Computer Science and Application 计算机科学与应用, 2025, 15(1), 177-186
Published Online January 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.151018

Hans汉斯

基于CLIP模型的以图搜图方法

杨晓涵

河北地质大学信息工程学院，河北 石家庄

收稿日期：2024年12月23日；录用日期：2025年1月20日；发布日期：2025年1月28日

摘　要

近年来，人工智能技术在计算机视觉和自然语言处理领域的飞速发展，促进了两者间的深度融合，极大地拓展了智能系统的技术边界和应用前景。这种跨领域整合不仅推动了技术创新，也为诸多新颖研究和应用开辟了新的路径。本文提出了一种针对猫狗数据集和铁路相关数据集的图像检索方法——CLIP-Retrieval，旨在解决公开和专业领域中复杂背景、多角度拍摄等带来的图像检索挑战。CLIP-Retrieval利用CLIP模型的图像编码器作为核心架构，通过提取图像特征并构造相似度矩阵，计算不同图像之间的相似度分数，根据排序结果展示最相关的图像。为验证CLIP-Retrieval的鲁棒性和稳定性，我们进行了对比实验和抗干扰实验。实验结果显示，该算法在性能上有显著提升，具备良好的图像检索效果。具体而言，CLIP-Retrieval能够有效应对不同数据集中的复杂背景、姿态变化等问题，提供准确且高效的检索服务。

关键词

CLIP模型，以图搜图，图像检索

Image-to-Image Search Method Based on CLIP Model

Xiaohan Yang

College of Information Engineering, Hebei GEO University, Shijiazhuang Hebei

Received: Dec. 23rd, 2024; accepted: Jan. 20th, 2025; published: Jan. 28th, 2025

Abstract

In recent years, the rapid advancement of artificial intelligence technology in both computer vision and natural language processing (NLP) has facilitated deep integration between these fields, significantly expanding the technological boundaries and application prospects of intelligent systems. This cross-domain convergence not only drives technological innovation but also paves new paths for a myriad of novel research and applications. This paper introduces CLIP-Retrieval, an image

文章引用: 杨晓涵. 基于 CLIP 模型的以图搜图方法[J]. 计算机科学与应用, 2025, 15(1): 177-186.
DOI: 10.12677/csa.2025.151018

retrieval method designed specifically for the Cats vs. Dogs dataset and railway-related datasets. The goal is to address the challenges posed by complex backgrounds and multi-angle photography in both public and specialized domains. CLIP-Retrieval leverages the image encoder of the CLIP model as its core architecture, extracting image features and constructing a similarity matrix to compute similarity scores between different images. Based on the sorted results, it displays the most relevant images. To verify the robustness and stability of CLIP-Retrieval, we conducted comparison studies and interference resistance experiments. Experimental results show significant performance improvements, demonstrating excellent image retrieval effects. Specifically, CLIP-Retrieval effectively handles complex backgrounds and pose variations across different datasets, providing accurate and efficient retrieval services.

Keywords

CLIP Model, Image-to-Image Search, Image Retrieval

1. 引言

随着我国铁路轨道交通领域的迅猛发展，科技创新过程中产生的图像数据量急剧增长。在日常生产中，由于拍摄角度的差异，同一目标对象可能产生多角度的图像记录。而猫与狗数据集[1]中的图像展示了多种背景环境和拍摄角度，包括室内和室外场景、不同光照条件以及不同的拍摄距离。这些变化增加了图像的多样性，使得模型必须学会识别猫和狗的关键特征，而不仅仅是依赖于背景或特定的拍摄条件。

传统方法依赖人工提取图像中的低级特征，在面对海量图像时，因背景复杂性和对象类别的多样性而表现欠佳[2]。随着计算机视觉技术的进步，基于深度学习的图像检索方法[3]——特别是以图搜图技术取得了显著进展。内容相关的图像检索(Content-Based Image Retrieval, CBIR)作为一种基于图像内容特征的技术[4]，通过分析颜色、纹理、形状等视觉属性来寻找与用户查询图像相似的其他图像。CBIR 的核心在于计算图像间的距离或相似性度量，以实现无需文本描述或标签的图像检索。近年来，卷积神经网络(CNN) [5]凭借其强大的特征提取能力，在图像检索领域得到了广泛应用，如 Sun 等[6]提出的 circle loss 最小化类内相似度与最大化类间相似度，提升了车辆检索、人脸检索等多种任务的准确度；Revaud 等[7]提出 list-wise loss 优化全局检索排名；Ding 等[8]提出的 FaceNet 直接学习图像到欧几里得空间的映射，并通过空间距离衡量图像相似度。

