2023）通过组合式表示学习获取更加结构化的特征表征；LSFSL（Padmanabhan等，2023）利用形状先验增强了特征的泛化性并降低对纹理模式的过度依赖；SSL-ProtoNet（Lim等，2024）则将自监督预训练与自蒸馏机制相结合，并嵌入到原型网络中，有效缓解了小样本场景下的过拟合问题，同时提升了分类性能。

此外，数据增强（Data Augmentation）策略通过合成额外训练样本以扩充数据分布，从而缓解模型在小样本条件下的过拟合风险。RFS（Tian等，2020）的研究表明，先在大规模数据集上进行预训练以获得高质量的通用表征，再在小样本任务上进行微调即可取得出色性能，这充分说明了高质量表示学习与数据扩充对小样本学习的重要性。MixtFSL（Afrasiyabi等，2021）通过为每个类别引入多个特征子空间来丰富类内样本的分布多样性；MetaDiff（Zhang等，2024）利用条件扩散模型模拟小样本训练过程中的优化轨迹，为模型生成多样化的适应路

径；CADS(Zhang等，2024)通过条件退火扩散采样进一步提升了生成样本的多样性，为小样本学习提供了更加丰富的合成数据支持。

尽管上述方法仅依赖于视觉模态信息，在小样本学习场景下已取得了一定进展，但由于训练样本本身极其有限，它们在刻画类别间局部差异方面仍然存在明显不足，模型性能在极低样本(尤其是1-shot)条件下仍受到数据稀缺性的严重制约。这一局限也直接推动了后续将语义信息与视觉特征相结合的多模态小样本学习方法的发展。

1.2 基于图文的小样本学习方法

为缓解训练样本匮乏的问题，近年来研究者逐渐将类别名称、属性描述等语义信息引入小样本学习，并借助预训练语言模型或现有知识库融合视觉与语义特征，从而提升新型类别的识别能力。例如，KTN(Peng等，2019)利用类别名称的词嵌入将基础类别知识迁移到新类别，从而提升极小样本条件下的分类性能；AM3(Xing等，2019)引入自适应模态混合机制，通过注意力机制加权融合视觉特征与类别名称词向量，指导类别原型学习并增强类内表征的判别性；TRAML(Li等，2020)则根据类别语义相似度自适应调整分类决策边界(即损失函数中的边距)，更好地区分易混淆类别。

随后，研究者探索了更复杂的多模态融合策略。例如，CMGNN(Liu等，2021)将类别语义嵌入作为图结构中的“元节点”，通过图神经网络传播与视觉特征交互，提升小样本分类性能；另有一些工作在训练中施加视觉－语义对比对齐约束，借助预训练文本编码器的上下文知识学习更具泛化性的类概念表征。

零样本学习(zero-shot learning，ZSL)和广义零样本学习(generalized zero-shot learning，GZSL)同样关注视觉特征与语义信息之间的关联建模。近年来，相关研究进一步强调视觉—语义一致对齐的重要性，例如Jiang等(2025)通过视觉提示与语义提示协同增强语义相关的视觉特征，Chen等(2025)则利用大语言模型生成属性语义，并通过局部视觉—语义对齐提升零样本识别能力。上述研究为跨模态语义对齐提供了重要参考。然而，ZSL/GZSL通常假设新类别没有标注图像样本，而本文面向小样本图像分类任务，测试阶段每个新类别仍包含少量支持样本。因此，本文在视觉原型基础上引入多角度细粒度语义，并通过多分支对齐模块增强语义与视觉特征的交互，从而提升类别原型的判别性与鲁棒性。

此外，大规模视觉－语言预训练模型的出现为小样本学习提供了新思路。例如，对比语言-图像预训练(contrastive language-image pre-training，CLIP)(Radford等，2021)模型通过自然语言监督学习到图文对齐的表示，可直接利用文本提示实现零样本分类；CoOP(Zhou等，2022)方法通过优化可学习的上下文向量替代固定模板，自适应调整文本提示，从而提高了CLIP(Radford等，2021)在小样本任务中的性能。进一步，LPE-CLIP(Yang等，2023)方法基于语义引导潜在部件嵌入思想，利用类别语义知识生成类别特定的潜在部件滤波器，并通过潜在部件发现和部件级相似度度量增强类别表示；此外，SP-CLIP(Chen等，2023)方法通过语义提示将预训练文本编码器提取的类别语义表示注入视觉特征提取过程；最近，简单的SimpleFSL(Zhou等，2025)框架通过可学习提示显式激活预训练语言模型的零样本泛化能力，并采用加性跨模态融合和自蒸馏机制构建判别性更强的语义增强类表示，从而在小样本条件下显著提升分类性能。

综上，现有基于图文协同的小样本学习方法表明，语言先验能够在样本稀缺条件下为类别原型提供有效补充。然而，从语义建模和跨模态融合两个角度来看，这类方法仍存在一定共性局限。首先，在语义来源与语义粒度方面，许多方法主要依赖类别名称、固定模板或简短提示构造文本表示，其语义信息通常停留在类别整体概念层面，难以充分覆盖局部属性、外观细节以及易混淆类别之间的细微差异。其次，在融合机制方面，现有方法多采用注意力加权、特征拼接或加性融合等相对浅层的交互方式，虽然能够引入语义先验，但难以有效缓解文本语义与视觉表征之间的模态鸿沟，也难以充分建模粗粒度语义与细粒度语义之间的互补关系。因此，如何构建兼具稳定性与判别性的多粒度类别语义，并设计更充分的跨模态对齐机制，是进一步提升图文小样本分类性能的关键。基于此，本文提出TSMA-Net，通过图文语义关联模块挖掘并融合类别名称语义与多角度细粒度语义，同时利用多分支对齐模块实现语义表示与视觉特征的深层交互，从而增强小样本场景下类别原型的鲁棒性与判别能力。

2 研究方法

本节对提出的TSMA-Net框架进行整体介绍，并对其两个关键组成部分进行系统阐述。图文语义关联模块负责从视觉与文本两个模态中提取并构建互补语义信息：一方面，通过视觉骨干网络的预训练获取稳定且具有判别性的视觉特征表示；另一方面，通过语义信息挖掘模块，以类别名称为语义锚点，利用提示词工程引导大语言模型离线生成细粒度语义描述，并通过语义精炼与语义适配获得更加稳健的类别文本表示。在此基础上，为了多粒度语义对齐设计了多分支对齐机制，将融合后的语义表示映射到多个独立子空间，使其能够在不同语义粒度上与视觉特征进行交互与对齐，从而更充分地建模类别的整体语义与细节差异；同时结合残差重校准结构抑制跨模态噪声，进一步提升类别原型的判别性与模型对新类别的泛化能力。TSMA-Net的整体框架结构如图1所示。

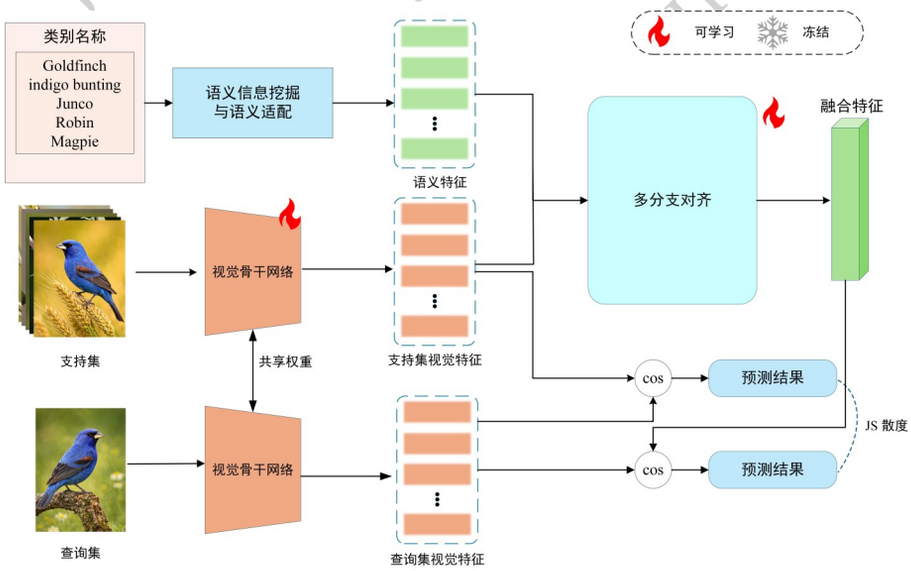


图1 TSMA-Net框架总体结构示意图

Figure 1 Overall architecture of the TSMA-Net framework

2.1 图文语义关联

2.1.1 视觉信息的挖掘与预训练

参考已有工作(Chen等，2023；Zhou等，2025)，本文首先在训练数据集上对视觉骨干网络Visformer-Tiny进行预训练，以加快后续元训练阶段的收敛速度并获得更稳定的视觉表征。在预训练阶段，所有基准类样本输入视觉骨干网络进行特征提取，并在基准类别监督下进行优化，使模型学习具有良好判别性的视觉特征。经过该阶段训练后，视觉编码器能够为后续图文语义关联与多分支对齐提供可靠的视觉表示，其在整体框架中的位置如图1中视觉骨干网络所示。

