中图法分类号:TP301.6　　**文献标识码**:　　**文章编号**:1006-8961(XXXX)XX-0001-14

论文引用格式:Zhang Ke, Gao Mingwei, Zheng Zhaoye, Li Shuoshi, Nie Ding, Zhou Shuai. XXXX. Transmission line bolt defect classification method integrating vision-language feature alignment and textual decoupling. Journal of Image and Graphics, XX(XX):0001-0014(张珂, 高明伟, 郑朝烨, 李硕士, 聂鼎, 周帅. XXXX. 融合图文特征对齐和文本解耦的输电线路螺栓缺陷分类方法. 中国图象图形学报, XX(XX):0001-0014)[DOI:10.11834/jig.250468]

融合图文特征对齐和文本解耦的输电线路螺栓缺陷分类方法

张珂[1,3,4],高明伟[2,5],郑朝烨[5],李硕士[1,3,4],聂鼎[6],周帅[6]

1. 华北电力大学燕赵电力实验室,保定　071003; 2. 云南电网有限责任公司人才工作站,昆明　650521; 3. 华北电力大学河北省电力物联网技术重点实验室,保定　071003; 4. 电力物联智慧化技术河北省工程研究中心,保定　071003; 5. 华北电力大学电子与通信工程系,保定　071003; 6. 云南电网有限责任公司电力科学研究院,昆明　650521

摘　要:**目的**　视觉-语言模型在缺乏视觉信息的输电线路螺栓缺陷分类中潜力巨大,然而以图像-文本对比学习为代表的视觉-语言模型预训练算法需要大规模的图像和文本数据,实际螺栓缺陷分类研究中数据规模通常较小且缺少丰富的描述文本。为提升视觉-语言模型在螺栓缺陷分类上的适配度,本文提出一种融合图文特征对齐和文本解耦的输电线路螺栓缺陷分类方法。**方法**　首先,本文深入分析了图像-文本对比学习不适配螺栓缺陷分类任务的原因,归结为类内语义重叠的影响;然后,针对零样本CLIP模型中螺栓缺陷类间相似度较高的问题,提出文本特征解耦(text feature decoupling, TFD)损失函数使不同缺陷类别文本在特征空间有效解耦;在此基础之上,针对图像-文本对比学习中的类内语义重叠问题,提出适配螺栓缺陷分类的文本锚点引导的图像特征对齐方法(text-anchor guided visual feature alignment, TA-VFA),实现同类别图像和文本特征对齐。同时,为抑制过拟合问题并提升模型泛化能力,提出渐进式图像编码器微调策略(progressive fine-tuning for the image encoder, PFT),逐层解冻图像编码器的Transformer层参数以实现稳定适配。**结果**　在螺栓缺陷数据集上的实验结果表明,VL-Bolt模型能够有效降低类间相似度并实现图像文本特征对齐,分类准确率达到92.2%,相比基线模型提升3.9%。同时,VL-Bolt在与CLIP-Adapter、WiSE-FT等迁移方法的对比中表现更优。相比于端到端微调,渐进式图像编码器微调策略能够有效稳定训练过程,抑制过拟合问题。进一步在Flowers102、StanfordCars与Caltech101三个公开数据集上进行验证,VL-Bolt的准确率均较端到端微调模型取得显著提升,证明了该方法的鲁棒性。**结论**　本文所提出的方法有效提升了视觉-语言模型在螺栓缺陷分类任务中的性能,为视觉-语言模型在输电线路巡检中的实际应用提供了新的思路。

关键词:输电线路;螺栓缺陷分类;图文特征对齐;视觉-语言模型;特征解耦;迁移学习

Transmission line bolt defect classification method integrating vision-language feature alignment and textual decoupling

Zhang Ke[1,3,4], Gao Mingwei[2,5], Zheng Zhaoye[5], Li Shuoshi[1,3,4], Nie Ding[6], Zhou Shuai[6]

1. Yanzhao Electric Power Laboratory of North China Electric Power University, Baoding 071003, China; 2. Talent Workstation of Yunnan Power Grid Co., Ltd., Kunming 650521, China; 3. Hebei Key Laboratory of Power Internet of Things Technology; North China Electric Power University, Baoding 071003, China; 4. Hebei Engineering Research Center of Intelligent Technology for Power Internet of Things, Baoding 071003, China; 5. Department of Electronic and Communication Engineering, North China Electric Power University, Baoding 071003, China; 6. Electric Power Research Institute, Yunnan Power Grid Co., LTD., Kunming 650521, China

收稿日期:2025-09-25;**修回日期**:2025-12-24

基金项目:国家自然科学基金(62076093,62206095,61871182);中央高校基本科研业务费专项资金(2023JG002,2022MS078,2023JC006);

Supported by:National Natural Science Foundation of China(62076093,62206095,61871182),

