中图法分类号： **文献标识码：** **文章编号：** 1006-8961(XXXX)XX-0001-13

论文引用格式： Cao Silu, Liu Gaozhi, Xi Meijuan, Zhang Xinpeng, Qian Zhenxing. EthnicFashion: A Hybrid Dataset for Ethnic Clothing Classification[J/OL]. Journal of Image and Graphics, XXXX:1-13. DOI: 10.11834/jig.250613. (曹丝露, 刘高志, 奚美娟, 张新鹏, 钱振兴. EthnicFashion:用于民族服装分类的混合数据集[J/OL]. 中国图象图形学报, XXXX:1-13. DOI: 10.11834/jig.250613.)[DOI:10.11834/jig.250613]

EthnicFashion:用于民族服装分类的混合数据集

曹丝露,刘高志,奚美娟,张新鹏,钱振兴

复旦大学计算与智能创新学院,上海 200438

摘 要：目的 在文化数字化建设和民族文化保护需求不断增长的背景下,民族服饰的准确分类具有重要的研究与应用价值。传统上,民族服饰分类大多依赖专业人员的人工处理,这种方式成本高、耗时长且劳动强度大。因此,开发自动化的民族服饰分类方法已成为迫切需求。**方法** 本文提出了一个标注完善的民族服饰数据集 EthnicFashion,该数据集包含数千张图片,涵盖数十个民族。包括原始图像、类别标签、关键点标注以及分割掩码等相关标注信息。此外,考虑到数据稀缺和合适方法匮乏这两个该领域面临的主要挑战,我们提出将数据生成和上下文信息增强作为插件,应用于各类骨干网络,以提升分类模型的性能。具体而言,前者利用生成模型生成具有更鲜明民族风格特征的数据,用于数据增强;后者通过整合关键点检测与分割所识别的区域,实现上下文信息的增强。**结果** 在 EthnicFashion 数据集上,我们使用不同基线模型和所提方法进行了大量实验,结果表明该方法在民族服饰分类任务中表现优异。以当前表现最优的基线模型为例,引入本文方法后,Top-1 分类准确率提升 15% 以上,并在不同少样本设置下均取得稳定性能增益,显著优于原始基线模型。**结论** 本文所提出的少样本民族服饰分类方法,通过构建高质量的 EthnicFashion 数据集,并结合数据生成与上下文信息增强技术,有效缓解了数据稀缺问题,显著提升了民族服饰分类的准确性与泛化能力。数据集已上传至地址:10.57760/sciencedb.j00240.00167。

关键词： 民族服装数据集;民族服饰图像分类;少样本;生成模型;双分支网络

EthnicFashion: A Hybrid Dataset for Ethnic Clothing Classification

Cao Silu, Liu Gaozhi, Xi Meijuan, Zhang Xinpeng, Qian Zhenxing

College of Computer Science and Artificial Intelligence, Fudan University, Shanghai 200438, China

Abstract: Objective Ethnic clothing classification is a crucial yet underexplored task in the field of computer vision, with profound implications for cultural heritage preservation, digital archiving, and intelligent fashion applications. Traditionally, the classification and identification of ethnic clothing have relied heavily on manual processing by domain experts, which is both time-consuming and resource-intensive. These manual approaches require not only significant labor costs but also deep domain-specific knowledge, which is difficult to scale. Moreover, existing large-scale fashion classification datasets and models are primarily designed for modern, everyday clothing, focusing on generic attributes such as sleeve length, collar style, or overall shape. Such approaches are insufficient for ethnic clothing, which is characterized by fine-grained stylistic elements, cultural symbolism, and unique accessory combinations. There is therefore an urgent need to develop efficient, automatic, and scalable ethnic clothing classification methods that can handle data scarcity and capture rich contextual information embedded in ethnic garments. **Method** To address this gap, we propose EthnicFashion, a newly constructed, well-annotated dataset that contains 3,800 high-quality images representing 40 distinct ethnic groups, spanning

收稿日期： 2025-12-04;**修回日期：** 2026-05-08

基金项目： 国家重点研发计划(项目编号:No. 2023YFF0905000)

Supported by: the National Key R&D Program of China under Grant 2023YFF0905000

Chinese, other Asian, and European origins. Each image in the dataset is carefully collected, manually cleaned, and labeled to ensure high data quality and accurate ethnic category annotations. The dataset includes detailed cultural and stylistic diversity, such as characteristic headpieces, patterns and jewelries. To alleviate the problem of limited real-world data, we further introduce a data generation pipeline that leverages state-of-the-art generative diffusion models, specifically DreamBooth fine-tuned on Stable Diffusion, to synthesize new, realistic images of ethnic clothing. These generated samples enrich the diversity of training data and emphasize key ethnic style attributes, significantly mitigating the issue of data scarcity. In addition to data generation, we design a contextual information enhancement module that serves as a plugin applicable to various backbone architectures. This module unifies landmark detection and segmentation to construct more semantically meaningful representations. Landmark annotations capture structural details of garments, such as collars, cuffs, hems, and neckline corners, while segmentation focuses on accessories and larger contextual regions like headpieces and jewelries. By taking the union of these two masks, we obtain a complementary and comprehensive representation that preserves both structural accuracy and cultural details. This unified mask is applied to the original image to guide the model's attention to culturally meaningful regions, effectively reducing background noise and enhancing classification performance. To ensure balanced and stable model training, especially in few-shot scenarios, we also design a cyclic sampling strategy. Rather than mixing real and generated data randomly, which may lead to overfitting on synthetic data, the cyclic sampling method maintains a 1:1 ratio of real to generated images during training. Real images are cycled through multiple times to match the quantity of generated samples, ensuring that high-quality real data consistently influences the learning process. This strategy preserves the richness of generated data while preventing performance degradation due to synthetic noise. We evaluate our method extensively on the EthnicFashion dataset using multiple backbone models, including ResNet101, DenseNet201, Swin-Transformer V2, Tip-Adapter, and AMU-Tuning. Both conventional CNN and Transformer-based architectures, as well as few-shot adaptation methods, are examined to demonstrate the general applicability of our plugins. We use standard Top-1, Top-3, and Top-5 accuracy as evaluation metrics under various few-shot settings (2-shot, 4-shot, 6-shot, 8-shot, 10-shot). Experimental results show that integrating data generation and contextual enhancement yields substantial performance improvements across all settings. For example, on Swin-Transformer V2, Top-1 accuracy improves from 31.12% to 47.06%, and similar trends are observed for other models. These results confirm that the proposed approach is model-agnostic and can serve as a performance booster for different classification backbones. We further validate the generalization capability of our method on multiple subsets of the EthnicFashion dataset—ChineseEthnic, AsianEthnic, and EuropeanEthnic—as well as on miniDeepFashion, a subset of DeepFashion adapted to few-shot scenarios. In all cases, the proposed method consistently outperforms baselines, highlighting its ability to handle both ethnic and non-ethnic clothing classification tasks. Ablation studies further confirm the effectiveness of each component: data generation alone significantly boosts accuracy by enriching the data distribution, contextual enhancement focuses the model on key visual features, and cyclic sampling ensures stable training progress. **Result** The proposed framework significantly enhances classification accuracy, robustness, and generalization in few-shot ethnic clothing classification scenarios. By synthesizing training data using generative models, we effectively alleviate the challenge of data scarcity, which has long been a bottleneck in this field. The contextual enhancement module enables the model to better capture intricate cultural patterns and structural features, leading to more discriminative feature representations. Additionally, the cyclic sampling strategy contributes to stable and efficient training, particularly beneficial in few-shot scenarios. Our experiments demonstrate that combining these strategies leads to consistent improvements across various backbones and datasets. For example, Swin-Transformer V2 achieves Top-1 accuracy improvements of over 15% compared to the baseline, and similar gains are observed in Top-3 and Top-5 accuracy as well. **Conclusion** This work pioneers an automated, few-shot ethnic clothing classification framework that combines a high-quality, culturally diverse dataset with generative data augmentation and contextual enhancement. The construction of the EthnicFashion dataset provides a valuable resource for the research community, addressing the lack of well-annotated, ethnically diverse clothing datasets. The integration of generative modeling with diffusion techniques allows us to scale data effectively, capturing unique ethnic styles with high fidelity. Moreover, by enhancing contextual information through the union of landmark and segmentation masks, we empower classification models to focus on culturally significant visual cues. The proposed cyclic sampling strategy ensures the effective utiliza-