2.1.2 语义信息挖掘与语义适配模块

在小样本图像分类任务中，仅依赖类别名称构造文本提示虽然能够提供一定的语义先验，但其表达通常较为粗粒度，难以充分刻画类别的局部结构、外观属性及与相近类别之间的细粒度差异。尤其在1-shot或5-shot场景下，视觉样本数量极少，单纯依靠支持样本往往难以建立稳定且具有判别性的类别原型。为此，本文在图文语义关联框架中引入语义信息挖掘模块，以类别名称为语义锚点，通过提示词工程引导GPT-4(Achiam等，2023)离线生成更丰富的类别描述，并进一步与类别名称语义进行融合，从而获得更稳定且更具区分能力的类别文本表示。该模块的整体流程如图2所示，包括多角度语义生成、语义精炼以及与类别名称语义的适配融合。

不同于直接采用单一模板生成文本描述，本文从多个互补角度对类别语义进行挖掘。如图2所示，本文的提示工程围绕类别名称构造三个互补角度的提示模板，以保证生成语义既包含类别整体概

念，又覆盖局部细节、类间区分线索以及实例变化信息。具体而言，Angle 1 为“全局与局部”（Global vs Local）角度，关注类别的整体轮廓和主要外观特征，以及局部纹理、标记区域和细粒度细节；Angle 2 为“共性与独特性”（Commonality vs Uniqueness）角度，关注该类别通常共有的典型特征，以及区别于相近类别的判别性特征；Angle 3 为“静态与动态特征”（Static vs Dynamic Features）角度，关注颜色、形状等相对稳定的视觉属性，以及纹理模式、姿态或外观形式上可能出现的变化。为增强生成过程的可复现性，本文采用如下提示模板：[Angle 1: Global vs Local. Overall silhouette and dominant appearance. Local textures, markings, and fine-grained details. Angle 2: Commonality vs Uniqueness. Typical traits shared by the category. Distinct traits separating it from similar classes. Angle 3: Static vs Dynamic Cues. Relatively stable attributes such as color and shape. Possible variations in pattern, pose, or appearance.]

后文式（2）中的语义精炼操作 $R(\cdot)$ 通过离线大语言模型指令完成，而非人工规则过滤或训练阶段的可学习模块。具体而言，本文将多角度候选描述输入大语言模型，并采用如下精炼 Prompt 对其进行压缩和规范化：[Remove redundant or repetitive descriptions. Resolve ambiguous or weakly discriminative expressions. Keep concise, relevant, and discriminative cues.] 该过程用于去除冗余或重复描述，修正歧义性或弱判别性表达，并保留简洁、相关且具有区分性的视觉语义线索。

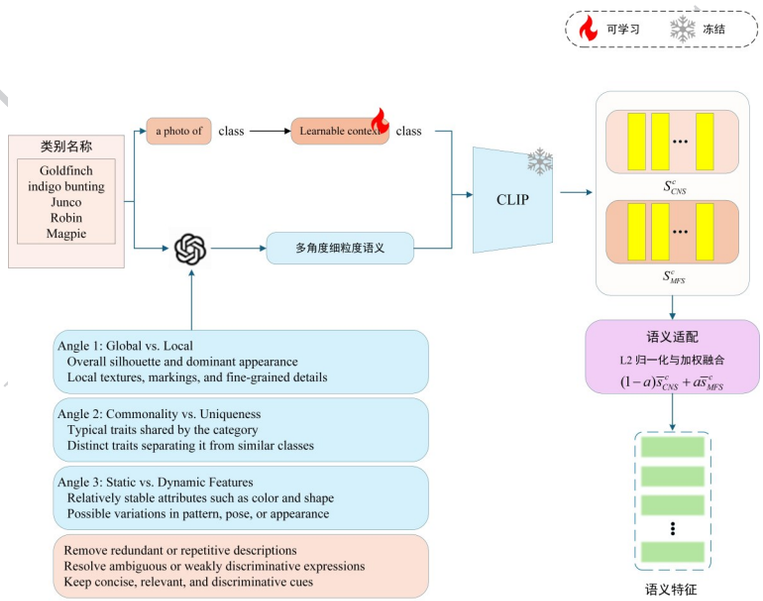


图2　语义信息挖掘与语义适配流程示意图

Figure 2　Illustration of the semantic information mining and semantic adaptation process

以类别“Junco”为例，Angle 1 生成 small sparrow-like bird, compact body, short conical bill, pale belly 等整体与局部特征；Angle 2 生成 dark hood or gray head, white outer tail feathers, contrast between dark upperparts and pale underparts 等类间判别线索；Angle 3 生成 gray, brown, black and white plumage variations 等相对稳定属性及实例变化描述。经过语义精炼和压缩后，最终得到的多角度细粒度语义文本为：“A junco is a small sparrow-like bird with a compact body, short conical bill, gray or dark head, pale belly, and distinctive white outer tail feathers, often showing gray, brown, black, and white plumage variations. ”。

具体而言，围绕类别 c，分别从整体与局部、共性与差异、静态属性与变化特征三个角度构造提示模板，并将其输入大语言模型生成候选描述。设第 m 个角度对应的提示模板记为 $P_m(\cdot)$，则该角度下生成的语义文本可表示为：

$$T_c^{(m)} = G(P_m(c)), m = 1,2,3 \tag{1}$$

式中，$G(\cdot)$ 表示大语言模型的生成过程，$T_c^{(m)}$ 表示类

别c在第m个角度下得到的候选语义描述。上述三个角度分别关注类别的整体轮廓与局部细节、典型共性与独有特征,以及相对稳定的属性与可能出现的模式变化。通过多角度提示,不同描述之间能够形成互补,从而提升类别语义的覆盖范围与表达粒度。

需要指出的是,多角度生成的候选文本虽然包含更丰富的语义线索,但其中也可能存在冗余表述、重复信息或局部歧义。如果直接将这些原始文本输入文本编码器,容易引入开放语义噪声,削弱后续跨模态对齐的稳定性。因此,本文进一步对候选描述进行语义精炼。具体来说,针对每个角度下生成的候选文本$T_c^{(m)}$,利用一致性检查和压缩式重写策略对其进行修正,删除无助于类别判别的冗余信息,并保留与外观属性和类别区分最相关的描述,得到精炼后的语义文本:

$$\tilde{T}_c^{(m)} = R(T_c^{(m)}, c), m = 1,2,3 \tag{2}$$

式中,$R(\cdot)$表示语义精炼操作。该过程并非训练阶段显式优化的目标函数,而是基于提示词工程对大语言模型输出进行规范化约束,其核心目的在于使生成文本同时满足“语义相关、描述具体、表达简洁、具有区分性”等要求。

将三个角度下经语义精炼后的候选文本输入大语言模型进行整合与压缩,生成类别c最终的多角度细粒度语义文本T_f^c。该过程可表示为:

$$T_f^c = G\left(\tilde{T}_c^{(1)}, \tilde{T}_c^{(2)}, \tilde{T}_c^{(3)}\right) \tag{3}$$