Abstract: **Objective** Intelligent analysis and processing of UAV-captured images using computer vision has become the mainstream approach for power transmission line inspection. As the most widely used connection components in transmission lines, bolts play a critical role in structural integrity. However, due to complex environmental conditions, bolts are prone to defects such as losing nuts and losing pins, which may pose serious threats to the safe and stable operation of transmission systems. Therefore, bolt defect classification based on computer vision techniques is an essential task to ensure transmission reliability. Visual-language models (VLMs) hold great potential in bolt defect classification tasks where visual information is limited. However, vision-language pretraining approaches based on image-text contrastive learning typically require large-scale datasets and diverse textual descriptions. For instance, general-domain multi-modal datasets often comprise millions of image-text pairs, while bolt defect datasets are characterized by small-scale data and limited textual descriptions. As a result, contrastive learning approaches cannot be directly applied to such tasks. To expand the applicability of vision-language models in bolt defect classification, this paper proposes a transmission line bolt defect classification method integrating vision-language feature alignment and textual decoupling (VL-Bolt). **Method** We begin by analyzing the alignment behavior of image-text pairs within a training batch and identify a key limitation of standard image-text contrastive learning: intra-class semantic overlap. When multiple samples from the same class are present in a batch, image-text contrastive learning incorrectly treats identical class descriptions as negative pairs, hindering effective alignment between images and their textual counterparts. This issue is particularly pronounced in small-scale bolt defect datasets with limited textual diversity. Then, we adopt the CLIP ViT-B/32 model as the backbone. However, due to the lack of power domain knowledge in CLIP's pretraining and the high similarity between different bolt defect features, its performance in fine-grained classification is constrained. To address this, we introduce a text feature decoupling (TFD) loss, which explicitly disentangles textual features of different defect categories in the feature space. Building upon the decoupled textual features, we further design a text-anchor guided visual feature alignment (TA-VFA) strategy tailored to bolt defect classification. In TA-VFA, decoupled text features serve as anchors to guide the alignment of their corresponding image features, facilitating effective separation of image representations across different defect categories. This architecture inherently resolves the intra-class semantic overlap issue, making it well-suited to the bolt defect classification task. Moreover, considering the limited scale of the bolt defect dataset, we propose a progressive fine-tuning strategy for the image encoder (PFT) to mitigate overfitting and improve generalization. Specifically, only Transformer layer 11 is unfrozen during early training, and additional layers are gradually unfrozen in later stages. This enables the model to prioritize learning high-level semantic features relevant to bolt defects while retaining low-level structural information and the generalization ability gained during pretraining. **Results** In comparative experiments, VL-Bolt achieves a classification accuracy of 92.2% on the bolt defect dataset, surpassing the baseline by 3.9% and outperforming multiple ImageNet-pretrained vision models as well as four recent bolt defect classification models fine-tuned on bolt datasets. Compared with the multi-modal MUCO-BD model, the proposed VL-Bolt model exhibits a slight gap in classification accuracy; however, it achieves approximately 29.9% fewer trainable parameters. Moreover, when the VL-Bolt model is pre-trained with domain-specific knowledge following the same configuration as MUCO-BD, its performance surpasses that of MUCO-BD, demonstrating the superior adaptability of the proposed approach. Furthermore, we compare our transfer strategy against four recently proposed transfer learning methods. Results demonstrate that VL-Bolt achieves superior accuracy and generalization performance. From the perspectives of algorithm principle and structure, the main reason for the poor classification performance of the CLIP-Adapter, CoOp, CoCoOp, and Tip-Adapter models lies in the excessively high inter-class similarity of bolt defects. Compared with WiSE-FT, a transfer method designed to enhance model generalization, VL-Bolt outperforms it by 2.2%, which demonstrates the generalization advantage of the proposed transfer method in this study. To verify the robustness of the proposed method, this study applies VL-Bolt to three publicly available datasets and compares its classification performance with that of end-to-end fine-tuning. On the fine-grained classification datasets Flowers102 and StanfordCars, VL-Bolt achieves significant improvements, particularly on StanfordCars, where it surpasses the end-to-end fine-tuning model by 10.7%. Furthermore, on the Caltech101 dataset, which is characterized by large inter-class variations, VL-Bolt still demonstrates strong task adaptability, further validating the robustness of the proposed approach. Ablation studies confirm the contribution of each component of our method and its bolt classification task compatibility. The TA-VFA architec-

ture enables the avoidance of the intra-class semantic overlap issue, and its integration leads to a significant improvement in accuracy. The text feature decoupling mechanism can effectively mitigate inter-class confusion to a certain extent. Notably, the progressive image encoder fine-tuning strategy guides the model to focus on target semantics, which contributes the most prominent improvement to the accuracy. To identify the optimal layer depth for progressive fine-tuning, we analyze training loss and validation accuracy curves, and find that unfreezing only Transformer layers 5 to 11 achieves the best performance. Compared with end-to-end fine-tuning, the progressive image encoder fine-tuning strategy effectively stabilizes the training process and mitigates overfitting. In visualization experiments, the near-orthogonality of inter-class text feature similarities validates the effectiveness of the proposed text feature decoupling loss. Moreover, using t-SNE, we visualize the feature spaces learned by image-text contrastive learning, CLIP ViT-B/32 and VL-Bolt. VL-Bolt shows better inter-class separability with more natural boundary distributions, striking a reasonable balance in feature representation. Finally, Grad-CAM visualizations further reveal that VL-Bolt focuses more on bolt targets and defect regions while reducing attention to irrelevant background areas, thereby providing a solid foundation for downstream classification tasks. **Conclusion** The proposed VL-Bolt significantly improves the classification performance of vision-language models in bolt defect classification tasks. Specifically, the TFD loss and TA-VFA architecture demonstrate strong task adaptability and effectively mitigate the challenge of intra-class semantic overlap while reducing the burden of data preparation. Moreover, the progressive fine-tuning strategy for the image encoder suppresses overfitting and directs the model's attention toward defect-relevant regions. Overall, this work provides new insights into the practical application of vision-language models in power transmission line inspection.