相比以上方法，CLIP 模型[9]能够捕捉图像中的高级语义特征，有效应对复杂的背景信息和视角变化的影响。尽管 CLIP 模型已在图像分类与文本－图像检索等任务中展现出广泛应用前景，但在铁路交通图像检索方面，其潜力尚未得到充分挖掘。铁路轨道交通图像检索具有特殊性，如复杂背景、抓拍角度变化、目标对象姿态多样等因素，以及数据库中大量相似图片的存在，对检索结果产生了显著影响。这些特性使得直接应用现有的深度学习模型无法达到理想的检索效果。

鉴于此，本文提出采用 CLIP 模型的图像编码器，以高效提取公开数据集和铁路轨道交通专业数据集中的图像特征，从而提升图像检索的准确性和鲁棒性。CLIP 模型能够捕捉高级语义特征，有效克服背景信息复杂和视角变换带来的挑战。我们通过计算图像特征之间的余弦相似度来比较相似度分数，根据排序结果输出相似图像。这种方法不仅提高了检索精度，还为铁路轨道交通图像的分类和管理提供了强有力的支持。

2. CLIP 模型

CLIP (Contrastive Language-Image Pre-training)是一个基于图像和文本两个不同模态领域的数据训练出来的表征模型。部分研究表明，CLIP 具有强大的零样本学习能力[10]，其在开放图像数据集(如 ImageNet)上识别和分类图像方面表现出色，尤其是处理未曾见过的类别。CLIP 模型面对特定图像检索任务时，无需进行额外训练，就可以执行图像检索任务。意味着可以将其广泛应用于多种类型的图像检索任务，而且无需每次都重新训练模型。这不仅提高了效率，也扩展了其适用范围，包括那些难以获得足够标注数据的任务。CLIP 结构如下图 1 所示。

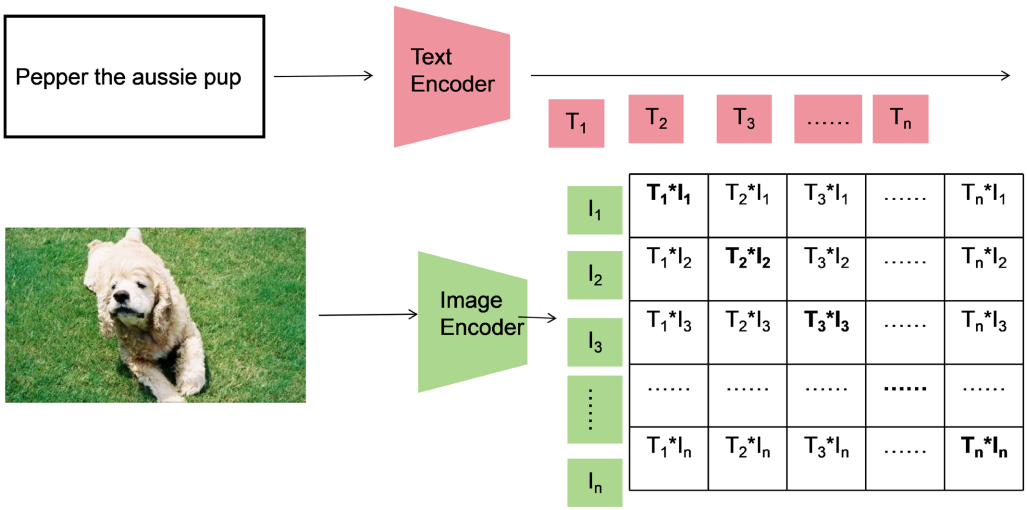

Figure 1. CLIP framework
图 1. CLIP 架构

CLIP 架构由两个主要部分组成：文本编码器和图像编码器。这两个编码器被联合训练被联合优化，目的是预测给定样本(图像，文本)的正确配对。在图 1 的左侧，文本编码器接收文本输入(例如：“Pepper the aussie pup”)，并将其转化为一系列文本特征 $(T_1,T_2,T_3,\cdots,T_n)$，捕捉文本的深层语义信息。相应地在图 1 的右侧，图像编码器处理图像输入，并生成图像的特征表示 $(I_1,I_2,I_3,\cdots,I_n)$，表示图像的视觉信息。图中的矩阵表示图像特征 $(I_1,I_2,I_3,\cdots,I_n)$ 与文本特征 $(T_1,T_2,T_3,\cdots,T_n)$ 的点积运算结果，被用来衡量它们之间的相似度。