式中,f表示多角度细粒度语义文本。在获得最终的多角度细粒度语义文本后,本文将其输入CLIP文本编码器$E(\cdot)$,得到类别c对应的细粒度语义表示:

$$\boldsymbol{s}_f^c = E(T_f^c) \tag{4}$$

$\boldsymbol{s}_f^c$能够更充分地编码类别的整体外观、局部纹理以及与相近类别之间的差异性线索,因此在类间区分方面更具优势。

尽管多角度细粒度语义能够提供更丰富的判别信息,但其来源于大语言模型生成文本,仍可能受到生成偏差与开放语义噪声的影响。相比之下,由类别名称直接构造的文本表示通常更加稳定,能够更可靠地刻画类别中心概念。因此,本文进一步引入类别名称语义,并与多角度细粒度语义进行自适应融合。对于类别名称c,采用与CLIP提示学习一致的方式,引入可学习上下文序列$\{v_i\}_{i=1}^{l}$,构造类别名称文本序列:

$$T_n^c = [\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_l, c] \tag{5}$$

将其输入文本编码器,可得到类别名称语义表示:

$$\boldsymbol{s}_n^c = E(T_n^c) \tag{6}$$

由于$\boldsymbol{s}_n^c$与$\boldsymbol{s}_n^c$分别来源于类别名称和多角度语义描述,两者在表示尺度上可能存在差异。为减小尺度不一致对融合结果的影响,本文先对两类语义表示进行l_2归一化:

$$\overline{\boldsymbol{s}}_n^c = \frac{\boldsymbol{s}_n^c}{\left\|\boldsymbol{s}_n^c\right\|_2}, \quad \overline{\boldsymbol{s}}_f^c = \frac{\boldsymbol{s}_f^c}{\left\|\boldsymbol{s}_f^c\right\|_2} \tag{7}$$

在此基础上,构造类别c的最终融合语义表示为:

$$\boldsymbol{s}_c = (1-a)\overline{\boldsymbol{s}}_n^c + a\overline{\boldsymbol{s}}_f^c \tag{8}$$

式中,$a \subseteq [0,1]$为融合因子,用于调节类别名称语义与多角度细粒度语义的相对贡献。当a较小时,最终语义表示更偏向于稳定的类别名称语义;当a较大时,最终语义表示更强调多角度语义带来的细粒度判别信息。通过这种融合方式,本文在类别中心概念的稳定性与细粒度描述的区分性之间建立了有效平衡,从而得到更加稳健且更具判别能力的类别文本表示。该表示随后被送入后续多分支对齐模块,与视觉特征进行深度交互和融合,以提升小样本场景下的类别原型表征能力与最终分类性能。

2.2 多分支对齐模块

承接上一节,图文语义关联模块已经为每个类别c构建了融合后的多粒度语义表示$\boldsymbol{s}_c$,该表示同时包含类别名称语义提供的稳定类别中心信息,以及多角度细粒度语义提供的局部属性与差异线索。在此基础上,多分支对齐模块进一步将$\boldsymbol{s}_c$与视觉骨干网络提取的样本视觉特征$f(x)$进行跨模态对齐与深层交互,从而将文本语义先验有效注入视觉原型构建过程。然而,若直接对文本语义向量进行简单降维并与视觉特征相加,往往难以实现理想的跨模态融合。一方面,直接降维可能导致文本语义在变换过程中被压缩,从而损失对视觉识别至关重要的细粒度语义信息;另一方面,文本与视觉特征原本处于结构差异显著的表征空间,简单相加会放大模态间的不一致性,引入额外噪声,并使跨模态的尺度偏移难以有效消除。如图3所示,为解决上述问题,本文设计了多分支对齐模块,使文本语义能够在多个独

立子空间中被分解、重组并与视觉表征进行深层次融合。

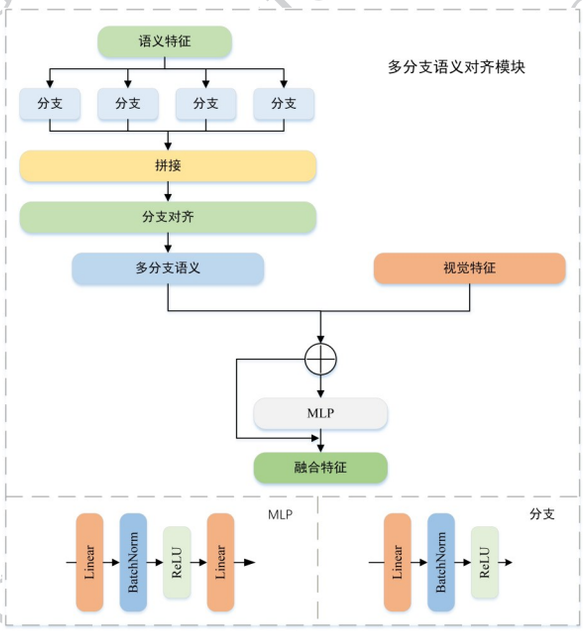


图3　多分支对齐模块结构示意图

Figure 3　Illustration of the architecture of the multi-branch semantic alignment module

具体而言，多分支对齐可统一表示为：

$$\boldsymbol{u}=f(\boldsymbol{x})+\boldsymbol{W}_f Concat\begin{pmatrix}\sigma(\boldsymbol{W}_1\boldsymbol{s}_c),\sigma(\boldsymbol{W}_2\boldsymbol{s}_c)\\ ,\cdots,\sigma(\boldsymbol{W}_H\boldsymbol{s}_c)\end{pmatrix} \tag{9}$$

式中，$\boldsymbol{W}_i$为第i个语义头的投影矩阵，$\sigma(\cdot)$为非线性激活函数，$\boldsymbol{W}_f$用于将多分支语义表示压缩到与视觉特征相同的维度。通过式(9)，多分支信息被整合至视觉特征空间，从而在统一空间内显式注入细粒度语义线索。

为了进一步抑制跨模态对齐过程中的噪声并突出关键语义成分，在$\boldsymbol{u}$上引入瓶颈式残差重建结构，对初步融合表示进行再校准。该过程可概括为：

$$\boldsymbol{v}=\boldsymbol{u}+\boldsymbol{W}_2\sigma(\boldsymbol{W}_1\boldsymbol{u}) \tag{10}$$

式中，$\boldsymbol{W}_1$、$\boldsymbol{W}_2$为瓶颈结构中的降维与升维矩阵，$\sigma(\cdot)$为非线性变换。输出$\boldsymbol{v}$为语义增强后的重构原型。

然后基于支持集构建两种类别原型，以从视觉空间与语义增强空间分别刻画类别特征。对于第i类，其K个支持样本的视觉特征均值定义为视觉原型：

$$\boldsymbol{c}_i^{\mathrm{v}}=\frac{1}{K}\sum_{j=1}^{K}f(\boldsymbol{x}_{ij}) \tag{11}$$

式中，$f(\cdot)$表示视觉编码器。同时，将支持样本经过语义融合与瓶颈式残差结构得到的语义增强特征求均值，定义为语义增强原型：

$$\boldsymbol{c}_i^{\mathrm{s}}=\frac{1}{K}\sum_{j=1}^{K}\boldsymbol{v}(\boldsymbol{x}_{ij}) \tag{12}$$

基于上述两种类别原型，分别构建视觉路径分类器与语义增强路径分类器。对于查询样本$\boldsymbol{x}_{\mathrm{q}}$，两条路径上的预测分布可统一写作：

$$\hat{y}_{\mathrm{q}}^{m}=\frac{\exp(<f(\boldsymbol{x}_{\mathrm{q}}),\boldsymbol{c}_i^{m}>/\tau)}{\sum_{j=1}^{N}\exp(<f(\boldsymbol{x}_{\mathrm{q}}),\boldsymbol{c}_j^{m}>/\tau)} \tag{13}$$

式中，$m=\mathrm{v}$和$m=\mathrm{s}$分别对应视觉原型与语义增强原型构建的分类器，N表示当前episode中的类别数，τ为温度系数。

最终训练损失定义为：

$$L=CE(\hat{\boldsymbol{y}}_{\mathrm{q}}^{\mathrm{v}},\boldsymbol{y}_{\mathrm{q}})+CE(\hat{\boldsymbol{y}}_{\mathrm{q}}^{\mathrm{s}},\boldsymbol{y}_{\mathrm{q}})+\lambda JS(\hat{\boldsymbol{y}}_{\mathrm{q}}^{\mathrm{v}},\hat{\boldsymbol{y}}_{\mathrm{q}}^{\mathrm{s}}) \tag{14}$$