Key words: transmission line; bolt defect classification; vision-language feature alignment; vision-language model; feature decoupling; transfer learning

0 引　言

近年来，我国在输电线路巡检领域的研究取得了显著发展，利用计算机视觉技术对无人机航拍图像进行智能分析与处理成为主流的巡检方式（赵振兵等，2020）。其中，螺栓作为输电线路中应用最广泛的连接部件，常用于连接金具和承力构件，起到关键的固定作用。然而，输电线路长期处于野外中，受复杂外部环境影响（赵振兵等，2021），连接各个部件的螺栓易产生螺母缺失、脱销等多种缺陷。螺栓体积小且缺陷类型复杂，致使螺栓缺陷识别更加困难，这些缺陷若未能及时发现，将对输电线路的安全稳定运行造成严重威胁（李学渊等，2022）。因此，通过计算机视觉技术提升螺栓缺陷识别准确率是确保输电线路正常工作的必要研究。

当前，输电线路巡检图像主要依赖直升机和无人机承载摄像头拍摄（顾超越等，2020），受拍摄距离远和螺栓尺寸小等因素影响，实际获取的螺栓图像常常分辨率较低（廖瑞金等，2018）。部分研究从提升螺栓图像质量方面做出探索：戚银城等人（2023）利用螺栓之间相似度较高的特点，将特征迁移引入螺栓图像超分辨率处理中，获取更加清晰的螺栓超分图像，从而提升分类准确率；Zhang等人（2024）提出了一种基于双判别器结构和伪增强策略的螺栓缺陷图像生成方法，通过生成高保真的合成图像，显著提升了分类性能。部分研究考虑从增强螺栓语义信息入手：赵振兵等人（2021）基于螺栓与螺母之间的关联，利用门控图神经网络构建栓母对知识图谱指导栓母对缺陷分类；张姝等人（2021）利用图像切割和数据增强处理技术，增大螺栓在图像中的占比，再利用YOLOv3模型实现螺栓缺陷识别。部分研究采用优化模型结构的方法：Liu等人（2022）提出了一种融合注意力机制与宽残差网络的螺栓缺陷分类方法，充分利用关于螺栓的先验知识。

近年来，随着多模态学习技术的发展，利用视觉与文本信息协同建模成为提升模型认知能力的新趋势（Zhang等，2024）。多模态深度学习旨在融合图像、语言、音频等多源数据，提高模型处理复杂任务的能力（Ramachandram等，2017）。其中，语言-图像对比预训练模型（contrastive language-image pre-training，CLIP）利用从网络获取的4亿个图文对，通过图像-文本对比学习进行预训练，实现图像文本对齐（Radford等，2021）。从网络获取的图文对数据集

文本描述多样且样本数量巨大，赋予了模型强大的语义理解能力。但是，零样本CLIP模型在细粒度语义感知和推理方面具有局限性（Jiang等，2023；Khan等，2023；Momeni等，2023），而不同螺栓缺陷类别的语义特征又高度相似，因此该模型无法直接用于螺栓缺陷分类。此外，在实际研究中，螺栓缺陷的样本数量和文本描述多样性无法达到预训练数据集的规模，这使得模型难以通过图像-文本对比学习充分学习到螺栓缺陷相关的语义信息和先验知识，并且容易导致严重的过拟合问题。因此，仅利用螺栓缺陷数据集进行图像-文本对比学习会导致次优的效果。

近期，有研究尝试解决上述问题。MUCO-BD模型（multi-modal contrastive learning for bolt defect classification）（张珂等，2025）通过引入额外的输电线路金具图像并构建多样化文本描述，为图像-文本对比学习提供了丰富的数据，使模型在预训练中学习到大量输电线路金具知识。该方法在提升螺栓缺陷分类性能方面表现突出，但进行大规模数据扩充和文本构建会带来显著的计算开销和较高的标注成本。

在降低类间特征相似度方面，相关研究提供了新的启发。Li等人（2019）在细粒度检索任务中，将每一类别的特征分布通过类别中心进行建模，并在类别中心间施加解相关约束，有效降低了类间特征相似度，提升了模型对细粒度差异的区分能力。Lezama等人（2018）提出一种正交低秩嵌入（orthogonal low-rank embedding）损失，用于将每个类别的特征压缩到一个子空间，同时不同类别子空间之间接近正交，从而增强类间区分。这种“类间去相关”思想为多模态任务中的特征分离提供了有益参考。

部分研究就图像-文本模态对齐的方法进行探索。ViLT模型（Kim等，2021）采用统一的Transformer架构，将图像patch与文本token共同输入模型，通过跨模态注意力机制实现了显式的图像-文本特征对齐，显著提升了语义融合能力。UniCL（Yang等，2022）通过统一的多模态对比学习框架，将监督分类与跨模态对齐有机结合，为多模态预训练和微调提供了新的思路。

为提升视觉-语言模型在螺栓缺陷分类上的适配度，本文提出基于特征解耦和多模态对齐的螺栓缺陷分类方法（transmission line bolt defect classification method integrating vision-language feature alignment and textual decoupling，VL-Bolt）。该方法在仅利用螺栓缺陷数据集、单条文本描述的条件下，完成了细粒度的螺栓缺陷分类，并有效缓解了过拟合问题。首先，本文指出在螺栓缺陷数据集上进行图像-文本对比学习容易受到类内语义重叠的影响，从而引发对比学习冲突；然后，针对螺栓缺陷类间特征相似度高的问题，提出文本特征解耦损失函数，使不同缺陷描述之间接近正交，从而实现文本特征解耦。接着，基于已解耦文本特征，提出一种文本锚点引导的图像特征对齐方法，实现了不同螺栓缺陷图像特征在特征空间中的有效分离。此外，本文提出渐进式图像编码器微调策略，通过逐层解冻图像编码器的Transformer层参数以实现稳定适配，有效缓解了过拟合问题。