tion of both real and synthetic data without sacrificing generalization. Our results confirm that few-shot ethnic clothing classification can be significantly improved through synergistic use of data generation, contextual enhancement, and strategic training, achieving state-of-the-art performance across multiple architectures. This work not only contributes to advancing the field of fashion image classification but also provides a technical foundation for applications in cultural heritage preservation, intelligent recommendation systems, and cross-cultural understanding. Future work may explore the integration of multimodal signals such as textual descriptions or historical metadata, as well as extending this framework to real-time classification in cultural tourism or digital museum applications. The dataset has been uploaded to the following DOI: 10.57760/sciencedb.j00240.00167.

Key words: EthnicFashion dataset; ethnic clothing image classification; few-shot; generative model; dual branch network

公式章 1 节 1 **中图法分类号**:(此号在中国图书馆分类法中查) **文献标识码**:A **文章编号**:1006-8961(2025)-

论文引用格式:Cao Silu, Liu Gaozhi, Xi Meijuan, Zhang Xinpeng, Qian Zhenxing. EthnicFashion: A Hybrid Dataset for Ethnic Clothing Classification[J/OL]. Journal of Image and Graphics. DOI: 10.11834/jig.250613.(曹丝露,刘高志,奚美娟,张新鹏,钱振兴. EthnicFashion:用于民族服装分类的混合数据集[J/OL]. 中国图象图形学报. DOI: 10.11834/jig.250613.)

0 引 言

民族服饰是各民族文化遗产中极具价值的组成部分。它不仅承载着独特的美学价值,还传递着重要的历史与社会信息,是传统习俗、信仰及地域身份的重要象征。近年来,受虚拟博物馆、在线教育平台、文化保护项目及电子商务服务等新兴应用的推动,从图像中自动识别和分类民族服饰的需求日益增长。然而,这一任务长期以来未得到足够重视,目前主要依赖人工分类,这种方式耗时费力,且需要专业知识支撑,无法满足民族服装相关应用的现实需求。

总的来说,开发民族服饰自动分类系统面临两大核心挑战:一是缺乏覆盖多种民族服饰、标注完善的综合数据集;二是缺乏能有效应对该领域固有复杂性与数据稀缺性的定制化方法。在数据集方面,尽管目前存在一些用于服饰分类的数据集,如DeepFashion(Liu 等,2016a)、FashionMNIST(Xiao 等,2017)和FashionAI(Zou 等,2019)等,但它们主要聚焦于现代日常服饰。这些数据集通常根据上衣、连衣裙、裤子等宽泛类别,或袖长、领口类型等风格属性对服饰进行分类。虽然这些数据集推动了主流服饰分类任务的显著发展,但在应用于民族服饰分类时仍存在明显局限性。首先,它们极少覆盖传统民族服饰,而民族服饰本身具有更强的多样性和文化特异性;其次,现有数据集是针对现代服饰设计的,数据易于获取,通常每个类别都包含至少数千个样本,而民族服饰由于数据收集和标注的固有难度,难以达到这样的样本规模。

同样,在算法方面,现有服饰分类技术在通用服饰识别任务中表现出色。这些方法包括:结合全局与局部视觉特征捕捉服饰细节的方法(Liu 等,2016)、基于纹理并采用图表示法分析织物重复图案与纹理的技术(Thewsuwan & Horio,2018),以及近年来结合属性标注、融入空间和通道引导注意力机制的多模态模型(Wan 等,2022)。尽管这些方法对现代日常服饰有效,但它们严重依赖于大规模标注的数据集,因此不适用于民族服饰分类。针对这一问题,已有研究尝试通过结合人体检测与多任务学习(Wu 等,2019)或融合视觉风格与标签约束(Zhang 等,2021)等方式提升少数民族服装的识别与解析性能。而在少样本场景下,部分方法(Tang 等,2024;Zhang 等,2022)通过利用CLIP(Radford 等,2021)等大型模型的先验知识,提供了一种有潜力的解决方案。然而,我们的零样本(zero-shot)实验结果表明,现有视觉语言大模型在本领域任务中存在显著局限:CLIP 的准确率低至 15.88%,性能更强的 Qwen-VL-Plus 也仅为 27.25%。这恰恰印证了 CLIP 研究者对其模型短板的判断——当遇到预训练阶段未充分表征的概念(即处理分布外数据)时,模型性能会显著下降。这一局限性对民族服饰分类等任务构成了关键挑战,模型往往缺乏准确识别特定民族(尤其

是少数民族)服饰风格所需的领域特定知识,因此这类模型并非该任务的理想选择。如表1所示,即便是最新的少样本微调大模型的方法RAFT(Wu等,2024)、TIMO(Li等,2025)效果也不尽如人意。

表1　基于大模型的少样本微调的分类准确率(%)

Table 1　Classification accuracy(%) of few-shot fine-tuning based on large models

方法	2-shot	4-shot	6-shot	8-shot	10-shot
CLIP-finetuned	16.38	20.50	24.56	26.69	32.56
RAFT	19.75	24.56	30.00	32.25	34.69
TIMO	23.38	29.69	34.00	34.19	38.63

为克服上述局限性,我们构建了一个新的民族服饰数据集EthnicFashion,该数据集分为三个子集:中国民族服饰子集(ChineseEthnic)包含20个中国民族的2000张图片;亚洲民族服饰子集(AsianEthnic)包含中国以外10个亚洲民族的900张图片;欧洲民族服饰子集(EuropeanEthnic)包含多个欧洲民族的900张图片。数据集的构建采用了数据爬取与人工收集相结合的方式,数据来源包括百度图片①、视觉中国②、Flipkart③、亚马逊④、《时尚芭莎》杂志,以及小红书⑤等社交平台上用户上传的图片。为确保类别标签的准确性,我们对所有图片进行了人工清洗和筛选。为了进一步解决数据稀缺问题,我们设计了一个数据集自合成阶段,通过数据生成技术合成了额外样本。这种方法不仅通过增加样本数量来扩充数据集,还能让生成模型专注于捕捉民族服饰独特的文化特异性特征,提升了数据集的多样性和代表性。