式中，$CE(\cdot)$为交叉熵损失，$JS(\cdot,\cdot)$表示Jensen-Shannon散度，λ为蒸馏损失权重。

本文采用Jensen-Shannon散度约束视觉路径与语义增强路径的预测分布一致性，而非采用单向KL散度。主要原因在于，两条路径在本文中并不存在固定的教师-学生关系，而是共同参与分类决策的协同分支。JS散度具有对称性和有界性，能够对两条预测分布进行更加稳定的双向一致性约束，从而减少单向KL散度可能引入的路径偏置。

3　实　验

3.1　实验数据

miniImageNet（Vinyals 等，2016）和 tieredImageNet（Ren 等，2018）数据集都基于 ImageNet ILSVRC-2012（Deng 等，2009）构建，是小样本图像分类中使用最广泛的基准数据集。miniImageNet包含100个类别、60000张图像，图像分辨率为84×84。数据集按标准划分为64个训练集类别、16个验证集类别和20个测试集类别。由于类别覆盖范围广且图像内容多样，miniImageNet在小样本设置下具有较高的识别难度。tieredImageNet规模更大，包含608个类别与约779165张图像，并按语义层级组织为34个大类。训练集、验证集与测试集分别由互不重叠的大类构成，从而在语义层面实现严格隔离。该划分方式显著提升了跨类别迁移与泛化的挑战性，更

适用于评估模型在语义分布差异较大的新类场景中的鲁棒性。

CIFAR-FS(Lee等,2019)和FC100(Oreshkin等,2018)数据集都基于CIFAR-100构建,CIFAR-FS包含100个类别、60000张32×32像素图像,划分为64个训练集类别、16个验证集类别和20个测试集类别,该数据集尺寸较小、训练效率较高,常用于小样本方法的快速验证与对比评估。FC100将CIFAR-100的100个类别按20个超类进行划分,训练集使用12个超类,测试集类别与训练集在超类层面保持较强隔离。降低了训练与测试类别之间的语义重叠程度,对模型的跨语义泛化能力提出更严格要求,常用于检验小样本方法在细粒度语义迁移方面的有效性。

3.2 实验细节

本实验采用Visformer-Tiny(Chen等,2021)作为视觉特征提取器,并配合随机裁剪和颜色抖动等标准图像增强方法。输入图像的默认尺寸为224×224像素。文本编码器部分采用预训练的CLIP(Radford等,2021)模型,仅使用其中的文本编码器,并通过模板"a photo of a {class name}"对类别名称进行扩展。在文本语义构建过程中,本文采用GPT-4作为大语言模型,离线完成类别多角度细粒度语义描述的生成、语义精炼与整合压缩。此外,本文对基于提示词工程生成并经语义精炼后的细粒度语义文本进行编码,利用对比语言-图像预训练(contrastive language-image pre-training, CLIP)(Radford等,2021)模型将其转换为文本特征,作为视觉-文本对齐的辅助信息。在预训练阶段,使用AdamW(Loshchilov等,2017)优化器,学习率设为5e-4、权重衰减设为5e-2。在miniImageNet、CIFAR-FS、FC100数据集上进行了800轮训练,tieredImageNet数据集上进行了400轮训练。元训练阶段继续采用相同的优化配置,但将视觉特征提取器学习率降低至1e-6,多分支对齐模块学习率设置为5e-4,并在所有数据集上进行了100轮训练。在评估过程中,从新类别中随机采样2000个测试episode。全部实验在单张RTX3090 GPU上完成,训练过程中使用了自适应学习率调度器动态调整学习率。最后,结果报告各阶段模型的准确率及其95%置信区间。

3.3 主要结果

表1和表2分别汇总了本文所提出的TSMA-Net在四个基准数据集上5-way 1-shot和5-way 5-shot小样本学习任务中的性能表现,并将其与当前多种代表性FSL方法进行了对比。对比方法既包括仅依赖视觉信息的单模态方法,也包括引入语义信息的多模态方法。实验结果表明,本文方法在四个数据集上均取得了优异的分类性能,尤其在更具挑战性的5-way 1-shot设置下表现更为突出。例如,在miniImageNet数据集上,本文方法相较于采用相同骨干网络的SimpleFSL在1-shot和5-shot任务上分别提升2.11%和0.47%;在tieredImageNet数据集上分别提升2.35%和0.49%;在CIFAR-FS数据集上分别提升1.21%和0.28%;在FC100数据集上分别提升0.54%和0.10%。

本文方法性能提升的关键在于图文语义关联与多分支对齐的协同作用。一方面,语义信息挖掘模块利用提示词工程和语义精炼策略,从类别名称出发生成更具细粒度特征和类别判别性的语义文本,并进一步与类别名称语义进行融合,从而获得更加稳健且更具区分能力的类别文本表示;另一方面,多分支对齐模块将融合语义映射到多个独立子空间,与视觉特征进行深度交互和对齐,有效增强了跨模态类别原型的表征能力。与SimpleFSL中主要采用加性跨模态融合的方式相比,TSMA-Net在语义表示构建和跨模态对齐两个层面均进行了更细致的建模,因此能够更充分地发挥语义信息在小样本场景中的辅助作用。

从语义增强方法内部的比较来看,本文方法相较于SP-CLIP和SimpleFSL的提升更能体现多粒度语义构建与深度对齐机制的有效性。SP-CLIP主要通过语义提示将类别文本信息注入视觉特征,其语义来源仍以类别名称及提示语义为主。相比之下,TSMA-Net进一步引入由大语言模型生成并经语义精炼的多角度细粒度描述,使类别文本表示不仅包含稳定的类别中心概念,还包含局部属性、外观细节和类间差异线索。因此,在相同Visformer-Tiny骨干网络下,本文方法相较SP-CLIP在miniImageNet、tieredImageNet、CIFAR-FS和FC100的1-shot任务上分别提升4.24%、6.90%、5.01%和2.68%,说明多粒度语义信息能够在极低样本场景下提供更充分的判别先验。进一步地,SimpleFSL已经采用可学习提示、跨模态融合和自蒸馏策略,是较强的语义增强基线。与SimpleFSL相比,本文方法仍在四个数据集

上取得稳定提升，尤其在miniImageNet和tieredImageNet的1-shot任务上分别提升2.11%和2.35%。这表明，仅引入语义信息并进行简单融合仍不足以充分发挥语言先验的作用，而通过多分支子空间分解、深层交互和残差重校准，可以更有效地缓解视觉—语义模态鸿沟，增强语义增强原型的判别能力。

进一步来看，本文方法在5-way 1-shot任务中的优势更为明显。这表明在样本极度稀缺的情况下，细粒度语义信息能够为类别原型提供更可靠的先验补充，从而有效弥补视觉监督不足带来的表征缺失。相比之下，在5-shot条件下，由于支持集中的视觉信息更加充分，语义信息带来的相对增益有所减弱，但仍能够稳定提升性能。这一现象说明，语义信息挖掘模块与多分支对齐模块在低样本条件下具有更高的重要性，也进一步验证了本文方法在极低样本场景中的有效性与泛化能力。

表1　miniImageNet和tieredImageNet上的结果(%)