1 螺栓缺陷数据集类内语义重叠

本节就CLIP模型采用的图像-文本对比学习不适配螺栓缺陷分类的原因进行分析。螺栓缺陷数据集（bolt defect classification dataset，BDCD）包含正常螺栓、缺销螺栓、缺螺母螺栓、螺栓缺失4种螺栓类别，共计4740张，该数据集为图像数据集，文本描述需要人工设计。目前，通用领域常用的预训练数据集规模较大，如：多模态数据集LAION-400M包含约4亿个图文对；传统监督预训练数据集ImageNet-1K图片数量约为128万。可见螺栓缺陷数据集与通用领域的预训练数据规模差距较大。

CLIP模型的预训练数据集是从互联网收集的4亿个图像文本对，图像数量庞大，文本描述内容丰富、表述形式多样，在数据丰富的条件下，图像-文本对比学习表现出强大的泛化能力。然而，螺栓缺陷数据集规模较小、文本描述相对单一，类内语义重叠可能会引发对比学习冲突，影响图像文本对齐效果。

图1展示了在螺栓缺陷数据集上进行图像-文本对比学习时出现类内语义重叠的情况。例如，对于第1张图像“正常螺栓”，图像-文本对比学习会最大化其与“A photo of normal bolt”的相似度（粉红色箭头），最小化与其他类别文本如“A photo of bolt losing”等的相似度（蓝色虚线箭头）。然而，当该批次中出现另一张同属“正常螺栓”的图像（第4张）时，其对应的描述“This is a normal bolt”与“A photo of normal bolt”虽然语义相同，但是在当前对比学习框

架下“This is a normal bolt”仍会被视为负样本，进而引发冲突：模型在提高图像与“A photo of normal bolt”的相似度的同时，会错误地降低其与“This is a normal bolt”的相似度（深红色箭头）。这种一个批次中的类内语义重叠会导致同类图像文本无法有效对齐。

该问题在样本数量有限、类别数量较少且文本描述多样性不足的螺栓缺陷数据集上尤为突出。此前一些研究并未直接关注该问题，但是采用的部分策略在一定程度上起到了缓解作用。CLIP的大规模数据训练通过随机打乱可以降低批次内同类样本的共现概率；UniCL通过将图像文本两种数据源整合到统一的图像-文本-标签空间为模型指明同类文本；MUCO-BD通过扩充金具数据集和设计多样化文本描述来解决图像样本不足和文本形式单一问题。

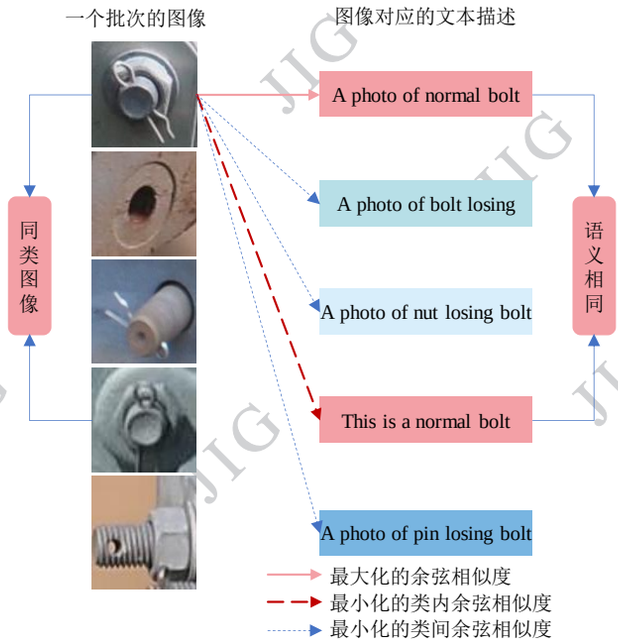


图1　类内语义重叠示意图

Fig. 1　Intra-class semantic overlap

2　VL-Bolt算法原理

本文方法结构如图2所示，主干选用CLIP ViT-B/32模型。训练分为两个阶段：第一阶段，采用文本特征解耦损失微调CLIP模型的文本编码器，促使不同类别缺陷描述文本在特征空间中充分