如图1所示,民族服饰通常具有独特且稳定的风格特征,例如具有代表性的头饰、服饰款式、色彩,以及带有民族图腾的图案等。受这一发现启发,我们首先为图片添加关键点和分割掩码标注,然后通过整合关键点检测与分割的信息来增强上下文信息。这种整合方法既能捕捉细粒度细节,又能获取更广泛的上下文特征,显著提升了分类准确率。其中,关键点检测可精确识别服饰上的关键点,从而确定服饰区域;分割则能捕捉头饰、首饰等对民族服饰分类至关重要的重要配饰,两者形成互补效应。大量实验表明,将数据生成和上下文信息增强作为插件应用于基线模型后,基线模型在民族服饰分类任务中的性能得到了显著提升。借助所提出的新数据集,我们的研究为推进民族服饰分类领域的发展奠定了基础。

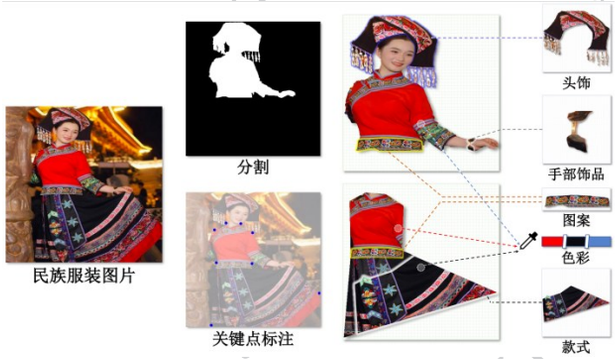

图1　民族服饰分析

Fig. 1　Analysis of ethnic clothing

本文的主要贡献总结如下:1)提出了标注完善的用于民族服饰分类的数据集EthnicFashion,涵盖40个民族,共3800张图片。同时,为解决数据稀缺问题,我们通过数据生成技术合成了额外数据,并引入了一种简单有效的循环采样策略以提升准确率。2)通过整合关键点检测与分割的信息增强上下文信息,显著提升了特征提取效果与分类准确率。3)据我们所知,本文首次聚焦于民族服饰分类的自动化研究,大量实验表明,所提方法能显著提升所有基线模型在该任务上的性能。

1　EthnicFashion数据集

EthnicFashion数据集涵盖40个不同民族的3800张高质量图片,分为三个主要子集:中国民族服饰子集(ChineseEthnic)、亚洲民族服饰子集(不含

①https://image.baidu.com
②https://www.vcg.com
③https://www.flipkart.com
④https://www.amazon.in
⑤https://www.xiaohongshu.com

表2　EthnicFashion数据集构成

Table 2　Composition of the EthnicFashion Dataset

子集名称	类别数量	典型类别名称	每类图像数量
中国民族服饰	20	白族、傣族、汉族、藏族、苗族等	100
亚洲民族服饰	10	阿拉伯、印度、日本、泰国、土耳其等	90
欧洲民族服饰	10	英格兰、希腊、意大利、挪威、波兰等	90

中国，AsianEthnic）和欧洲民族服饰子集（European-Ethnic）。具体数据集类别构成和数量如表2所示。该数据集旨在通过服饰风格来呈现文化多样性的视觉特征，为服饰分类、文化识别、跨文化视觉理解等领域的研究提供支持。

1.1　数据集收集与清洗

在构建EthnicFashion数据集时，我们采用了多源收集策略。大部分图片通过网络爬取获得，数据来源包括百度图片、视觉中国、Flipkart、亚马逊等常用图像搜索引擎和电子商务平台。这些平台提供了多样化的视觉内容，有助于确保数据覆盖不同地区和民族。为进一步丰富数据集，并减少搜索引擎结果可能存在的偏差或冗余，我们还从《时尚芭莎》等时尚杂志，以及小红书等社交平台人工收集了额外图片。这些补充来源提供了更多真实场景或编辑场景下民族服饰的多样化、时效性表征。初步收集完成后，所有图片都经过了严格的清洗过程：我们对每张图片进行人工检查，验证民族类别标签的正确性，并评估图片整体质量。低分辨率、模糊、标签错误的图片、重复的图片以及可能存在版权问题的图片均被剔除，以确保数据集的高可用性。最终获得了包含3800张图片的高质量数据集，为后续任务奠定了可靠基础。

1.2　数据集标注

EthnicFashion数据集采用互斥类别标签进行标注，每张图片仅归属于一个民族。类别集合规模适中，既适合训练标准分类模型，又能保持风格的丰富多样性。具体而言，数据集根据地理和文化区域分为三个子集：中国民族服饰子集包含20个类别标签，包括白族、傣族、满族、苗族、土家族等；亚洲民族服饰子集（不含中国）包含10个类别标签，如印度、日本、泰国等民族；欧洲民族服饰子集包含10个类别标签，如英格兰、希腊、意大利等民族。

1.3　数据集自合成

如前所述，民族服饰分类的一大主要挑战是高质量标注数据的获取受限。为解决这一问题，我们采用了基于生成模型的数据增强策略，能够合成保留传统服饰关键视觉特征的新训练样本。

图2　数据集中真实图片和对应生成图片示例

Fig. 2　Examples images of real samples in EthnicFashion and corresponding generated samples

近年来，面向特定目标的生成模型（尤其是文本-图像生成模型）在保持细粒度目标保真度方面展现出了强大能力。在本研究中，我们采用DreamBooth（Ruiz 等，2023）方法，该方法旨在微调Stable Diffusion（Rombach 等，2022）等大型文本-图像扩散模型，仅需少量参考图像即可引导生成过程（详见图3中数据集自合成阶段）。首先，我们通过零样本生成获得与真实图像数量相同的初步参考图像集；然后，仅使用训练集中的真实图像对扩散模型进行微调，其他训练参数保持默认设置。微调过程中，提示词的结构为“a photo of <label > clothing”。这种定制化微调能够生成更细致、上下文更准确的民族服饰表征。如图2所示，尽管生成图像在面部细节等局部方面仍不如真实图像精细，但它们能够清晰展现不同民族服饰风格的独特特征。考虑到样本总量较少且数据分布与自然图像差异较大，传统的自动评估指标，如Fréchet Inception Distance（FID）、Inception Score（IS），难以准确反映图像质量。因此，我们