Table 1　Results (%) on miniImageNet and tieredImageNet

	Methods	Backbone	miniImageNet		tieredImageNet	
			5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
Visual	MAML(Finn等,2017)	ResNet-12	58.05 ± 0.10	72.41 ± 0.20	63.85 ± 0.76	81.57 ± 0.56
	Meta-AdaM(Sun等,2023)	ResNet-12	59.89 ± 0.49	77.92 ± 0.43	65.31 ± 0.48	85.24 ± 0.35
	SetFeat(Afrasiyabi等,2022)	ResNet-12	68.32 ± 0.62	82.71 ± 0.46	73.63 ± 0.88	87.59 ± 0.57
	PBML(Fu等,2024)	ResNet-12	63.60 ± 0.70	81.94 ± 0.44	70.64 ± 0.72	85.39 ± 0.40
	ProtoNet(Snell等,2017)	ResNet-12	60.34 ± 1.20	80.54 ± 1.13	69.63 ± 0.53	84.82 ± 0.36
	DeepEMD(Zhang等,2020)	ResNet-12	65.91 ± 0.82	82.41 ± 0.56	71.16 ± 0.87	86.03 ± 0.58
	PRSN(Dong等,2025)	ConvNet	62.50 ± 0.86	75.73 ± 0.58	64.56 ± 0.84	79.11 ± 0.72
	Align(Afrasiyabi等,2020)	Wide-ResNet	65.92 ± 0.60	82.85 ± 0.55	74.40 ± 0.68	86.61 ± 0.59
	CORL(He等,2023)	ResNet-12	65.74 ± 0.53	83.03 ± 0.33	73.82 ± 0.58	86.76 ± 0.53
	LSFSL(Padmanabhan等,2023)	ResNet-12	64.67 ± 0.49	81.79 ± 0.18	71.17 ± 0.52	86.23 ± 0.22
	SSL-ProtoNet(Lim等,2024)	ConvNet	52.58 ± 0.45	70.87 ± 0.36	55.14 ± 0.49	74.23 ± 0.40
	RFS(Tian等,2020)	ResNet-12	62.02 ± 0.63	79.64 ± 0.44	71.52 ± 0.69	86.03 ± 0.49
	MixtFSL(Afrasiyabi等,2021)	ResNet-12	63.98 ± 0.79	82.04 ± 0.49	70.97 ± 1.03	86.16 ± 0.67
	MetaDiff(Zhang等,2024)	ResNet-12	64.99 ± 0.77	81.21 ± 0.56	72.33 ± 0.92	86.31 ± 0.62
	CADS(Zhang等,2024)	ResNet-12	66.56 ± 0.19	82.74 ± 0.13	72.04 ± 0.22	86.47 ± 0.15
Semantic	KTN(Peng等,2019)	Conv-128	64.42 ± 0.72	74.16 ± 0.56	74.16 ± 0.56	-

表1续表

Methods	Backbone	miniImageNet		tieredImageNet	
		5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
AM3(Xing等,2019)	ResNet-12	65.30 ± 0.49	78.10 ± 0.36	69.08 ± 0.47	82.58 ± 0.31
TRAML(Li等,2020)	ResNet-12	67.10 ± 0.52	79.54 ± 0.60	-	-
CMGNN-DPGN(Liu等,2021)	ResNet-12	71.38 ± 0.51	82.60 ± 0.47	72.89 ± 0.49	84.92 ± 0.48
LPE-CLIP(Yang等,2023)	ResNet-12	71.64 ± 0.40	79.67 ± 0.32	73.88 ± 0.48	84.88 ± 0.36
SP-CLIP*(Chen等,2023)	Visformer-T	73.19 ± 0.72	82.39 ± 0.56	75.66 ± 0.91	87.25 ± 0.63
SimpleFSL*(Zhou等,2025)	Visformer-T	75.32 ± 0.37	82.75 ± 0.31	80.21 ± 0.43	87.83 ± 0.32
本文	Visformer-T	**77.43 0.35**	**83.22 0.29**	**82.56 0.40**	**88.32 0.31**

注:*表示相关实验基于其公开代码重新实现,±显示95%的可信区间,-表示原文未报告该项结果或该项不适用,黑体表示最优结果。

3.4 模型分析

3.4.1 消融实验

为验证本文方法各组成部分的有效性,在miniImageNet和CIFAR-FS数据集上进行了消融实验,结果如表3所示。不同于仅从模块层面进行消融,本文进一步将语义信息划分为类别名称语义和多角度细粒度语义两部分,并结合多分支对齐模块进行分析,以考察粗粒度语义与细粒度语义在小样本分类中的作用及其互补关系。

仅使用视觉骨干网络时,模型在miniImageNet数据集上的5-way 1-shot和5-way 5-shot准确率分别为64.08%和80.45%,在CIFAR-FS数据集上分别为71.54%和86.22%。说明在样本极为有限的条件下,仅依赖视觉信息难以形成稳定且具有充分判别性的类别表征。

在引入类别名称语义并结合多分支对齐模块后,模型性能得到明显提升。在miniImageNet数据集上,1-shot和5-shot准确率分别达到76.46%和83.08%;在CIFAR-FS数据集上分别达到85.55%和88.86%。该结果表明,类别名称语义虽然属于粗粒度语义,但能够提供相对稳定的类别中心概念,从而增强类别原型的语义约束能力。

当引入多角度细粒度语义并结合多分支对齐模块时,模型性能进一步提升,在miniImageNet数据集上的准确率达到77.29%和83.16%,在CIFAR-FS数据集上达到86.12%和88.92%,整体优于仅使用类别名称语义的设置。说明多角度细粒度语义能够更充分地补充类别的局部属性、外观细节及类间差异等判别性信息,从而提升小样本场景下的类别区分能力。

完整模型在同时融合类别名称语义、多角度细粒度语义和多分支对齐模块后取得了最优结果。在miniImageNet数据集上,1-shot和5-shot准确率分别达到77.43%和83.22%;在CIFAR-FS数据集上分别达到86.50%和89.09%。这一结果表明,类别名称语义与多角度细粒度语义之间具有良好的互补性:前者提供较稳定的类别整体语义,后者补充更具区分性的细节语义,二者在多分支对齐模块作用下实现有效协同,从而共同提升了小样本场景下类别原型的鲁棒性与判别能力。

这一结论也可以从图4的融合因子分析中得到进一步验证。随着融合因子a从0增大到1,模型性能并非单调提升,而是呈现先上升后下降的趋势,并在$a = 0.4$附近达到最优。这说明类别名称语义和多角度细粒度语义并不是简单的替代关系。类别名称语义能够提供稳定的类别中心约束,避免生成式

表2　CIFAR-FS和FC100上的结果(%)

Table 2　Results (%) on CIFAR-FS and FC100

Methods	Backbone	CIFAR-FS		FC100	
		5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
Meta-AdaM（Sun 等，2023）	ResNet-12	-	-	41.12 ± 0.49	56.14 ± 0.49
PBML(Fu等,2024)	ResNet-12	73.07 ± 0.59	85.51 ± 0.41	47.92 ± 0.49	**62.96　0.51**
ProtoNet（Snell 等，2017）	ResNet-12	72.20 ± 0.70	83.50 ± 0.50	37.50 ± 0.60	52.50 ± 0.60
PRSN(Dong等,2025)	ConvNet	66.28 ± 0.68	80.10 ± 0.56	42.55 ± 0.43	53.43 ± 0.40
Align（Afrasiyabi 等，2020）	Wide-ResNet	-	-	45.83 ± 0.48	59.74 ± 0.56
CORL(He等,2023)	ResNet-12	74.13 ± 0.71	87.54 ± 0.51	44.82 ± 0.73	61.31 ± 0.54
LSFSL（Padmanabhan等,2023）	ResNet-12	73.45 ± 0.27	87.07 ± 0.17	43.60 ± 0.11	60.12 ± 0.17
SSL-ProtoNet（Lim 等，2024）	ConvNet	60.41 ± 0.52	76.52 ± 0.38	-	-
RFS(Tian等,2020)	ResNet-12	73.90 ± 0.80	86.90 ± 0.50	44.60 ± 0.70	60.90 ± 0.60
CADS(Zhang等,2024)	ResNet-12	73.23 ± 0.21	87.67 ± 0.14	-	-
LPE-CLIP（Yang 等，2023）	ResNet-12	80.62 ± 0.41	86.22 ± 0.33	-	-
SP-CLIP*（Chen 等，2023）	Visformer-T	81.49 ± 0.75	88.28 ± 0.57	46.32 ± 0.72	59.43 ± 0.71
SimpleFSL*（Zhou 等，2025）	Visformer-T	85.29 ± 0.36	88.81 ± 0.31	48.46 ± 0.38	59.72 ± 0.39
本文	Visformer-T	**86.50　0.34**	**89.09　0.30**	**49.00　0.38**	59.82 ± 0.39

注：*表示相关实验基于其公开代码重新实现，±显示95%的可信区间，-表示原文未报告该项结果或该项不适用，黑体表示最优结果。

表3　在1-shot和5-shot设置下，miniImageNet和CIFAR-FS数据集上不同模型变体的消融实验结果

Table 3　Ablation results of different model variants on miniImageNet and CIFAR-FS under the 1-shot and 5-shot settings