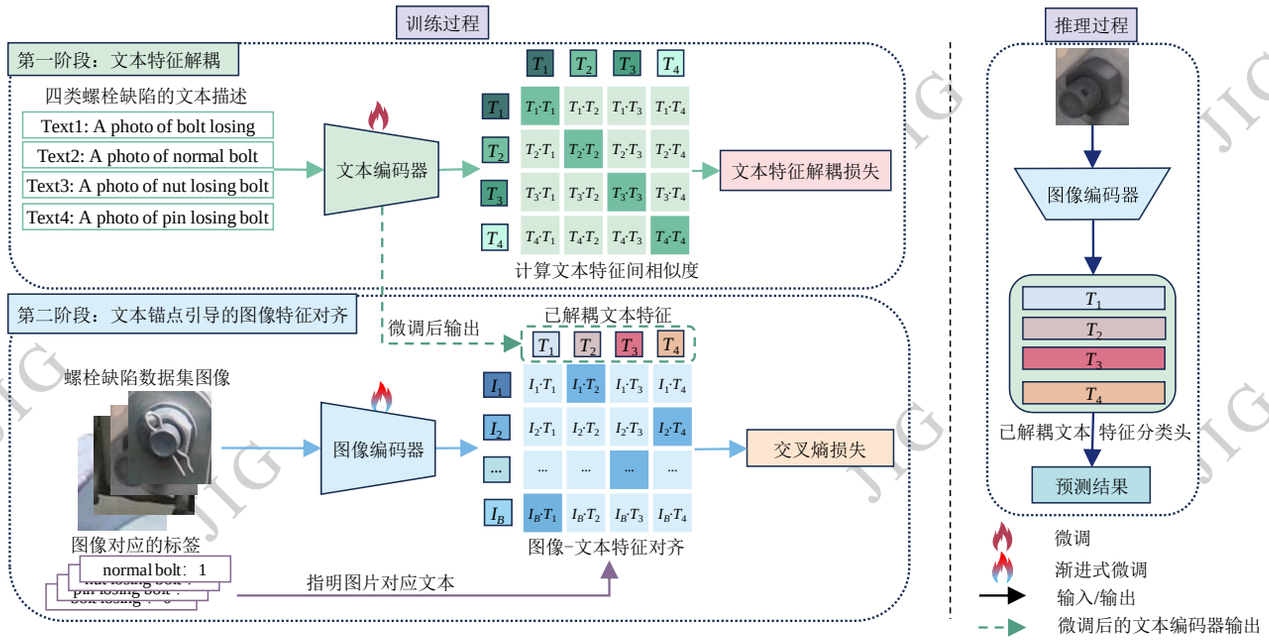


图2　VL-Bolt算法概览

Fig. 2　An overview of VL-Bolt

解耦；第二阶段，将已解耦文本特征作为类别锚点，利用交叉熵损失函数实现图像特征与对应的缺陷文本特征对齐，并结合渐进式图像编码器微调策略缓解过拟合问题。在推理阶段，将微调后的图像编码器直接迁移作为主干，将已解耦文本特征作为分类头权重，输入图像的特征与文本特征进行矩阵

乘法,输出预测结果。

2.1 文本特征解耦

针对零样本CLIP模型中螺栓缺陷类间相似度高的问题,VL-Bolt算法的第一阶段通过微调文本编码器,将缺陷对应的文本特征在特征空间中有效解耦,然后将已解耦文本特征作为后续图像文本特征对齐的类别锚点,从而避免使用大量样本和多样化文本描述来学习类间区分性。如图2第一阶段所示,通过固定模板"A photo of [cls]"为螺栓缺陷生成Text1~Text$N$文本描述,N为类别数。螺栓缺陷文本描述经过文本编码器后得到文本特征$\boldsymbol{T}_1$~$\boldsymbol{T}_N$。基于文本特征,构建文本特征解耦(text feature decoupling, TFD)损失函数,如式(1)。

$$L_{\mathrm{TFD}}=\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\langle \boldsymbol{T}_i,\boldsymbol{T}_j\rangle-\delta_{ij})^2 \tag{1}$$

式中,$\langle \boldsymbol{T}_i,\boldsymbol{T}_j\rangle$表示余弦相似度;$\delta_{ij}$的单位矩阵$\boldsymbol{I}_{N\times N}$的元素,如式(2)。

$$\delta_{ij}=\begin{cases}1 & i=j\\ 0 & i\neq j\end{cases} \tag{2}$$

利用L_{TFD}微调文本编码器使各类别文本特征间近似正交。

2.2 文本锚点引导的图像特征对齐

针对图像-文本对比学习中的类内语义重叠问题,文本锚点引导的图像特征对齐(text-anchor guided visual feature alignment, TA-VFA)将已解耦的文本特征作为锚点微调图像编码器,使图像特征对齐文本特征,实现图像特征的分离。该方法具有结构优势,有效避免类内语义重叠的影响。如图2中第二阶段所示,$\boldsymbol{T}_j(j=1,2,3,\cdots,N)$为已解耦的缺陷文本特征;训练集中的图像经过图像编码器生成特征向量$\boldsymbol{I}_i(i=1,2,3,\cdots,B)$,$B$为批尺寸。图像与文本特征之间的相似度计算如式(3)。

$$s_{ij}=\langle \boldsymbol{I}_i,\boldsymbol{T}_j\rangle \tag{3}$$

相似度构成矩阵$\boldsymbol{S}\in\mathbf{R}^{B\times N}$,如式(4)。

$$\boldsymbol{S}=[s_{ij}]_{B\times N} \tag{4}$$

采用交叉熵损失函数对图像编码器进行微调,以提高图像特征向量与对应类别文本特征向量之间的相似度,同时降低与其他类别文本的相似度,从而实现图像和文本特征有效对齐。交叉熵损失函数如式(5)。

$$L=\frac{1}{B}\sum_{i=1}^{B}\mathrm{CrossEntropy}(\boldsymbol{S}_{i,:},y_i)=-\frac{1}{B}\sum_{i=1}^{B}\log\left(\frac{\exp(s_{iy_i})}{\sum_{j=1}^{N}\exp(s_{ij})}\right) \tag{5}$$

式中,y_i为批次中第i张图片的标签。

该架构聚焦于图像特征与各类别文本锚点之间的相似度,而非同一批次内其他图像对应的文本,能够有效规避类内语义重叠问题,后续可视化实验验证了图像特征的分离效果。

2.3 渐进式图像编码器微调策略

螺栓缺陷数据集与通用领域的预训练数据集规模有较大差距,导致模型在微调过程中不可避免地存在过拟合问题。为抑制过拟合问题,本文从图像编码器的结构出发,提出渐进式图像编码器微调策略(progressive fine-tuning for the image encoder, PFT)。CLIP ViT-B/32模型的图像编码器结构如图3(a)所示。Raghu等(2021)通过可视化分析指出ViT模型低层主要关注图像的局部纹理与背景信息,高层侧重于全局结构和目标语义的提取。基于此,本文在微调初始阶段仅解冻Transformer高层,冻结低层,使模型优先学习与螺栓分类任务相关的高层语义特征,同时保留底层结构信息和预训练的泛化能力,实现稳定地适应下游任务

