摘要

图像检索是计算机视觉基础任务之一，虽然已经经过近三十年的研究，取得长足进步，但依然难以实现精细化的检索。精细化视觉检索旨在从数据库中找到与查询图像精细语义相似的目标图像，涵盖了实例检索任务（如行人重识别）与细粒度检索任务（如鸟类识别）。此任务在军事侦查、工业生产、城市安防、医疗诊断等领域都有广泛应用，如从遥感图像中识别特定舰船或从视频监控中锁定特定嫌疑人。因此，精细化视觉检索是当前工业界、学术界一个研究热点。

现有的深度学习方法在精细化视觉检索任务上取得了一定的成果。然而，深度学习依赖于样本-标签对进行训练，并且往往仅能利用模型输出的高维全局特征，导致现有方法在精细化检索上依然存在以下三个未充分解决的难点。首先，常规深度学习方法仅将模型的最终特征用于图像检索。由于特征分辨率不足，这些方法对图像微小差异的辨别力不足。其次，此类图像标注复杂且易出错，标签噪声问题突出，加之数据采集难度大，常形成长尾分布。传统有监督训练方式对这些问题的鲁棒性不足。最后，由于精细图像的种类繁多，数据库规模较大，直接采用深度网络生成的高维实值特征进行检索效率较低。因此，本文创新性地引入深度模型解耦学习策略，针对精细化视觉检索的三大难点——特征判别力弱、训练鲁棒性差及检索效率低，提出一套全面解决方案，主要贡献概述如下：

针对特征判别力弱的问题，本文提出多尺度特征解耦学习方法，通过解耦和学习多尺度特征，提升深度特征的细节判别力。首先，本文提出双路网络架构分别处理高低分辨率图像，提升特征细节表达能力与跨分辨率匹配精度。其次，本文引入文本指导的全局-局部特征学习机制，即使在缺乏局部标注的情况下，也能借助局部信息文本提示引导模型聚焦于高区分性的局部区域，进一步提高特征对关键局部细节的理解能力。在多个精细化检索数据集上本文方法均能提升特征判别力。例如，所提出的方法在网购商品数据集 SOP 上，rank-1 准确率在基线方法基础上提升了 15.6%。

针对训练鲁棒性差的问题，本文提出样本-标签解耦优化方法，通过解耦训练集中的样本与标签的绑定关系，获得更准确的训练目标。首先，本文设计基于邻域注意力的标签纠正方法，融合样本邻域标签信息，打破样本-标签的一对一映射，预测更准确训练目标，减轻噪声标签引起的过拟合。其次，本文提出基于特征记忆的子标签学习策略，学习多个子标签，用特征记忆隔离噪声样本、平衡长尾分布，并强化图像特征。在含噪声和长尾分布的数据上，所提出的方法展现出了对噪声标签、长尾分布的鲁棒性。例如，在含有 50% 标签噪声的 Market1501 行人重识别数据集中，基线方法性能由

无噪声情况下的 94.4% rank-1 准确率下降至 69.2%，而本方法在相同噪声比例下仍能维持 92.2% 的高准确率。

针对检索效率低的问题，本文提出多语义属性解耦编码方法，通过从解耦后的多语义特征中选择出高判别力特征，编码出紧致二值特征用于高效检索。本文利用条件扩散模型不同层对应不同语义的特性，将深度网络输出解耦为多语义特征，并通过信息熵筛选出信息丰富的部分，进而编码为既紧凑又具高辨识力的二值特征。在多种图像检索任务上本文方法都展现出优秀性能，如在鸟类细粒度识别数据集 CUB 上，本文将 32 位二值特征的性能由基线方法的 17.1% 提升至 30.1%。此外，本文整合上述研究成果，设计了一种同时优化特征判别力、鲁棒性与紧致性的学习方法，确保二值特征在有偏数据集训练下性能接近实值特征，且相似度计算速度提升为原本的 414 倍。

综上，本文构建了一套基于深度模型解耦学习的精细化视觉检索方法，通过解耦深度学习中的特征学习、模型训练以及特征编码的过程，有效提升了精细化视觉检索的特征判别力、对有偏数据的训练鲁棒性及检索效率，对推动图像检索领域的进步具有积极意义。

关键词：精细化视觉检索，解耦学习方法，特征判别力，训练鲁棒性，检索效率

Deep Disentanglement Learning for Fine-Grained Visual Retrieval

Mao Shunan (Computer Applied Technology)
Directed by: Prof. Zhang Shiliang

ABSTRACT

Image retrieval is one of the fundamental tasks in computer vision. Despite nearly thirty years of research and significant progress, fine-grained retrieval remains challenging. Fine-grained visual retrieval aims to find target images in a database that are semantically similar to a query image, covering tasks such as instance retrieval (*e.g.*, person re-identification) and fine-grained retrieval (*e.g.*, bird species recognition). This task has wide applications in military reconnaissance, industrial production, urban security, medical diagnosis, *etc.*, such as identifying specific ships from remote sensing images or locking onto specific suspects from video surveillance. Therefore, fine-grained visual retrieval is currently a hot research topic in both industry and academia.