分别开展了两个相互独立的主观评估实验，从评分分布与可区分性两个角度对生成图像质量进行分析。首先，我们邀请多名用户对混合图像集进行10分制独立评分（1分最低，10分最高）。评分过程中不告知图像来源，以避免先验偏差。统计结果表明，生成图像的平均得分为5.04分，真实图像的平均得分为6.33分。两者之间存在一定差距，但整体分布明显重叠，说明生成样本在整体视觉质量上已接近真实样本水平。随后我们又邀请用户对图像是否为生成图像进行判断，将主观判定结果与真实标签进行对比，最终得到混淆矩阵如表3所示，从结果可以看出，生成图像被误判为真实图像的比例达到47.6%，整体分类准确率约为59.2%，接近随机猜测水平，这说明用户难以稳定区分真实和生成样本，两者视觉表现上区分度较低。综合来看，虽然生成图像在局部细节上仍然存在一定瑕疵，但整体质量已足以用于数据扩展。在此基础上，生成数据不仅缓解了少样本带来的稀缺问题，也有效保留了民族服饰的标志性特征。

表3　真实与生成图像主观判定的混淆矩阵

Table 3　Confusion Matrix for Subjective Classification of Real and Generated Images

	预测为真实	预测为生成
真实	329	171
生成	238	263

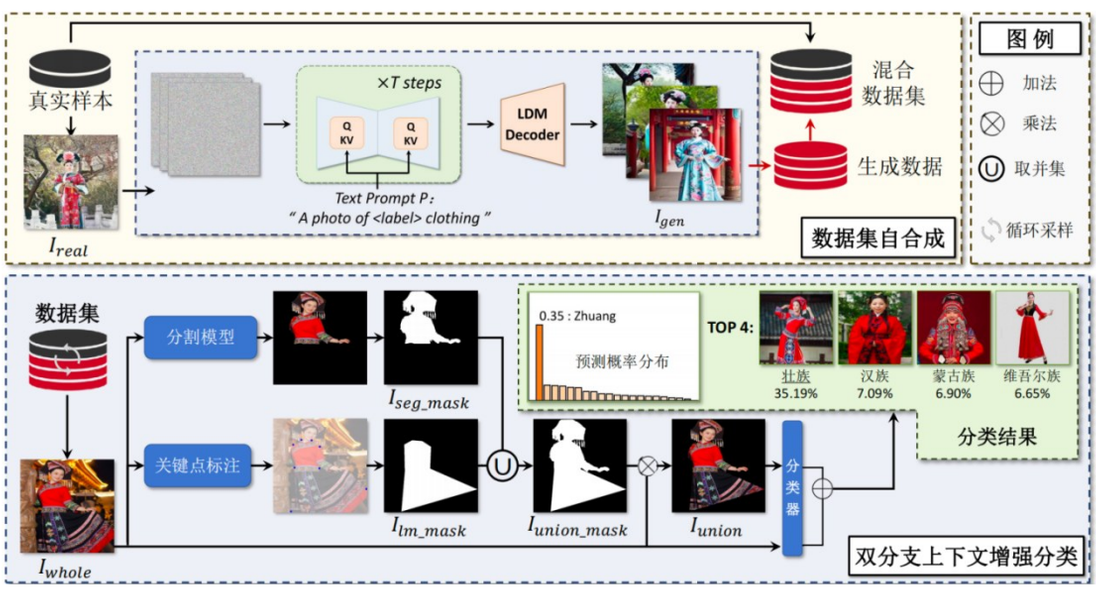

图3　方法整体框架图

Fig. 3　The overall framework of the proposed method

2　双分支上下文增强分类

2.1　方法概述

我们将民族服饰分类任务视为标准的多类别图像分类任务。图3展示了我们方法的整体框架，该框架包含两个主要阶段：数据集自合成阶段和双分支上下文增强分类阶段。在数据集自合成阶段，我们利用生成模型，基于真实图像$I_{\text{real}} \in \mathbb{R}^{3\times H\times W}$合成一系列图像$I_{\text{gen}} \in \mathbb{R}^{3\times H\times W}$。生成的图像仅用于训练过程，尤其适用于真实民族服饰数据收集成本高或范围受限的场景。在上下文增强分类阶段，我们旨在通过引导模型关注语义信息丰富的区域来提升分类准确率。首先，我们对真实图像I_{real}和生成图像I_{gen}进行处理，利用现有工具提取关键点和分割掩码$I_{\text{seg_mask}} \in \{0,1\}^{H\times W}$；在此基础上，生成基于关键点的掩码$I_{\text{lm_mask}} \in \{0,1\}^{H\times W}$，并计算统一区域掩码，同时聚焦分割图像和关键点裁剪图像的上下文信息。最终得到统一表征$I_{\text{union}} \in \mathbb{R}^{3\times H\times W}$，并将其与原始图像$I_{\text{whole}} \in \mathbb{R}^{3\times H\times W}$作为并行分支输入分类模型，进行特征提取。最后，将两个分支的预测分数相加，得到最终结果。

2.2　上下文增强分类

传统上，服饰领域的分类任务将整幅图像作为

输入，依赖能在尺寸、姿态、背景等变化中保持泛化能力的强大特征提取器，再通过分类器给出最终预测。然而，民族服饰结构复杂，不同民族的服饰具有显著不同的特征区域。为解决这一问题，我们设计了一种上下文信息增强方法，使模型能够关注对准确区分民族服饰至关重要的细粒度和场景特定信息。此外，我们还引入了循环采样策略，以解决训练过程中真实数据与生成数据可能存在的不平衡问题，进一步提升分类准确率。

2.2.1　采样策略

尽管生成图像有助于弥补数据不足，但它们的质量通常低于真实图像。在训练过程中直接将生成图像与真实图像混合，可能导致模型过拟合于生成样本（尤其是当生成样本数量远超过真实样本时），进而导致泛化能力下降和分类性能降低。为解决这一问题，我们提出了一种循环采样策略（如图4所示），能在整个训练过程中有效平衡数据质量和多样性。具体而言，我们不对所有样本进行随机打乱，而是按顺序读取生成图像，且每张生成图像仅使用一次；同时，从真实图像集中循环读取真实图像，直到所有生成图像都被处理完毕。这样，在每个数据加载周期中，真实样本与生成样本的比例始终保持1∶1，通过复制真实图像实现了与生成图像规模的匹配。这种策略使模型既能持续接触并学习高质量真实样本的特征，又能接触到生成图像的多样性特征。该策略不仅避免了模型对质量较低的生成数据的过拟合，还确保了训练过程中真实图像的有意义特征得以保留，从而有助于模型实现更稳定的收敛，提升整体性能。

2.2.2　样本标注

关键点标注：关键点指服饰重要区域上的点位，如领口、下摆、袖口的边角等。与人体姿态关节不同，这些关键点以服饰为中心，用于精确定位服饰结构。每张图像标注8个此类关键点。为在保证标注质量的同时兼顾效率，我们对真实图像进行人工关键点标注；对于生成图像，则利用公开的预训练模型①（Liu 等，2016）自动提取服饰关键点，实现了对合成数据的可扩展结构化标注。