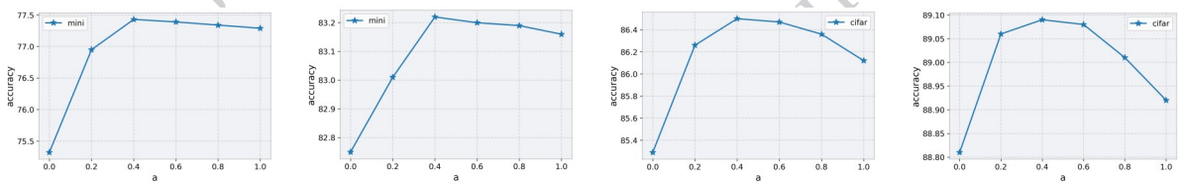
类别名称语义	多角度细粒度语义	多分支对齐模块	miniImageNet		CIFAR-FS	
			5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
			64.08 ± 0.85	80.45 ± 0.57	71.54 ± 0.87	86.22 ± 0.59
√		√	76.46 ± 0.37	83.08 ± 0.29	85.55 ± 0.36	88.86 ± 0.30
	√	√	77.29 ± 0.34	83.16 ± 0.29	86.12 ± 0.36	88.92 ± 0.29
√	√	√	**77.43　0.35**	**83.22　0.29**	**86.50　0.34**	**89.09　0.30**

注：√表示采用对应语义信息或模块，类别名称语义和多角度细粒度语义分别对应粗粒度语义和细粒度语义，黑体表示最优结果。

细粒度文本中的开放语义噪声占据主导；多角度细粒度语义则能够补充类名难以表达的局部属性和差异线索，增强类别间判别能力。因此，适度融合两类语义能够在语义稳定性与细节判别性之间取得平衡，这也解释了表3中完整模型优于仅使用单一语义来源的模型变体。

3.4.2　融合因子a的影响

为量化多角度细粒度语义表示与类别名称语义在统一语义表示中的相对贡献，本文对融合公式中的融合因子a进行敏感性分析。图4给出了不同a取值下模型在miniImageNet与CIFAR-FS的1-shot/5-shot任务表现。当a从0增加到1时，模型准确率呈先上升后下降的趋势，并在$a = 0.4$时达到最高。例如，miniImageNet 5-way 1-shot的准确率在$a = 0.4$时取得峰值77.43%，然后在$a > 0.4$时略有降低。该现象表明，适度引入细粒度语义能有效补充类名语义的粗粒度缺陷，为跨模态对齐提供更具判别性的视觉属性线索，从而提升原型表征质量；而当a过大时，细粒度语义的高信息密度可能在语义空间中占据主导，带来描述冗余或噪声干扰，进而削弱与视觉特征的稳定对齐。相反，a过小会限制细粒度语义的贡献，使模型更依赖类名语义，难以充分缓解类名歧义与细粒度差异不足的问题。因此，选择适中的融合因子$a = 0.4$可以在语义信息的稳定性和判别性之间取得良好平衡，实现最佳的融合效果。本模型也据此将$a = 0.4$作为经验最优设置。



（a）　miniImageNet数据集1-shot结果　（b）　miniImageNet数据集5-shot结果　（c）　CIFAR-FS数据集1-shot结果　（d）　CIFAR-FS数据集5-shot结果

（a）　1-shot on miniImageNet　（b）　5-shot on miniImageNet　（c）　1-shot on CIFAR-FS　（d）　5-shot on CIFAR-FS

图4　不同融合因子a对miniImageNet，CIFAR-FS的影响

Figure 4　Effect of different fusion factors a on miniImageNet and CIFAR-FS

3.4.3　线性投影数量的影响

多分支对齐模块中线性投影的数量也会影响模型性能。本文在不同投影分支个数下评估了模型，在miniImageNet 5-way 1-shot上的结果如表4所示：当线性投影分支数由1增加到4时，模型性能整体呈上升趋势，说明适当增加投影分支有助于将融合语义表示分解到多个独立子空间中，从而捕获不同语义成分与视觉特征之间的对应关系。分支数较少时，模型只能在有限子空间内进行语义变换和跨模态交互，难以充分表达多角度细粒度语义中包含的外观属性、局部结构和类间差异信息，因此表征能力受到限制。当分支数继续增加到6时，性能反而下降。这可能是因为过多分支会导致每个子空间分配到的有效特征维度减小，使语义表示过度碎片化；同时，额外分支也会引入冗余参数和重复语义响应，在小样本episode训练条件下更容易放大跨模态噪声或产生过拟合，从而削弱模型泛化能力。综合来看，4个投影分支在语义表达能力、参数复杂度和跨模态对齐稳定性之间取得了较好平衡，因此本文将其作为默认设置。

3.4.4　有无残差重校准的影响

在多分支对齐模块中，引入瓶颈式残差重校准结构对初步跨模态融合特征进行再校准，以抑制由模态差异带来的噪声并强化有效语义成分。为验证该结构的贡献，在保持其余配置不变的条件下，对比“有/无残差重校准”两种设置，结果如表5所

表4　不同分支数量设置下模型在miniImageNet和CIFAR-FS数据集上的性能对比结果

Table 4　Performance comparison of the model under different numbers of branches on miniImageNet and CIFAR-FS

分支数量	miniImageNet		CIFAR-FS	
	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
1	77.13	83.11	86.48	89.03
2	77.09	83.19	86.43	89.08
3	77.37	83.08	86.39	89.05
4	**77.43**	**83.22**	**86.50**	**89.09**
6	76.91	83.12	86.44	89.03

注：黑体表示最优结果。

示。实验表明，残差重校准带来小幅但一致的性能改进，说明其在抑制融合噪声方面具有一定辅助作用。在miniImageNet 5-way 1-shot上，加入残差重校准后准确率为77.43%，相比不使用该模块时的77.04%有所提升；在CIFAR-FS上也由86.46%提升至86.50%，总体呈现小幅但可复现的提升趋势。可见，残差重校准通过瓶颈降维与升维操作自适应地重新分配了融合特征的通道权重，有效抑制了跨模态融合过程中的噪声并突出了关键语义信息。与未经过重校准的特征相比，加入该模块后得到的语义增强原型在视觉模态和语言模态之间的一致性更高，判别能力更强，因而能够带来更优的分类性能。

表5　是否引入残差重校准结构的消融实验结果

Table 5　Ablation results on whether to introduce the residual recalibration structure

残差重校准	miniImageNet		CIFAR-FS	
	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
有	**77.43**	**83.22**	**86.50**	**89.09**
无	77.04	83.20	86.46	89.07

注：黑体表示最优结果。

3.4.5　模型复杂度分析

为进一步评估所提出模块带来的计算开销，本文在相同硬件环境下统计SimpleFSL与TSMA-Net在5-way 1-shot设置下的参数量、FLOPs和平均推理时间。结果如表6所示。由于两种方法均采用Visformer-Tiny作为视觉骨干网络，其主要计算量均来自support/query图像的特征提取，因此二者FLOPs基本一致。与SimpleFSL相比，TSMA-Net的参数量由10.050M增加至10.383M，仅增加约3.31%，说明本文引入的多分支对齐和残差重校准结构具有较轻量的参数规模。

在推理时间方面，TSMA-Net由于引入多角度语义融合、多分支映射和残差重校准操作，平均推理时间由51.571ms/episode增加至100.804ms/episode。尽管推理时间有所上升，但TSMA-Net在miniImageNet 5-way 1-shot任务上获得了2.11%的准确率提升，表明该方法能够在可接受的复杂度增加下换取更稳定的分类性能。尤其在1-shot场景中，额外语义对齐操作能够有效补充支持样本不足带来的原型不稳定问题，因此体现出较好的性能—复杂度折中。

表6　SimpleFSL与TSMA-Net的模型复杂度对比

Table 6　Complexity comparison between SimpleFSL and TSMA-Net

方法	Params/M	FLOPs/G	Time/ms	Acc/%
SimpleFSL	10.050	101.623	51.571 ± 15	75.32
TSMA-Net	10.383	101.624	100.804 ± 16	77.43