然而，简单的点积并不能充分捕捉特征之间的相似性，为了更精确地检索到用户需求的结果，故采用余弦相似度来计算。具体来讲，余弦相似度是通过计算两个向量的点积并除以它们各自的模长得出结果的。最终，模型的目标是最大化相关图像和文本特征之间的余弦相似度，同时最小化不相关配对的相似度。通过这种方式，当给定一个文本描述或图像时，CLIP 能够在其训练过的多模态空间中找到最匹配的图像或文本。

3. CLIP-Retrieval 模型

本文采用 CLIP 模型的图像编码器作为图像特征提取结构，以提取图像特征。图像特征提取结构使用 Vision Transformer (ViT) [11]，提取特征向量后，经过 FFN 结构，对图像特征进行特征变换，捕捉更复杂的特征表示。然后建立特征矩阵，以便计算不同图像之间的余弦相似度[12]，匹配参考图像和数据库中所有图像，得出最为匹配的目标图像。整体架构如下图 2 所示。

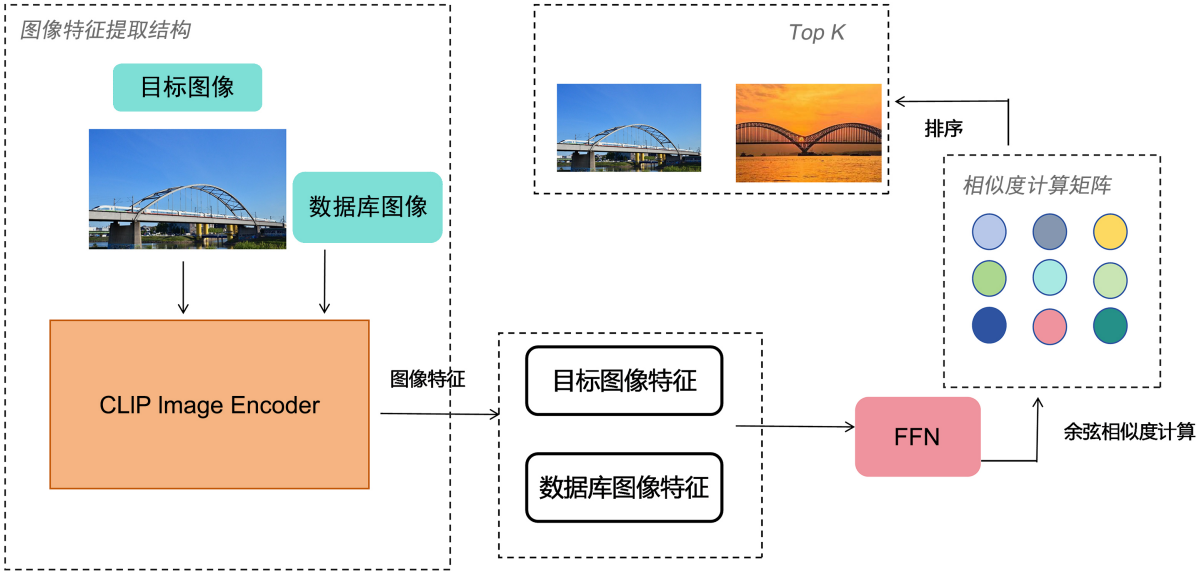

Figure 2. CLIP-retrieval framework
图 2. CLIP-retrieval 架构

3.1. 图像特征提取结构

图像特征提取结构基于 ViT，ViT 是一个将 transformer 直接应用于图像的模型架构。ViT 使用一维标记嵌入序列(Sequence of token embeddings)作为输入，输出为图像提取的特征向量。首先将输入图像 $x \in \Re^{H\times W\times C}$ 重塑为一个展平的 2 维 patches 序列 $x_p \in \Re^{N\times\left(P^2\cdot C\right)}$，其中(H, W)为原始图像分辨率，C 是原始图像通道数，P 是图像切割后的 patch，由此产生的图像 patch 数 $N = HW/P^2$。然后将 patch 展平，使用可训练的线性投影层将维度 $P^2\cdot C$ 映射为隐藏层维度，即 D 维，并将得到 patch Embeddings 输入至 Transformer 编码器中，经过多头注意力机制、层归一化和跳跃链接，就可以的到对应的图像表示。计算过程如下：