分割标注：除服饰本身外，头饰、首饰等具有文化意义的配饰也对图像语义有重要贡献。为捕捉这些元素，我们采用轻量级的前景-背景分割工具②从图像中分割出主要视觉目标。尽管存在SAM等更先进的分割模型，但我们发现它们在本场景中并不适用。实验中，SAM倾向于生成过于细致的细粒度掩码，这些掩码可能包含无关区域或边界碎片化的区域，往往需要额外人工修正才能分离出所需的服饰和配饰区域，这不仅增加了标注成本，还为后续处理引入了不必要的复杂性（而后续处理并不需要极高的精度）。相比之下，更简单的分割工具能够更快地实现更清晰的前景分离，足以满足本任务需求，且效率更高。

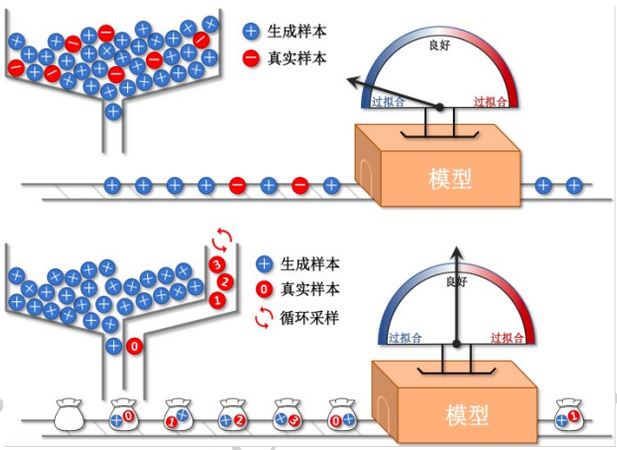

图4　循环采样方法示意图

Fig. 4　Illustration of our cyclic-sampling method

2.2.3　区域融合

如图3所示，基于关键点的区域可以精确定位服饰核心结构位置，如领口、袖口等，但往往只能覆盖小范围区域；相比之下，分割掩码具有更大的覆盖范围，能够囊括整体轮廓及其周边的配饰元素，但可能存在覆盖不完整或定位偏差的问题。我们通过对这两个区域进行融合操作，即取二者的并集。关键点确保结构的准确性，分割则补充周边或配饰细节，可以实现互补的效果。这种混合方法在精确性和完整性之间取得了平衡，弥补了单独使用任一方法的不足。以往利用关键点的研究（Goenka 等，2022；Liu 等，2016；Wang 等，2018）通常聚焦于单个关键点周围的小局部区域，这种策略对现代服饰有效，因为现代服饰的结构相对简单，图案具有重复性。然

①https://github.com/liuziwei7/fashion-landmarks

②https://github.com/danielgatis/rembg

而,对于民族服饰或传统服饰,这种方法并不适用。民族服饰通常具有复杂的纹理、不对称装饰和文化特定设计元素,这些细节往往分布在较大的不规则区域,而不是局限于单个关键点周围的区域。为了更好地捕捉这类服饰的结构和装饰丰富性,我们引入了基于轮廓的裁剪策略:通过将标注的关键点按顺序连接形成多边形轮廓,提取轮廓所包围的区域,从而确定感兴趣的区域。该区域用于生成二值化掩码,即基于关键点标注的掩码。其数学定义为:

$$I_{\text{lm\_mask}}(x,y)\begin{cases}1, \text{if}(x,y)\in \text{ContourRegion}(\boldsymbol{LM})\\ 0, otherwise\end{cases} \tag{1}$$

式中,ContourRegion($\boldsymbol{LM}$)点集包围的区域。

接下来,我们将基于关键点标注包围区域的掩码I_{lm_mask}与数据集中已通过分割模型标注得到的分割掩码$I_{\text{seg_mask}}$相结合。通过对关键点标注掩码和分割模型掩码进行逻辑或运算,即取两个掩码标记区域的并集,可以得到统一表征掩码

$$I_{\text{union\_mask}}(x,y)=I_{\text{lm\_mask}}(x,y)\vee I_{\text{seg\_mask}}(x,y) \tag{2}$$

将这一统一掩码$I_{\text{union_mask}}$应用于原始的完整图像I_{whole},可得到两标注区域合并后的前景图像$I_{\text{union}}\in\mathbb{R}^{3\times H\times W}$,该图像突出了模型更感兴趣的前景区域

$$I_{\text{union}}(x,y)=I_{\text{whole}}(x,y)\cdot I_{\text{union\_mask}}(x,y) \tag{3}$$

这一过程减少了背景噪声和无关细节的干扰,引导模型关注语义丰富的区域。为了在训练过程中充分利用整体视觉信息和聚焦的前景视觉信息,我们采用了双分支策略:将原始图像I_{whole}和基于掩码得到的前景图像I_{union}输入相同的特征提取器FE和分

类器cls,并在输出层对两者预测的结果分数进行相加融合,得到最终输出:

$$output_{\text{final}}=\text{cls}\left(\text{FE}\left(I_{\text{union}}\right)\right)+\text{cls}\left(\text{FE}\left(I_{\text{whole}}\right)\right) \tag{4}$$

该双分支融合机制包含两个层面:首先,在特征级,我们对按1:1比例采样的真实样本与生成样本分别提取图像特征,并将其在特征空间进行连接(concat)操作后输入分类器,以融合真实数据的高质量结构信息与生成数据带来的分布多样性,增强类别表达能力;其次,在分数层,原始图像与基于联合掩码的前景图像的两条分支经过上述"特征提取+分类器"流程得到各自的分数输出,再通过加和实现决策级融合,从而在保持模型轻量的同时,有效整合全局语义信息与关键区域增强信息,提高细粒度民族服饰分类的判别性能。

模型采用SoftTargetCrossEntropy损失函数进行优化,分别计算每个分支的损失并求和,得到最终的训练损失:

$$\begin{aligned}loss_{\text{final}}=&\text{STCE}\left(output_{\text{whole}},label_{\text{whole}}\right)\\&+\text{STCE}\left(output_{\text{union}},label_{\text{union}}\right)\end{aligned} \tag{5}$$

式中,$output_{(\cdot)}$表示完整/前景图像经过双分支的输出结果,$label_{(\cdot)}$表示真实标签值,STCE表示SoftTargetCrossEntropy,即软目标交叉熵函数,常用于处理概率分布不均匀的多分类问题。它在计算时不会将所有概率质量集中到单一类别,而是允许非目标类别保留一定概率,从而降低梯度的陡峭程度并起到正则化作用。当训练样本有限、类间外观相似且标注噪声不可避免的时候,硬标签容易促使模型过度自信并快速记忆训练样本,从而导致过拟合;而STCE通过"软化"监督信号,使优化过程更平滑,有助于提升泛化能力与训练稳定性。此外,由于本文同时引入生成样本参与训练,软目标也能够在一定程度上缓解生成数据与真实数据之间分布差异带来的训练偏置,使模型不会对某一类样本产生过度拟合。

重要的是,为确保推理效率,我们采用了简化的测试流程:推理阶段只需要使用原始图像,而无需对测试图像进行关键点的标注和送入分割模型进行分割。这种训练输入与测试输入的解耦设计,使我们能够在不增加部署阶段计算成本的前提下,保留训练带来的优势。因此,训练过程中结合关键点标注和分割掩码,不仅提供了更丰富的监督信息,提升了模型对背景干扰的鲁棒性,最终还实现了性能的提升。