为确认螺栓缺陷分类的最优解冻层数,本文采用自适应渐进式解冻机制,逐层解冻Transformer层并观察验证集准确率,保留准确率最大轮次的权重。整体微调流程如图3(b)所示,实验结果表明,在冻结Transformer 0~4层,仅微调5~11层时,达到最高分类准确率。

2.4 模型推理

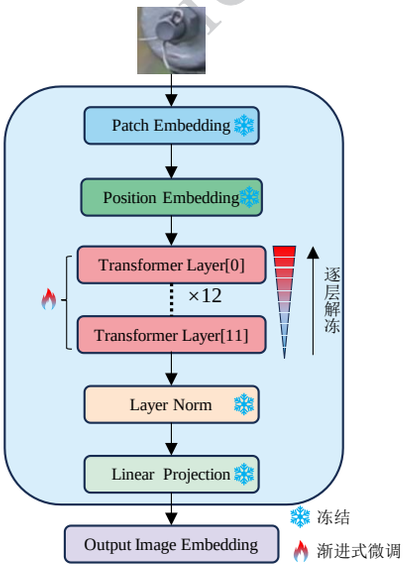
如图2右侧所示,在模型推理阶段,本文直接迁移第二阶段微调后的图像编码器权重以初始化主干网络。每一批次的输入图像经图像编码器后,输出图像特征矩阵$\boldsymbol{I}\in\mathbf{R}^{b\times d}$,其中$b$表示批尺寸,$d$为嵌入维度。然后,由于图像与文本特征已经对齐,将第一阶段通过特征解耦获得的文本特征$\boldsymbol{T}_1$~$\boldsymbol{T}_N$拼接为文本特征矩阵$\boldsymbol{T}\in\mathbf{R}^{N\times d}$作为分类头。分类预测分数$\hat{\boldsymbol{S}}\in\mathbf{R}^{b\times N}$计算公式如式(6)。

$$\hat{\boldsymbol{S}}=\boldsymbol{I}\cdot\boldsymbol{T}^{\top} \tag{6}$$

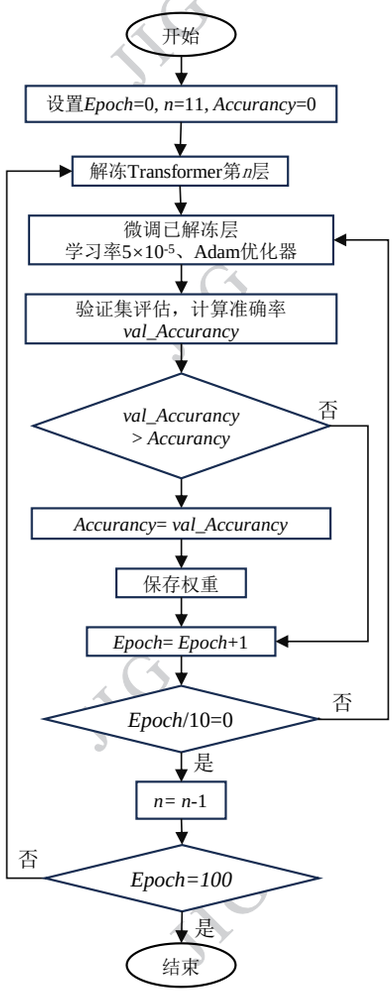
最后,选取每一行中相似度分数最高的类别索

引，作为图像的预测类别标签，如式(7)。

$$\hat{y}_i = \arg\max_j \hat{S}_{ij}, \quad i = 1,2,\cdots,b \tag{7}$$



(a) CLIP ViT-B/32图像编码器结构



(b) 渐进式微调流程

((a) the architecture of the CLIP ViT-B/32 image encoder; (b) progressive fine-tuning workflow)

图3 渐进式图像编码器微调策略示意图

Fig. 3 Progressive fine-tuning for the image encoder

3 实验与设置

3.1 实验环境和参数设置

本文实验基于PyTorch1.12.0框架，使用设备为单块NVIDIA GeForce RTX 3090。第一阶段微调文本编码器采用Adam优化器，学习率为10^{-5}，训练轮数为15轮。第二阶段微调图像编码器，输入图像的分辨率为224×224，采用2.3节渐进式微调策略，训练总轮次共为100轮，前10轮仅解冻Transformer第11层，而后每10轮逐步向低层解冻1层，采用Adam优化器，学习率为5×10^{-5}，批尺寸为128。

3.2 数据集

本文使用的数据为近年通过无人机搭载的摄像机对多条500kV输电线路巡检时拍摄得到的航拍图像。为突出螺栓主体供模型学习，本文使用labelimg软件对巡检图像进行标注，采用的数据格式为PASCAL VOC，标注依据为《架空输电线路设备缺陷影像标注规范(试行)》。依据标注框类别，从标注好的巡检图像中裁剪出螺栓及其缺陷图像，按ImageNet格式构建本文实验中所使用的螺栓缺陷数据集。数据集中包含来自不同连接处、不同拍摄角度的螺栓及其缺陷图像。本文将螺栓缺陷归结为4类，如图4所示，其中正常螺栓1180张、缺销螺栓2059张、螺母缺失螺栓526张、螺栓缺失975张。按照8:1:1比例划分训练集、验证