$$z_0 = \left[x_{class}; x_p^1 E; x_p^2 E; \cdots; x_p^N E;\right] + E_{pos} \tag{1}$$

$$z_l' = \mathrm{MSA}\left(\mathrm{LN}\left(z_{l-1}\right)\right) + z_{l-1} \tag{2}$$

$$z_l = \mathrm{MSA}\left(\mathrm{LN}\left(z_l'\right)\right) + z_l' \tag{3}$$

$$y = \mathrm{LN}\left(z_L^0\right) \tag{4}$$

其中，x_{class} 为可学习的类别标记，$x_p^N E$ 为对应的 patch Embedding，E_{pos} 是对应的位置编码。而 MSA、MLP、LN 为 transformer 的架构，分别代表多头注意力机制、多层感知机和归一化层，y 为得到的图像特征向量。ViT 的结构图如图 3 所示。

经过 ViT 结构之后，得到的图像再经过一个 FFN 层，对图像特征进行一个维度调整。然后再进入相似度计算模块进行特征之间的相似度计算。FFN 层结构包括一个全连接层(FC)、relu 激活层和第二个全连接层。计算过程的数学表达式为：

$$m = \mathrm{FC}\left(\mathrm{relu}\left(\mathrm{FC}\left(y\right)\right)\right) \tag{5}$$

其中，FC 表示为全连接，relu 表示为激活函数。

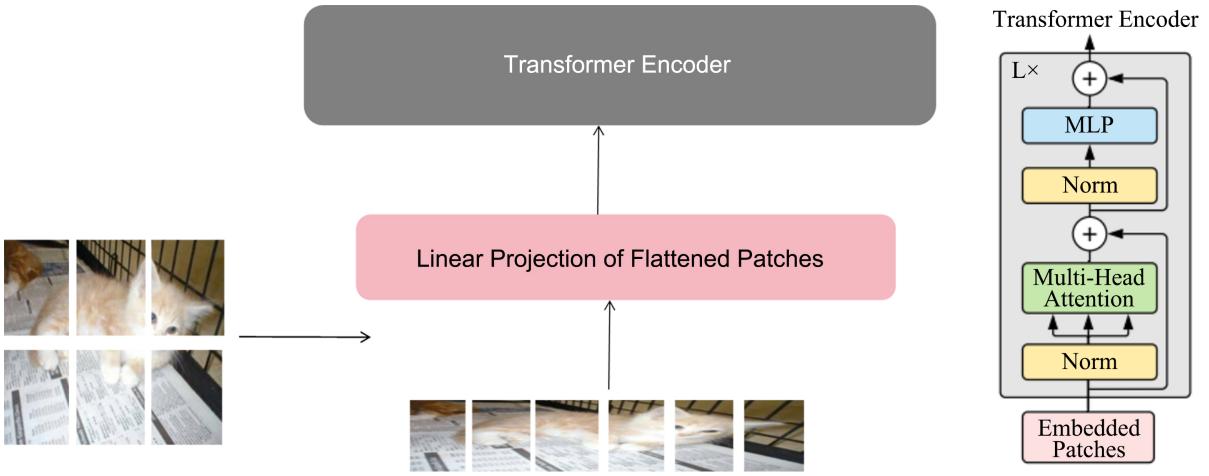

Figure 3. ViT framework
图 3. ViT 架构

3.2. 相似度计算

根据图像特征提取结构提取到的图像特征，构建余弦相似度矩阵。对于不同图像的特征，首先使用矩阵乘法计算图像特征向量之间的内积，计算过程为：

$$\hat{s} = a \cdot b \tag{6}$$

其中 a 和 b 均为图像特征提取模块得到的特征向量。然后计算每个图像特征向量的 L2 范数，计算过程为：

$$s = \|a\|_2 \cdot \|b\|_2 \tag{7}$$

之后就可以得到余弦相似度，定义为：

$$\cos(\theta) = \frac{\hat{s}}{s} \tag{8}$$

根据余弦相似度定义就可以构建一个余弦相似度矩阵。利用这个矩阵，选取分数最大的 K 个值，并根据对应的索引，就可以在数据库中找到相似的图像并输出。具体流程如上图所示。