3 实 验

3.1 实验设置与评价指标

由于目前缺乏专门面向民族服饰分类任务的公开基准模型或统一评测标准,为保证实验的代表性和可比性,我们选取了多种具有代表性的通用图像分类模型作为基线并进行插件化对比。这些模型涵盖不同的技术路线,包括基于卷积神经网络(convolutional neural networks, CNN)、基于Transformer架

表4　不同机制下的分类准确率(%)

Table 4　Classification accuracy(%) of different schemes

方法	2-shot			4-shot			6-shot			8-shot			10-shot		
	Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5
ResNet101	10.50	23.81	35.94	15.38	33.63	46.33	20.75	39.63	53.75	21.75	43.81	57.63	24.38	46.00	59.13
ResNet101+Ours	21.00	41.63	55.88	22.25	43.56	58.88	23.62	44.75	59.25	29.12	53.37	66.75	27.69	51.75	64.31
DenseNet201	8.50	21.00	33,50	13.19	29.31	43.50	17.44	36.06	50.62	20.88	42.94	56.13	22.69	43.88	57.44
DenseNet201+Ours	23.88	42.50	56.44	22.06	41.44	55.56	24.75	46.00	60.00	28.94	52.25	65.31	28.62	49.13	63.75
Tip-Adapter	23.88	43.31	57.44	22.38	42.81	56.75	28.63	50.50	65.06	29.63	53.25	66.13	34.44	59.00	72.25
Tip-Adapter+Ours	27.38	51.75	64.94	27.48	48.50	61.88	29.38	50.37	64.06	31.75	57.13	70.68	33.44	57.00	70.88
AMU-Tuning	9.62	23.12	35.18	12.19	26.93	39.18	14.56	30.50	42.31	16.00	31.68	44.75	16.75	33.69	45.50
AMU-Tuning+Ours	12.56	26.81	39.06	9.75	24.19	35.88	12.19	25.69	36.05	12.69	29.25	42.56	13.44	29.06	43.06
Swin-Transformer	7.75	21.06	32.50	13.19	30.50	43.25	19.25	38.94	53.38	26.81	49.94	65,13	31.12	56.00	67.75
Swin-Transformer+Ours	**35.75**	**57.88**	**70.06**	**37.88**	**60.13**	**71.63**	**41.06**	**65.94**	**76.38**	**47.94**	**71.50**	**80.31**	**47.06**	**69.13**	**79.13**

注：加粗字体为每列最优值。+Ours表明在原基准模型上添加了本文提出的两个模块。

构和面向少样本学习场景的模型，具体包括ResNet101（He 等，2016）、DenseNet201（Huang 等，2017）、Swin-Transformer V2（Liu 等，2022）、Tip-Adapter（Zhang 等，2022）和AMU-Tuning（Tang 等，2024）。其中，前三种是广泛使用的传统分类模型，后两种是专门为少样本图像分类场景设计的模型。

所有模型均在NVIDIA H800 GPU上进行训练。为确保公平比较，batch size设置为与少样本场景中的样本数相等。所有实验均采用AdamW优化器（Loshchilov & Hutter, 2019）：对于ResNet101和DenseNet201，学习率固定为1×10^{-4}；对于其他模型，则采用其原始论文中报告的默认超参数设置。

除非另有说明，实验默认采用Swin Transformer V2（Liu 等，2022）作为骨干网络。实验主要在我们提出的EthnicFashion数据集上进行评估，默认聚焦于中国民族服饰子集。评估时，预留80张真实图像作为测试集，其余真实图像和生成图像共同用于训练。为进一步验证我们框架的泛化能力，我们还在基于DeepFashion数据集构建的miniDeepFashion数据集上进行了补充测试。miniDeepFashion包含10个全身服饰类别，对于每个类别，若样本数少于90张，则保留所有图像；否则，每个类别随机采样90张图像。实验采用少样本设置：每个类别仅使用10张图像进行训练，其余图像用于测试，从而模拟数据稀缺场景下的模型表现。为全面评估分类性能，我们采用Top-1、Top-3和Top-5准确率作为评价指标。这些指标反映了模型将正确标签纳入前N个预测结果的能力，能够从不同预测严格程度层面，细致全面地评估模型性能。

3.2　性能对比

本节将我们提出的方法与多个基线模型的图像分类性能进行了对比，实验结果汇总于表4。我们的方法表现出显著的性能提升，与性能最佳的基线模型Tip-Adapter（Zhang 等，2022）相比，Top-1准确率至少提高了10%，在Top-3和Top-5指标上同样保持领先，这表明本文所提方法在少样本民族服饰分类场景下具有明显优势。此外，我们将数据生成和上下文增强模块整合到其他模型中，结果表明这些模块在传统分类模型上带来了稳定收益，效果良好，但对少样本模型的适配性并非完全理想。这主要是因为生成数据能为传统分类模型提供更多训练样本，从而提升模型性能；而少样本模型对数据质量和分布高度敏感，若生成数据未能准确表征原始数据分布，或包含噪声及无关变异，则模型可能过拟合于这些噪声，导致在真实测试数据上的性能下降。此外，在表1中，我们还补充了更多基于视觉语言大模型的对比实验，结果表明，在相同训练规模下，本文方案的性能同样优于该类基准模型。

表5　不同数据集的性能表现

Table 5　The performance of our proposed method on different datasets.

方法	ChineseEthnic			AsianEthnic			EuropeanEthnic			miniDeepFashion		
	Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5
Base	31.12	56.00	67.75	48.56	75.09	87.99	31.50	56.88	73.75	54.84	77.50	90.31
Ours	47.06	69.13	79.13	62.83	86.98	93.12	41.38	66.75	79.00	62.00	85.25	94.04

3.3　泛化能力

为验证模型的泛化能力，我们在四个不同数据集上测试了我们的方案，表5显示了各数据集上的图像分类性能。结果表明，我们提出的方法在各类民族服饰数据集上均显著提升了图像分类性能。此外，该方法在通用服饰分类数据集上也具有有效性。尽管由于现代服装分类任务较为简单，其性能提升幅度相较于在民族服饰数据集上略小，但此结果仍然证明了我们方法思想的有效性。

3.4　消融实验

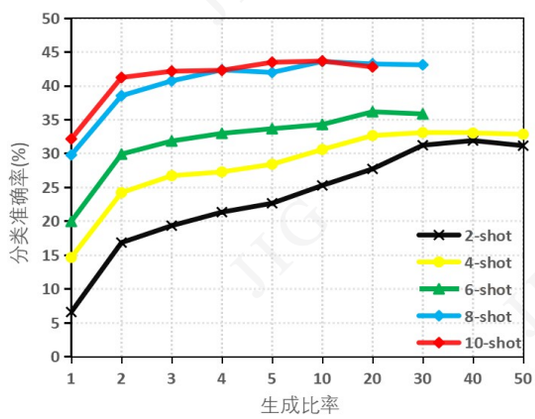

图5　不同生成比例下的分类准确率

Fig. 5　Classification accuracy(%) under different ratios of generated/real images.