Existing deep learning methods have achieved some success in fine-grained visual retrieval tasks. However, deep learning relies on sample-label pairs for training and often only utilizes high-dimensional global features output by the model, leading to three unsolved challenges in fine-grained retrieval. Firstly, conventional deep learning methods only use the final features of the model for image retrieval. Due to insufficient feature resolution, these methods lack the ability to discern minor differences in images. Secondly, image annotation in such tasks is complex and error-prone, leading to label noise issues. Due to the difficulty of data collection, training data often obey a long-tailed distribution. Traditional supervised training methods lack robustness to these issues. Finally, due to the wide variety of fine-grained images and the large scale of databases, directly using high-dimensional real-valued features generated by deep networks leads to low retrieval efficiency. Therefore, this thesis innovatively introduces a deep model disentanglement learning strategy, addressing three major challenges in fine-grained visual retrieval—weak feature discriminative power, poor training robustness, and low retrieval efficiency, proposing a comprehensive solution. The main contributions are outlined as follows:

To address the issue of weak feature discriminative power, this thesis proposes a multi-scale feature disentanglement learning method to enhance the detail discrimination of deep

features by disentanglement and learning multi-resolution features. Firstly, this thesis proposes a dual-path network architecture to process high and low-resolution images separately, improving the expressive ability of features and cross-resolution matching accuracy. Secondly, this thesis introduces a text-guided global-local feature learning mechanism, even in the absence of local annotations, enabling the model to focus on highly discriminative detail regions with the guidance of textual information, further enhancing the model's understanding of key local details. This method can improve the discriminative power on multiple fine-grained retrieval datasets. For example, the proposed method achieves a 15.6% improvement in rank-1 accuracy on the SOP dataset for online product images compared to baseline methods.

To address the issue of poor training robustness, this thesis proposes a sample-label disentanglement optimization method to obtain more accurate training targets by disentanglement the binding relationship between samples and labels in the training set. Firstly, this thesis designs a label correction method based on neighborhood attention, integrating sample neighborhood label information, breaking the one-to-one mapping between samples and labels, predicting more accurate training targets, and alleviating overfitting caused by noisy labels. Secondly, this thesis proposes a sub-label learning strategy based on feature memory, learning multiple sub-labels to isolate noisy samples, balance long-tailed distributions, and strengthen image features. On data with noise and long-tailed distributions, the proposed method demonstrates robustness to noisy labels and long-tailed distributions. For example, the rank-1 performance of the baseline method on the person re-identification dataset Market1501 is 94.4%, while it drops to 69.2% with 50% label noise. The proposed method still maintains high accuracy of 92.2% at the same noise ratio.

To address the issue of low retrieval efficiency, this thesis proposes a multiple semantic attribute disentanglement encoding method to efficiently retrieve highly discriminative features from decoupled multi-scale features and encode them into compact binary features. This thesis utilizes the multi-level characteristics of conditional diffusion models to decouple the output of deep networks into multi-level features and selects information-rich parts through information entropy, encoding them into compact and highly discriminative binary features. The method demonstrates excellent performance on various image retrieval tasks, such as improving the performance of 32-bit binary features from 17.1% to 30.1% on the CUB fine-grained bird recognition dataset compared to the baseline method. Additionally, this thesis integrates the above research results and designs a learning method that simultaneously optimizes feature discrimination, robustness, and compactness, ensuring that binary features perform close

to real-valued features under biased data set training and increasing similarity calculation speed by 414 times.

In summary, this thesis constructs a fine-grained visual retrieval method based on deep model disentanglement learning, effectively improving the detail discrimination of features, training robustness to biased data, and retrieval efficiency by disentanglement the feature learning, model training, and feature encoding processes in fine-grained visual retrieval, which is of great significance for advancing the field of image retrieval.

KEY WORDS: fine-grained visual retieval, disentanglement learning, discriminative power, robust training, compactness