4. 实验数据集

猫与狗数据集是计算机视觉领域中广泛使用的公开数据集之一，主要用于图像分类任务。该数据集由 Kaggle 提供，包含大量的猫和狗的彩色图像，大约有 25,000 张标记为“猫”或“狗”的图像，分为训练集和测试集两部分。训练集中有 12,500 张猫的图像和 12,500 张狗的图像，共计 25,000 张。测试集则包含 12,500 张未标注的图像，用于评估模型性能。部分数据展示如图 4 所示。

鉴于铁路轨道交通涉及的目标对象多样，本文选取了出现频率较高的 5 个标签作为图像检索的依据。通过网络下载获取图像数据，并人工删除重复项，最终构建了一个铁路相关图像的数据集。然后进行图像预处理，首先是去重，人工筛选并移除重复图像，确保数据集的质量。其次是等比例缩放，对图像进行等比例缩放，统一调整大小，确保图像不发生变形或扭曲。最后是数据集划分，将处理好的数据集按照一定比例划分为训练集和测试集，以支持模型训练和评估。

本次实验共设定了 5 个类别，每个类别包含 5 组图片，具体分类包括：接触网、高速列车、铁轨、铁路、高铁站等。部分数据展示如图 5 所示。


Figure 4. Cat and dog partial data display
图 4. 猫狗部分数据展示


Figure 5. Rail partial data display
图 5. 铁路部分数据展示

5. 消融实验

将本文算法的主干模型 CLIP 与主流图像提取特征算法，如 Resnet [13]、VGG [14]等架构基于猫狗数据集进行比较，并通过计算不同类别的相似度平均分数对比实验结果(表 1)。

Table 1. Algorithm comparison
表 1. 算法对比

主干网络	cat	dog
Resnet50	85%	84%
Resnet152	87%	86%
VGG	90%	89%
CLIP-Retrieval	94%	92%

6. 结果可视化

通过本文模型的高效处理和分析，能够确保各类图像的准确分类和检索，从而提供更为精准的图像检索服务。每组展示包括 7 张图片：最左侧的一张是用于检索的参考图像(标记为 query:image:*.jpg)，而右侧的 6 张是根据相似度检索出的目标图像。这些目标图像按照与参考图像的相似度降序排列，其中相似度分数(标记为 score: ***)越高表示两张图像越相似。每张目标图像下方有两行标注：第一行显示其与参考图像的相似度分数，第二行显示该图像在检索数据集中的文件名。可以根据需求调整检索出的相似图像的数量，如结果可视化所示，检索出 6 张相似图像。