生成图像与真实图像的比例：在数据集自合成阶段，我们利用生成模型合成额外样本以扩充数据集。由于目前尚无明确的定量标准来确定最佳数据集规模，我们通过实验测试了不同生成数据比例，以确定不同样本数设置下的最佳生成图像数量，并将其用于后续实验。结果如图5所示，随着生成数据比例的增加，分类性能先提升后下降。性能下降的原因可能是生成图像开始出现过度相似性，且生成数据引入的噪声超过了额外样本多样性带来的益处。值得注意的是，考虑到数据生成和标注的计算成本与时间成本，一旦性能开始下降，我们便停止进一步的数据生成。

数据生成与上下文信息增强的有效性：图6展示了数据生成和上下文信息增强的有效性。可以看出，这两个模块均能持续提升分类准确率。具体而言，数据生成最多可使基线模型的性能提升12%-24%，且样本数越少，性能提升越显著，这一结果证实了数据量对分类任务的重要性。此外，尽管上下文信息增强带来的性能提升略低于数据生成，但它在所有场景中均表现出有效性。

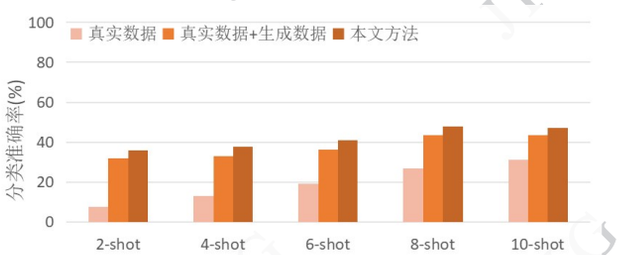

图6　数据生成和上下文信息增强对实验结果的影响

Fig. 6　Effects of data generation and contextual information enhancement

采样策略的有效性：我们还针对采样策略进行了额外的消融实验。具体而言，我们用真实图像循环采样策略替代了“普通采样”也即混合后随机采样的策略，确保输入模型的真实图像与生成图像比例始终保持1:1。在训练阶段，生成图像能够有效扩充数据分布，提高类别覆盖度与多样性，但在细节真实性和结构一致性方面仍不及真实图像。当真实样本数量远少于生成样本时，如果采用普通采样，即图像从真实图像和生成图像的混合池中随机选取，模型在迭代过程中将更频繁地接触生成图像，容易对生成分布产生过度拟合，从而削弱在真实测试数据上的泛化表现。为缓解这一问题，我们设计了循环采样策略，通过在训练过程中对真实样本与生成样本的输入比例进行显式约束，确保两者保持约1:1

的平衡，既保证了生成数据带来的分布扩展与类别补充优势，又持续引入高质量真实样本进行特征校正，避免训练偏向低质量分布。表6的实验结果表明，相较于普通随机采样，循环采样策略能够在少样本设置下稳定提升模型性能，说明合理的采样机制对于融合真实与生成数据至关重要。

表6　循环采样的消融实验

Table 6　Ablation study for sampling strategy

方法	2-shot	4-shot	6-shot	8-shot	10-shot
普通采样	32.38	30.00	31.44	41.56	42.94
循环采样	35.75	37.88	41.06	47.94	47.06

4　结　论

本文提出了一种新的民族服饰分类方法，该方法包括构建新数据集EthnicFashion（数据集地址：10.57760/sciencedb.j00240.00167），以及整合数据生成和上下文信息增强技术。其中，数据生成策略通过合成额外样本解决了数据稀缺问题，并采用了有效的循环采样策略；上下文信息增强通过利用关键点检测和分割技术优化了特征提取过程，从而提升了分类准确率。我们在EthnicFashion数据集上对该方法进行了评估，结果表明，该方法将分类准确率提升15%以上，并在不同少样本设置下均取得稳定性能增益，相较于现有方法达到了当前最优水平。

参考文献（References）

Anaby-Tavor A, Carmeli B, Goldbraich E, Kantor A, Kour G, Shlomov S, et al. Do not have enough data? Deep learning to the rescue![C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34: 7383-7390.[DOI:10.1609/aaai.v34i05.6233]

Azizi S, Kornblith S, Saharia C, Norouzi M and Fleet D J. Synthetic data from diffusion models improves ImageNet classification[EB/OL].(2023)[2024-06-01]. https://arxiv.org/abs./2304.08466.[DOI: 10.48550/arxiv.2304.08466]

Gao P, Geng S, Zhang R, Ma T, Fang R, Zhang Y, et al. Clip-adapter: better vision-language models with feature adapters[J]. International Journal of Computer Vision, 2024, 132(2): 581-595.[DOI:10.1007/s11263-023-01891-x]

Goenka S, Zheng Z, Jaiswal A, Chada R, Wu Y, Hedau V, et al. FashionVLP: vision language transformer for fashion retrieval with feedback[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 14105-14115.[DOI: 10.1109/CVPR52688.2022.01371]

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets[J]. Advances in Neural Information Processing Systems, 2014, 27: 1-9.[DOI: 10.1145/3422622]

He K, Zhang X, Ren S and Sun J. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.[DOI: 10.1109/IEEESTD.2001.92771]

He R, Sun S, Yu X, Xue C, Zhang W, Torr P, et al. Is synthetic data from generative models ready for image recognition?[C]//Proceedings of the International Conference on Learning Representations. 2023.[DOI:10.48550/arXiv.2210.07574]

Ho J, Jain A and Abbeel P. Denoising diffusion probabilistic models[J]. Advances in Neural Information Processing Systems, 2020, 33: 6840-6851.[DOI:10.48550/arXiv.2006.11239]

Hospedales T, Antoniou A, Micaelli P and Storkey A. Meta-learning in neural networks: a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(9): 5149-5169.[DOI:10.1109/TPAMI.2021.3079209]

Huang G, Liu Z, Van Der Maaten L and Weinberger K Q. Densely connected convolutional networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4700-4708.[DOI:10.1109/CVPR.2017.243]

Li Y, Guo J, Qi L, Li W and Shi Y. Text and image are mutually beneficial: enhancing training-free few-shot classification with CLIP[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2025, 39: 5039-5047.[DOI:10.48550/arXiv.2412.11375]

Liu Z, Hu H, Lin Y, Yao Z, Xie Z, Wei Y, et al. Swin Transformer V2: scaling up capacity and resolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 12009-12019.[DOI:10.48550/arXiv.2111.09883]

Liu Z, Luo P, Qiu S, Wang X and Tang X. DeepFashion: powering robust clothes recognition and retrieval with rich annotations[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1096-1104.[DOI: 10.1109/CVPR.2016.124]

Liu Z, Yan S, Luo P, Wang X and Tang X. Fashion landmark detection in the wild[C]//European Conference on Computer Vision. 2016: 229-245.[DOI:10.48550/arXiv.1608.03049]

Loshchilov I and Hutter F. Decoupled weight decay regularization[C]//Proceedings of the International Conference on Learning Representations. 2019.[DOI:10.48550/arXiv.1711.05101]