集和测试集。

3.3 对比实验

3.3.1 不同模型对比

本文的基线模型CLIP ViT-B/32是将CLIP的图像编码器后加入可训练的线性层作为分类头，并进行端到端微调。为验证VL-Bolt模型在螺栓缺陷分类任务中的有效性，本文选取了四种在ImageNet上预训练的视觉模型、基线模型、以及近年来提出的三种视觉螺栓缺陷分类模型作为对比对象。所有模型均在螺栓缺陷数据集上进行微调，结果如表1所示。从实验结果可以看出，VL-Bolt模型的分类准确率可达92.2%。与单模态视觉模型相比，视觉-语言模型在各项指标上均表现更优。相较基线模型CLIP

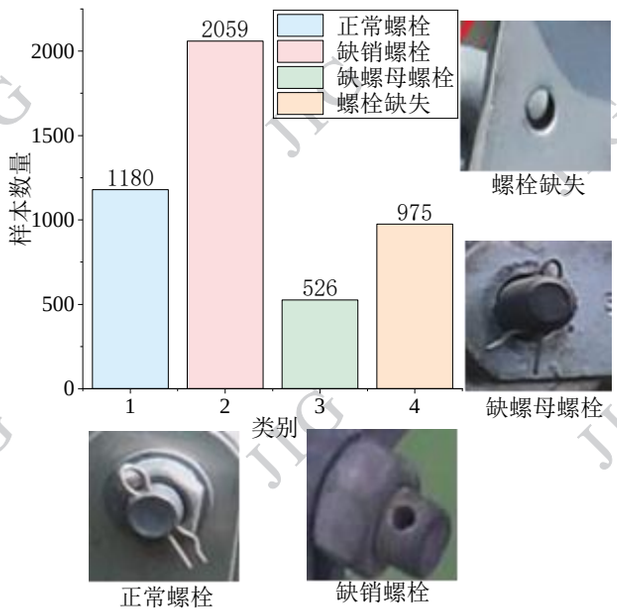


图4　螺栓缺陷数据集样本数量和示例

Fig. 4　Bolt defect dataset: sample counts and examples

ViT-B/32，VL-Bolt准确率、召回率、精准率和F1值上分别提升3.9%、3.4%、3.2%和3.8%。进一步对比以ViT为主干结构的ViT B/32、CLIP ViT-B/32和VL-Bolt三种模型，在模型参数量接近的情况下，VL-Bolt的性能优于另外两者。对于现有螺栓缺陷分类模型：改进WRN（Liu等，2022）、改进DenseNet（李学渊等，2022）和ATT-RN50（Lin等，2021），其指标虽然优于大多数视觉模型，但仍低于VL-Bolt，证明了本文多模态方法的优势。

最新的螺栓缺陷分类模型为多模态MUCO-BD模型（张珂等，2025），该模型采取“预训练+微调”的路线。MUCO-BD通过扩充输电线路金具图像样本，并设计多样化文本描述模板，构建了包含22295组图像-文本对的输电线路巡检图像-文本数据集（transmission line inspection image-text dataset，TLID），并在该数据集上进行图像-文本对比学习预训练。随后，MUCO-BD在螺栓缺陷数据集上进一步端到端微调，使模型能够学习到丰富的电力领域知识。而VL-Bolt仅在零样本CLIP上进行微调，准确率等指标较MUCO-BD有差距，但其训练参数量减少了约29.9%。为了进一步验证两者在性能上的差异，本文将VL-Bolt在同等配置下与MUCO-BD进行对比，结果如表2所示。可以观察到，在TLID上进行预训练并在BDCD上微调的设置下，VL-Bolt的准确率优于MUCO-BD。然而，当模型仅在BDCD上进行预训练与微调时，VL-Bolt的性能显著下降，MUCO- BD的性能接近端到端微调。该结果说明，仅利用BDCD数据集进行预训练反而会破坏原始预训练表征。由于VL-Bolt仅渐进式微调Transformer层并冻结其他

表1　不同模型的螺栓缺陷分类结果

Table 1　Results of different bolt defect classification models

模型	模态	参数量/10^6	训练参数量/10^6	召回率/%	精准率/%	F1值/%	准确率/%
ViT-B/32	图像	85.8	85.8	87.2	88.7	87.9	89.5
RN50	图像	23.5	23.5	88.3	89.9	89.1	90.4
EfficientNet	图像	10.7	10.7	86.1	87.4	86.7	87.8
MobileNet v3	图像	4.2	4.2	85.5	86.9	86.2	87.4
CLIP ViT-B/32	图像	86.2	86.2	85.9	88.0	86.8	88.3
改进DenseNet	图像	12.5	12.5	88.5	90.2	89.3	90.8
改进WRN	图像	11.5	11.5	88.0	88.9	88.4	89.8
ATT-RN50	图像	24.0	24.0	88.1	89.7	88.8	90.1
MUCO-BD-ViT	图像-文本	86.2	151.3	94.2	95.5	94.8	96.2
本文VL-Bolt	图像-文本	86.2	108.7	89.6	91.2	90.6	92.2

层，其性能对预训练表征依赖性更强，因此当原始预训练表征被破坏时，模型准确率下降更为明显。

表2　VL-Bolt与MUCO-BD的对比结果

Table 2　Comparison results between VL-Bolt and /MUCO-BD

方法	预训练	微调	准确率/%
MUCO-BD-ViT	BDCD	BDCD	88.5
	TLID	BDCD	96.2
VL-Bolt	BDCD	BDCD	80.5
	TLID	BDCD	96.8