首先是公开数据集，即猫与狗数据集中的部分测试结果，设置的 K 值为 5，可视化展示如图 6 所示。

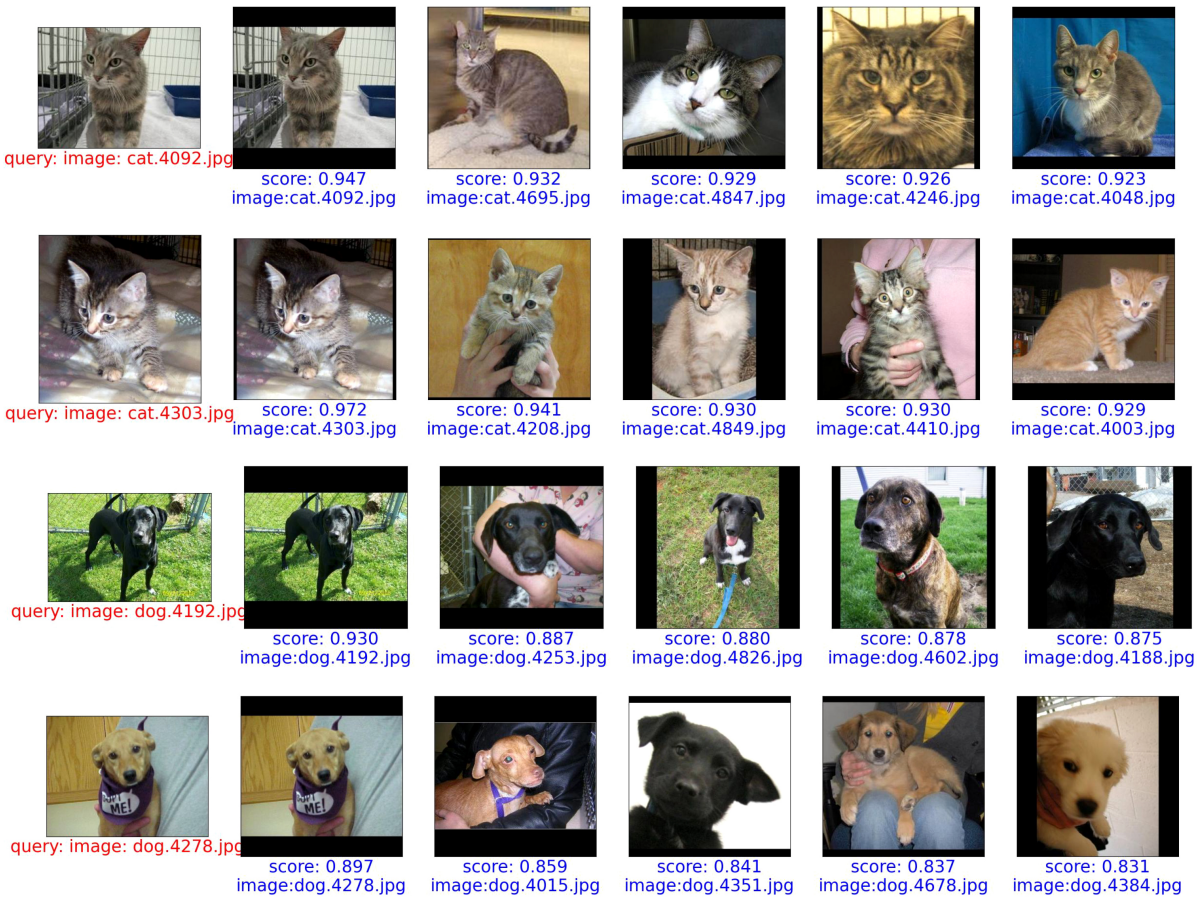

Figure 6. Visualization of cat and dog test result
图 6. 猫狗测试结果可视化

其次是铁路相关数据集，按照设置的类别数据顺序，依次按照接触网、高速列车、铁轨、铁路、高铁站进行测试结果可视化展示，设置的 K 值为 5，具体如图 7 所示。

在对 CLIP-Retrieval 模型于图像检索任务的测试过程中，我们采用了多类数据进行全面评估，涵盖公开数据集以及专业数据。实验结果表明，该模型在不同数据来源下均展现出卓越的性能。对于每一幅目标图像，CLIP-Retrieval 模型均能精准地在数据库中定位到与之相似的图像，有力地证明了其在图像检索应用中的有效性与可靠性，充分彰显了模型良好的泛化能力与适应性，为图像检索技术在复杂实际场景中的应用提供了坚实的支撑与保障。