Moreno-Barea F J, Jerez J M and Franco L. Improving classification accuracy using data augmentation on small data sets[J]. Expert Systems with Applications, 2020, 161: 113696.[DOI: 10.1016/j.

eswa.2020.113696]

Nawaz M M T, Hasan R, Hasan M A, Hassan M and Rahman R M. Automatic categorization of traditional clothing using convolutional neural network [C]//Proceedings of the IEEE/ACIS International Conference on Computer and Information Science. 2018: 98-103. [DOI:10.1109 /ICIS.2018.8466523]

Nichol A Q and Dhariwal P. Improved denoising diffusion probabilistic models [C]//Proceedings of the International Conference on Machine Learning. 2021: 8162-8171.[DOI: 10.48550/arXiv.2102.09672]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision [C]//Proceedings of the International Conference on Machine Learning. 2021: 8748-8763.[DOI:10.48550/arXiv.2103.00020]

Raghu A, Raghu M, Bengio S and Vinyals O. Rapid learning or feature reuse? Towards understanding the effectiveness of MAML [EB/OL]. (2019)[2024-06-01]. https://arxiv. org/abs/1909.09157. [DOI: 10.48550/arXiv. 1909.09157]

Ramesh A, Dhariwal P, Nichol A, Chu C and Chen M. Hierarchical text-conditional image generation with CLIP latents [EB/OL]. (2022)[2024-06-01]. https://arxiv. org/abs/2204.06125. [DOI: 10.48550/arXiv. 2204.06125]

Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B. High-resolution image synthesis with latent diffusion models [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 10684-10695.[DOI:10.1109/CVPR52688.2022.01042]

Ruiz N, Li Y, Jampani V, Pritch Y, Rubinstein M and Aberman K. DreamBooth: fine-tuning text-to-image diffusion models for subject-driven generation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 22500-22510. [DOI:10.48550/arXiv.2208.12242]

Saharia C, Chan W, Saxena S, Li L, Whang J, Denton E L, et al. Photorealistic text-to-image diffusion models with deep language understanding[J]. Advances in Neural Information Processing Systems, 2022, 35: 36479-36494.[DOI:10.48550/arXiv.2205.11487]

Shin S Y, Jo G and Wang G. 2023. A novel method for fashion clothing image classification based on deep learning[J]. Journal of Information and Communication Technology, 22(1): 127 - 148.[DOI:10.32890/jict2023.22.1.6]

Snell J, Swersky K and Zemel R. Prototypical networks for few-shot learning[J]. Advances in Neural Information Processing Systems, 2017, 30: 1-10.[DOI:10.48550/arXiv.1703.05175]

Tang Y, Lin Z, Wang Q, Zhu P and Hu Q. AMU-Tuning: effective logit bias for CLIP-based few-shot learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 23323-23333.[DOI:10.1109/ CVPR52733.2024.02201]

Thewsuwan S and Horio K. 2018. Texture-based features for clothing classification via graph-based representation[J]. Journal of Signal Processing, 22(6): 299 - 305.[DOI:10.2299/jsp.22.299]

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?[C]//Computer Vision - ECCV 2020: 16th European Conference, Glasgow, UK, August 23 - 28, 2020, Proceedings, Part XIV. Springer, 266 - 282. [DOI: 10.1007/978-3-030-58568-6_16]

Wan Y, Yan C, Zhang B and Zou G. 2022. Learning image representation via attribute-aware attention networks for fashion classification [C]//International Conference on Multimedia Modeling. Springer, 69 - 81.[DOI:10.1007/978-3-030-98358-1_6]

Wang H, Jie S and Deng Z. 2024. Focus your attention when few-shot classification[J]. Advances in Neural Information Processing Systems, 36.[DOI:10.5555/3666122.3668730]

Wang W, Xu Y, Shen J and Zhu S C. 2018. Attentive fashion grammar network for fashion landmark detection and clothing category classification [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 4271 - 4280. [DOI: 10.1109/CVPR.2018.00449]

Wu G, Chen J, Li Q, Zhang W, Zheng W and Wang R. Region Attention Fine-tuning with CLIP for Few-shot Classification. 2024. IEEE International Conference on Multimedia and Expo (ICME).[DOI: 10.1109/ICME57554.2024.10688204]

Wu S, Liu L, Fu X, Liu L and Huang Q. Human detection and multi-task learning for minority clothing recognition[J]. Journal of Image and Graphics, 2019, 24(4): 562-572 (吴圣美, 刘骊, 付晓东, 刘利军, 黄青松. 结合人体检测和多任务学习的少数民族服装识别[J]. 中国图象图形学报, 2019,24(4):562-572) [DOI:10.11834/jig.180500.]

Xiao H, Rasul K and Vollgraf R. 2017. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms [EB/OL]. [2017]. https://arxiv. org/abs/1708.07747. [DOI: 10.48550/arXiv. 1708.07747]

Xu J and Le H. 2022. Generating representative samples for few-shot classification [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 9003 - 9013.[DOI: 10.1109/CVPR52688.2022.00880]

Zhang Q, Liu L, Gan L, Fu X, Liu L and Huang Q. Clothing parsing of Chinese minorities via the fusion of visual style and label constraints [J]. Journal of Image and Graphics, 2021, 26(2): 402-414 (张茜, 刘骊, 甘霖, 付晓东, 刘利军, 黄青松. 融合视觉风格和标签约束的少数民族服装图像解析[J]. 中国图象图形学报, 2021,26(2):402-414)[DOI: 10.11834/jig.190655.]

Zhang R, Zhang W, Fang R, Gao P, Li K, Dai J, et al. Tip-adapter: training-free adaption of CLIP for few-shot classification[C]//Proceedings of the European Conference on Computer Vision. 2022:

493-510.[DOI:10.1007/ 978-3-031-19833-5_29]

Zhang Y, Zhu W, Tang H, Ma Z, Zhou K, Zhang L. 2024. Dual memory networks: A versatile adaptation approach for vision-language models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 28718 - 28728.[DOI: 10.1109/CVPR52733.2024.02713]

Zou X, Kong X, Wong W, Wang C, Liu Y and Cao Y. FashionAI: a hierarchical dataset for fashion understanding[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019.[DOI:10.1109/ CVPRW.2019.00039]

作者简介

曹丝露，女，复旦大学硕士研究生，主要研究方向为数字水印与文旅视觉。 E-mail:slcao23@m.fudan.edu.cn

钱振兴，通信作者，男，教授，入选国家高层次人才计划，主要研究方向为信息隐藏，多媒体信息安全等。 E-mail:zxqian@fudan.edu.cn

刘高志，男，复旦大学博士研究生，主要研究方向为数字水印。 E-mail:gzliu24@m.fudan.edu.cn

奚美娟，女，复旦大学硕士研究生，主要研究方向为文旅视觉。 E-mail:23210240336@m.fudan.edu.cn

张新鹏，男，教授，国家杰青，主要研究方向为多媒体信息安全、人工智能安全。 E-mail:zhangxinpeng@fudan.edu.cn