注：Params、FLOPs和推理时间均按1-shot episode统计。每个episode包含5张support图像和75张query图像。CLIP文本编码和GPT-4语义生成均为离线过程，不计入在线推理时间。

3.5　可视化分析

如图5所示，本文采用t-SNE对SimpleFSL与TSMA-Net的特征嵌入进行可视化对比。在200-shot设置下，SimpleFSL的特征分布更为离散，存在一定类间重叠；相比之下，TSMA-Net形成了更紧凑的类内聚集与更清晰的类间分离，表明图文语义关联与多分支对齐能够有效提升跨模态对齐质量并增强特征判别性。

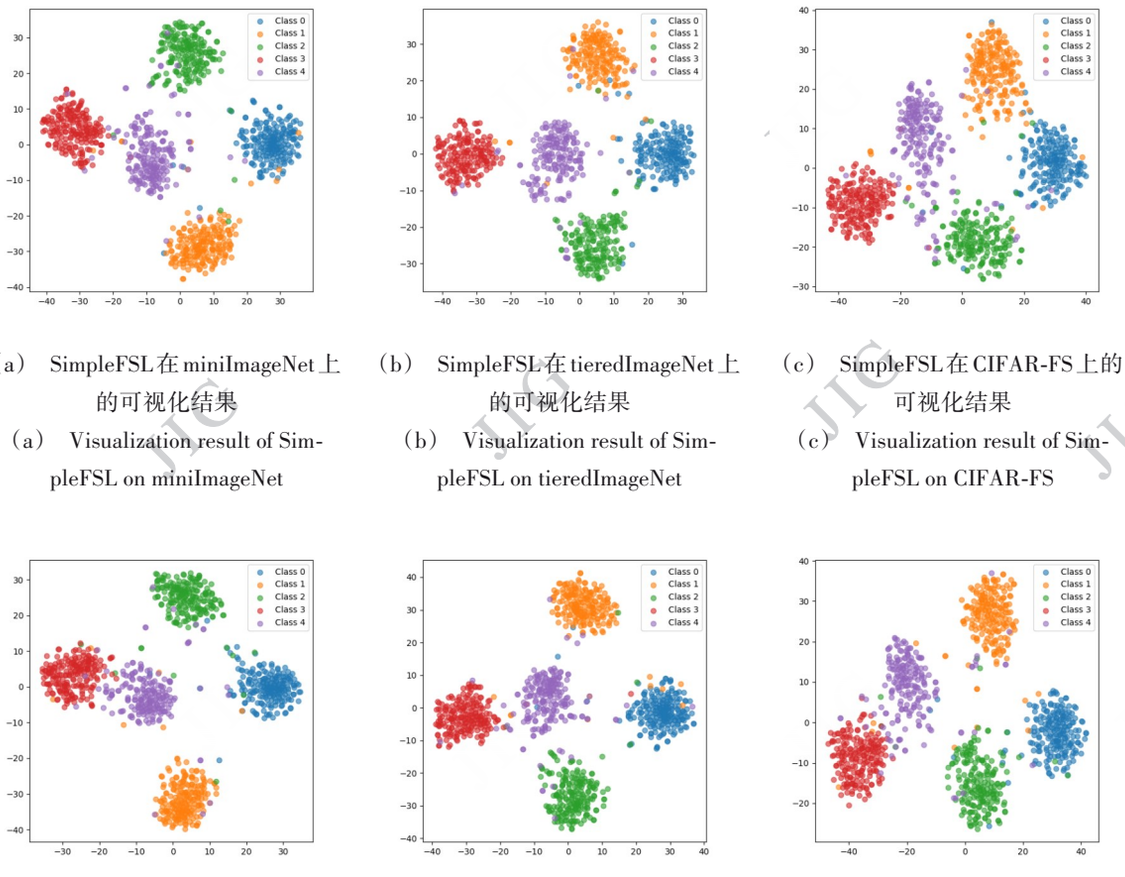
4　结　论

本文提出了TSMA-Net，一种面向小样本图像分类的图文语义多分支对齐网络。该方法从视觉与文本两个模态协同建模：首先，通过视觉骨干网络预训练获得稳定且具有判别性的视觉特征表示；其次，引入语义信息挖掘模块，以类别名称为语义锚点，利用提示词工程和语义精炼策略生成细粒度语义文本，并通过语义适配将其与类别名称语义进行融合，从而构建更加稳健且更具区分能力的类别文本表示；最后，通过多分支对齐模块将融合语义映射到多个独立子空间，与视觉特征进行深度交互和对齐，并结合残差重校准结构进一步提升跨模态特征表示的鲁棒性与判别性。

实验结果表明，TSMA-Net在miniImageNet、tieredImageNet、CIFAR-FS和FC100四个标准小样本学习基准数据集上均取得了稳定而显著的性能提升，尤其在5-way 1-shot任务中优势更为明显。这说明本文方法能够在极低样本条件下有效利用多粒度语义信息弥补视觉监督不足，从而增强类别原型的表征能力并提升最终分类性能。消融实验和参数敏感

性分析进一步验证了语义信息挖掘模块、多分支对齐模块以及融合策略的有效性。

尽管本文方法已经取得了较好的实验结果，但仍有若干方向值得进一步探索。首先，在语义信息挖掘方面，可以进一步研究更具任务适应性的提示词构造与筛选策略，以提升细粒度语义生成的稳定性和针对性。其次，在跨模态融合方面，可以在多分支对齐的基础上引入更强的交互建模机制，例如交叉注意力或动态门控策略，以增强语义信息在不同子空间中的自适应分配能力。未来，如何在保持模型稳定性的同时进一步提升语义质量与对齐效率，将是值得持续研究的重要方向。



（a） SimpleFSL在miniImageNet上的可视化结果
（a） Visualization result of SimpleFSL on miniImageNet

（b） SimpleFSL在tieredImageNet上的可视化结果
（b） Visualization result of SimpleFSL on tieredImageNet

（c） SimpleFSL在CIFAR-FS上的可视化结果
（c） Visualization result of SimpleFSL on CIFAR-FS

（d） TSMA-Net在miniImageNet上的可视化结果
（d） Visualization result of TSMA-Net on miniImageNet

（e） TSMA-Net在tieredImageNet上的可视化结果
（e） Visualization result of TSMA-Net on tieredImageNet

（f） TSMA-Net在CIFAR-FS上的可视化结果
（f） Visualization result of TSMA-Net on CIFAR-FS

图5 SimpleFSL（上）与TSMA-Net（下）在miniImageNet、tieredImageNet与CIFAR-FS上的t-SNE可视化对比

Figure 5 t-SNE visualization comparison between SimpleFSL (top) and TSMA-Net (bottom) on miniImageNet, tieredImageNet, and CIFAR-FS

参考文献（References）

Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman F L, et al. 2023. GPT-4 technical report[EB/OL]. [2026-03-12]. https://arxiv.org/abs/2303.08774

Afrasiyabi A, Lalonde J F and Gagné C. 2020. Associative alignment for few-shot image classification//Proceedings of the European Conference on Computer Vision. Cham: Springer: 18-35[DOI: 10.1007/978-3-030-58558-7_2]

Afrasiyabi A, Lalonde J F and Gagné C. 2021. Mixture-based feature space learning for few-shot image classification//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE: 9021-9031[DOI: 10.1109/ICCV48922.2021.00891]

Afrasiyabi A, Larochelle H, Lalonde J F and Gagné C. 2022. Matching feature sets for few-shot image classification//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE: 9014-9024 [DOI: 10.1109/CVPR52688.2022.00881]

Chen S M, Duan B, Khan S and Khan F S. 2025. Interpretable zero-shot learning with locally-aligned vision-language model//Proceedings of the IEEE/CVF International Conference on Computer Vision. Honolulu: IEEE: 478-487

Chen W, Si C, Zhang Z, Wang L, Wang Z and Tan T. 2023. Semantic prompt for few-shot image recognition[EB/OL]. [2026-03-12]. https://arxiv.org/abs/2303.14123

Chen Z, Xie L, Niu J, Liu X, Wei L and Tian Q. 2021. Visformer: the vision-friendly transformer//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE: 589-598 [DOI: 10.1109/ICCV48922.2021.00063]