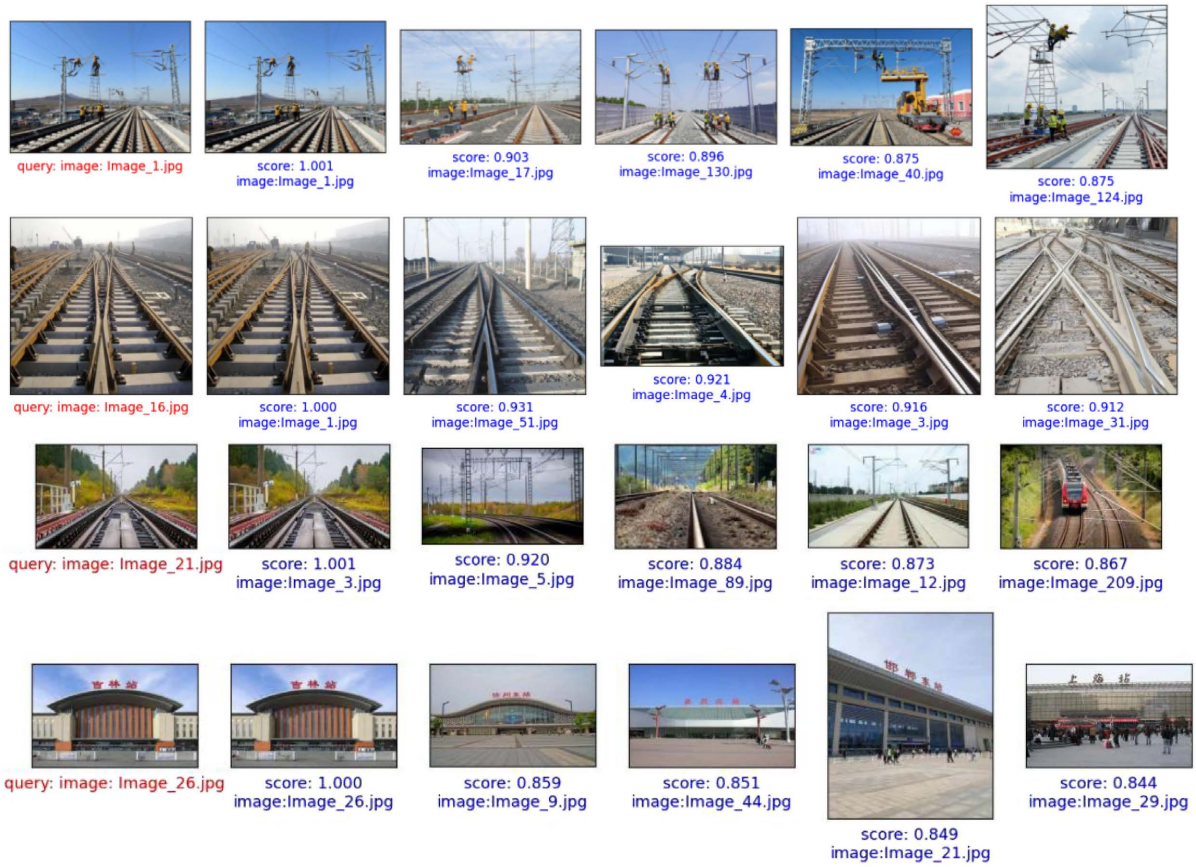

Figure 7. Visualization of high-speed train station test results
图 7. 高铁站测试结果可视化

7. 抗干扰实验

Figure 8. Roller coaster interference
图 8. 过山车干扰

在针对模型稳定性开展测试工作时，我们于铁路相关数据里添加了部分干扰数据。这些干扰数据提取自 COCO 数据集，涵盖了诸如游乐场过山车、铁架以及电线杆等可能会对模型产生干扰的图片类型，以此来检验模型在复杂数据环境下的稳定性表现。首先，在铁轨类别数据中添加游乐场过山车的图片，将本体(铁轨相关数据)与干扰项(过山车图片)按照约 2:1 的比例进行配置。随后，选取五组仅包含过山车的测试图像作为测试数据，测试结果如图 8 所示。

以下输入的测试图像均为铁轨，选取一组测试数据，测试结果如图 9 所示。

Figure 9. Rail track
图 9. 铁轨本体

从测试结果能够明显看出，即便在铁轨数据中增添了过山车这类干扰项，铁轨图像的检索效果依然未受显著影响，其精准度与效率保持在较为稳定的水平。接下来，针对接触网类别开展进一步测试，在其中加入铁架、电线杆的图片，并且将本体(接触网相关数据)与干扰项(铁架、电线杆)按照约 1:1 的比例进行混合。然后，选取两组为干扰图像(即铁架或电线杆图像)的测试数据，其测试结果如图 10，图 11 所示。

Figure 10. Power pole interference
图 10. 电线杆干扰

Figure 11. Steel framework interference
图 11. 铁架干扰

根据上述测试结果能够清晰地发现，在针对干扰项进行检测的过程中，未出现接触网相关图像被误检的情况。通过以上三组抗干扰实验测试，有力地证实了 CLIP-Retrieval 算法具备良好的稳定性，其在面对不同类型的干扰数据时，依然能够精准且稳定地运行，有效保障了图像检索等相关任务的准确性与可靠性，展现出该算法在复杂数据环境下卓越的抗干扰能力与稳健的性能表现。