Deng J, Dong W, Socher R, Li L J, Li K and Fei-Fei L. 2009. ImageNet: A large-scale hierarchical image database//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE: 248-255 [DOI: 10.1109/CVPR.2009.5206848]

Dong M, Li F, Li Z and Liu X. 2025. PRSN: prototype resynthesis network with cross-image semantic alignment for few-shot image classification. Pattern Recognition, 159: 111122[DOI: 10.1016/j.patcog.2024.111122]

Dong Y Y, Song B B and Sun W F. 2023. Local feature fusion network-based few-shot image classification. Journal of Image and Graphics, 28(7): 2093-2104 (董杨洋,宋蓓蓓,孙文方. 2023. 局部特征融合的小样本分类. 中国图象图形学报,28(7): 2093-2104) [DOI: 10.11834/jig.220079]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. 2020. An image is worth 16×16 words: Transformers for image recognition at scale[EB/OL]. [2026-03-12]. https://arxiv.org/abs/2010.11929

Finn C, Abbeel P and Levine S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks//Proceedings of the 34th International Conference on Machine Learning. Sydney: PMLR: 1126-1135

Fu M, Wang X, Wang J and Yi Z. 2024. Prototype bayesian meta-learning for few-shot image classification. IEEE Transactions on Neural Networks and Learning Systems, 36(4): 7010-7024 [DOI: 10.1109/TNNLS.2024.3403865]

He J, Kortylewski A and Yuille A. 2023. CORL: compositional representation learning for few-shot classification//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa: IEEE: 3879-3888 [DOI: 10.1109/WACV56688.2023.00388]

He K, Zhang X, Ren S and Sun J. 2016. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

He X J and Lin J F. 2022. Weakly-supervised object localization based fine-grained few-shot learning. Journal of Image and Graphics, 27(7): 2226-2239 (贺小箭,林金福. 2022. 融合弱监督目标定位的细粒度小样本学习. 中国图象图形学报,27(7): 2226-2239) [DOI: 10.11834/jig.200849]

Jiang H J, Li Z X, Yu X H, Hu Y L, Yin B C, Yang J and Qi Y K. 2025. Visual and semantic prompt collaboration for generalized zero-shot learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE: 20275-20285 [DOI: 10.1109/CVPR52734.2025.01888]

Lake B, Salakhutdinov R, Gross J and Tenenbaum J. 2011. One shot learning of simple visual concepts//Proceedings of the Annual Meeting of the Cognitive Science Society, 33. Boston: The Cognitive Science Society: 2568-2573

LeCun Y, Bengio Y and Hinton G. 2015. Deep learning. Nature, 521(7553): 436-444 [DOI: 10.1038/nature14539]

Lee K, Maji S, Ravichandran A and Soatto S. 2019. Meta-learning with differentiable convex optimization//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE: 10657-10665 [DOI: 10.1109/CVPR.2019.01091]

Li A, Huang W, Lan X, Feng J, Li Z and Wang L. 2020. Boosting few-shot learning with adaptive margin loss//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE: 12573-12581 [DOI: 10.1109/CVPR42600.2020.01259]

Lim J Y, Lim K M, Lee C P and Tan Y X. 2024. SSL-ProtoNet: self-supervised learning prototypical networks for few-shot learning. Expert Systems with Applications, 238: 122173[DOI: 10.1016/j.eswa.2023.122173]

Liu S, Xie Y, Yuan W and Ma L. 2021. Cross-modality graph neural network for few-shot learning//Proceedings of the 2021 IEEE International Conference on Multimedia and Expo. Shenzhen: IEEE: 1-6 [DOI: 10.1109/ICME51207.2021.9428405]

Loshchilov I and Hutter F. 2017. Decoupled weight decay regularization [EB/OL]. [2026-03-12]. https://arxiv.org/abs/1711.05101

Oreshkin B, Rodríguez López P and Lacoste A. 2018. TADAM: task dependent adaptive metric for improved few-shot learning// Advances in Neural Information Processing Systems, 31. Montréal, Canada: Curran Associates Inc [DOI: 10.48550/arXiv. 1805.10123]

Padmanabhan D C, Gowda S, Arani E and Zonooz B. 2023. LSFSL: leveraging shape information in few-shot learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. BC, Canada: IEEE: 4971-4980[DOI: 10.1109/CVPRW59228.2023.00525]

Peng Z, Li Z, Zhang J, Li Y, Qi GJ and Tang J. 2019. Few-shot image recognition with knowledge transfer//Proceedings of the IEEE/CVF

International Conference on Computer Vision. Seoul: IEEE: 441-449 [DOI: 10.1109/ICCV.2019.00053]

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 Event: PMLR: 8748-8763 [DOI: 10.48550/arXiv.2103.00020]

Ren M, Triantafillou E, Ravi S, Snell J, Swersky K, Tenenbaum J B, et al. 2018. Meta-learning for semi-supervised few-shot classification[EB/OL]. [2026-03-12]. https://arxiv.org/abs/1803.00676

Snell J, Swersky K and Zemel R. 2017. Prototypical networks for few-shot learning//Advances in Neural Information Processing Systems, 30. Long Beach, USA: Curran Associates Inc: 4080-4090

Sun S and Gao H. 2023. Meta-AdaM: an meta-learned adaptive optimizer with momentum for few-shot learning//Advances in Neural Information Processing Systems, 36. New Orleans, USA: Neural Information Processing Systems Foundation, Inc.: 65441-65455 [DOI: 10.52202/075280-2855]

Tian Y, Wang Y, Krishnan D, Tenenbaum J B and Isola P. 2020. Rethinking few-shot image classification: a good embedding is all you need//Proceedings of the European Conference on Computer Vision. Cham: Springer: 266-282 [DOI: 10.1007/978-3-030-58568-6_16]

Vinyals O, Blundell C, Lillicrap T, Kavukcuoglu K and Wierstra D. 2016. Matching networks for one shot learning//Advances in Neural Information Processing Systems, 29. Barcelona, Spain: Curran Associates Inc: 3630-3638

Wang X S, Lyu L X, Cheng Y H and Wang H Y. 2024. Attention set representation for multiscale measurement of few-shot image classification. Journal of Image and Graphics, 29(11): 3371-3382 (王雪松,吕理想,程玉虎,王浩宇. 2024. 注意力集合表示的多尺度度量小样本图像分类. 中国图象图形学报,29(11): 3371-3382) [DOI: 10.11834/jig.230763]

Xing C, Rostamzadeh N, Oreshkin B and Pinheiro P O. 2019. Adaptive cross-modal few-shot learning//Advances in Neural Information Processing Systems, 32. Vancouver, Canada: Curran Associates Inc: 4847-4857

Yang F Y, Wang R P and Chen X L. 2023. Semantic guided latent parts embedding for few-shot learning//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa: IEEE: 5447-5457 [DOI: 10.1109/WACV56688.2023.00541]

Zhang B, Luo C, Yu D, Li X, Lin H, Ye Y and Zhang B. 2024. MetaDiff: Meta-learning with conditional diffusion for few-shot learning//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver: AAAI: 16687-16695 [DOI: 10.1609/aaai.v38i15.29608]

Zhang C, Cai Y, Lin G and Shen C. 2020. DeepEMD: few-shot image classification with differentiable earth mover's distance and structured classifiers//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE: 12203-12213 [DOI: 10.1109/CVPR42600.2020.01222]

Zhang Y, Gong M, Li J, Feng K and Zhang M. 2024. Few-shot learning with enhancements to data augmentation and feature extraction. IEEE Transactions on Neural Networks and Learning Systems, 36(4): 6655-6668 [DOI: 10.1109/TNNLS.2024.3400592]

Zhou C, Yu Z, Yuan X, Zhou S, Bu J and Wang H. 2025. Less is more: a closer look at semantic-based few-shot learning. Information Fusion, 114: 102672[DOI: 10.1016/j.inffus.2024.102672]

Zhou K, Yang J, Loy C C and Liu Z. 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

王进,男,副教授,主要研究方向为人工智能与计算机视觉。E-mail: wj@ntu.edu.cn

丁新,男,副教授,主要研究方向为人工智能、图像处理和多视角合成。E-mail:xding@163.com

杜欣豫,男,硕士研究生,主要研究方向为小样本学习。E-mail:2441320039@stmail.ntu.edu.cn