8. 结论

本研究聚焦于新型深度学习框架的探索与开发，致力于扩展图像检索的应用范围，提升图像检索的

性能效果。在大规模异构数据集的对比学习进程中，我们力求促使模型于无明确标注情境下，依然能够精准高效地开展图像检索任务，这在以图搜图应用场景中具备尤为关键的价值。实验数据表明，CLIP-Retrieval 模型在公开数据集与专业数据集的测试中，皆斩获了较高的准确率，这一成果深刻彰显出其非凡卓越的图像理解实力。并且在抗干扰实验中，CLIP-Retrieval 也展示了良好的算法稳定性。然而，不可忽视的是，尽管 CLIP 模型已获取显著成效，但其自身仍存在一定局限性，例如对海量数据的高度依赖以及对高额计算资源的大量需求等方面。基于此，未来的研究工作将着重围绕探索新型算法与前沿技术展开，力求全方位提升 CLIP 的计算效率，从而使其能够在资源相对受限的应用环境中，切实发挥出更为卓越的实用性与适配性。

参考文献

[1] 燕程, 王志聪, 燕鹏. 基于 VGG 网络对猫狗数据集分类[J]. 数码设计(下), 2021, 10(5): 18.

[2] 王志瑞, 闫彩良. 图像特征提取方法的综述[J]. 吉首大学学报(自然科学版), 2011(5): 41-42.

[3] 赵学敏, 田生湖, 张潇璐. 基于深度学习的以图搜图技术在照片档案管理中的应用研究[J]. 档案学研究, 2020(4): 64-68.

[4] Choras, R.S. (2007) Image Feature Extraction Techniques and Their Applications for CBIR and Biometrics Systems. *International Journal of Biology and Biomedical Engineering*, **1**, 6-16.

[5] Rodriguez-Vazquez, A., Espejo, S., Dominguez-Castron, R., Huertas, J.L. and Sanchez-Sinencio, E. (1993) Current-mode Techniques for the Implementation of Continuous- and Discrete-Time Cellular Neural Networks. *IEEE Transactions on Circuits and Systems II*: *Analog and Digital Signal Processing*, **40**, 132-146. https://doi.org/10.1109/82.222812

[6] Sun, Y., Cheng, C., Zhang, Y., Zhang, C., Zheng, L., Wang, Z., *et al*. (2020) Circle Loss: A Unified Perspective of Pair Similarity Optimization. 2020 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 14-19 June 2020, 6397-6406. https://doi.org/10.1109/cvpr42600.2020.00643

[7] Revaud, J., Almazan, J., Rezende, R. and Souza, C.D. (2019) Learning with Average Precision: Training Image Retrieval with a Listwise Loss. 2019 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 5106-5115. https://doi.org/10.1109/iccv.2019.00521

[8] Ding, H., Zhou, S.K. and Chellappa, R. (2017) Facenet2expnet: Regularizing a Deep Face Recognition Net for Expression Recognition. 2017 12*th IEEE International Conference on Automatic Face & Gesture Recognition* (*FG* 2017), Washington DC, 30 May-3 June 2017, 118-126. https://doi.org/10.1109/fg.2017.23

[9] You, H., Zhou, L., Xiao, B., Codella, N., Cheng, Y., Xu, R., *et al*. (2022) Learning Visual Representation from Modality-Shared Contrastive Language-Image Pre-Training. *Computer Vision ECCV* 2022 17*th European Conference*, Tel Aviv, 23-27 October 2022, 69-87. https://doi.org/10.1007/978-3-031-19812-0_5

[10] Wang, W., Zheng, V.W., Yu, H. and Miao, C. (2019) A Survey of Zero-Shot Learning: Settings, Methods, and Applications. *ACM Transactions on Intelligent Systems and Technology*, **10**, 1-37. https://doi.org/10.1145/3293318

[11] Dosovitskiy, A., Beyer, L., Kolesnikov, A., *et al*. (2021) An Image Is Worth 16 x 16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations*, Vienna, 4 May 2021, 37.

[12] Nguyen, H.V. and Bai, L. (2011) Cosine Similarity Metric Learning for Face Verification. *Computer Vision ACCV* 2010 10*th Asian Conference on Computer Vision*, Queenstown, 8-12 November 2010, 709-720. https://doi.org/10.1007/978-3-642-19309-5_55

[13] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/cvpr.2016.90

[14] Simonyan, K. and Zisserman, A. (2015) Very Deep Convolutional Networks for Large-Scale Image Recognition. *International Conference on Learning Representations* (*ICLR*), San Diego, 7-9 May 2015, 1-14.