3.3.2　CLIP模型不同迁移方法对比

本研究将本文的迁移方法与近年提出的4种迁移方法进行对比，结果如表3所示。根据实验数据可知，本文方法的分类准确率优于其他迁移方法。从算法原理和结构角度进行分析，CLIP-Adapter（Gao等，2024）、CoOp（Zhou等，2022）、CoCoOp（Zhou等，2022）和Tip-Adapter（Zhan等，2022）模型分类效果不佳的主要原因为螺栓缺陷类间相似度过高：CLIP-adapter为主干添加的参数量不足以充分学习到类间差异；CoCoOp的动态提示依赖图像特征，螺栓图像类间相似度过高导致分类效果不佳；而CoOp将文本模板设为可学习参数，不依赖图像特征，相比于CoCoOp效果更优，但因未改进模型的图像侧，分类表现较CLIP-Adapter稍差；Tip-Adapter将每类训练图像的特征存储为缓存，在推理阶段通过计算输入图像与缓存的相似度来辅助分类，Tip-Adapter-F将缓存进行微调，效果相比于Tip-Adapter有所提升，但是Tip-Adapter模型未能直接降低类间相似度，因此不适配螺栓缺陷分类问题。WiSE-FT（Wortsman等，2022）通过将模型端到端微调的权重与零样本模型权重进行线性插值，有效提升了模型的泛化性，与另外3种迁移方法相比准确率最高，较基线模型提高1.7%，但本文方法仍领先2.2%，证明了本文迁移方法的泛化性优势。

3.4　公开数据集实验

为验证所提方法的鲁棒性，本研究将VL-Bolt模型应用于三个公开数据集，并与基线模型对比分类性能，结果如表4所示。在细粒度分类数据集Flowers102和StanfordCars上，VL-Bolt均取得显著提升，尤其在StanfordCars数据集上，相较端到端微调模型提升了10.7%。这一性能提升主要得益于渐进式图像编码器微调策略，使模型能够有效地聚焦于目标主体特征；同时，通用领域数据集样本规模较大，也有助于充分发挥TA-VFA结构优势。此外，在类间差异显著的Caltech101数据集上，VL-Bolt依然表现出较强的任务适配性，证明了该方法在多类别场景下的可拓展性与鲁棒性。

表3　不同迁移方法的螺栓缺陷分类结果

Table 3　The bolt defects classification results of different transfer methods

迁移方法	准确率/%
CLIP-Adapter	81.9
CoOp	79.4
CoCoOp	56.3
Tip-Adapter	41.2
Tip-Adapter-F	79.1
WiSE-FT α=0.5	89.4
WiSE-FT最优α	90.0
本文VL-Bolt	92.2

表4　公开数据集分类准确率

Table 4　Classification accuracy on public datasets

数据集	类别/类	最佳解冻层	准确率/%	
			端到端微调	VL-Bolt
Flowers102	102	6~11	97.8	99.3
StanfordCars	196	4~11	77.2	87.9
Caltech101	102	6~11	91.8	96.7

3.5　消融实验

1）为验证VL-Bolt模型各组成模块对螺栓缺陷分类性能的贡献，本文设计了一系列消融实验，结果如表5所示。由于螺栓缺陷数据集规模较小且类别间差异细微，受类内语义重叠问题影响，使用图像-文本对比学习微调CLIP模型难以学习到有效的类间差异；TA-VFA架构能够避免类内语义重叠问题，引入后准确率显著提升；文本特征解耦机制能够有效避免部分类间混淆；渐进式图像编码器微调策略引导模型聚焦于目标语义，提升准确率最为明显。此外，本文的改进方法未给模型带来额外的参数量和计算量。

2）为验证文本特征分类头为最佳分类头，本实验将第二阶段微调后的图像编码器接入一层可训练的线性层作为分类头，并冻结主干仅训练该层。实

表5　消融实验结果

Table 5　Results of ablation experiments

TA-VFA	TFD	PFT	参数量/10^6	计算量/10^9	准确率/%
√	√	√	86.2	17.6	92.2
√	–	√	86.2	17.6	91.7
√	√	–	86.2	17.6	89.6
√	–	–	86.2	17.6	88.9
–	–	–	86.2	17.6	22.3

验结果如表6所示，对比可知，文本分类头达到了与线性层分类头相同的准确率，均为92.2%。这表明图像特征与对应的文本特征高度对齐无需进一步微调，选用文本特征作为分类头兼具性能与效率。

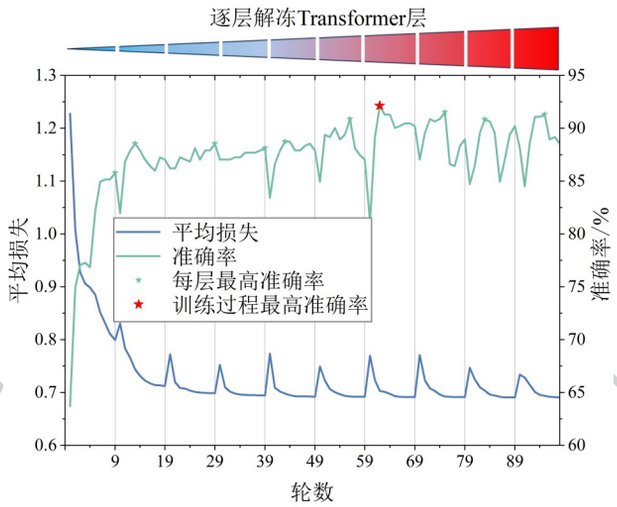
表6　不同分类头的螺栓缺陷分类准确率

Table 6　The bolt defect classification accuracy of different classification heads

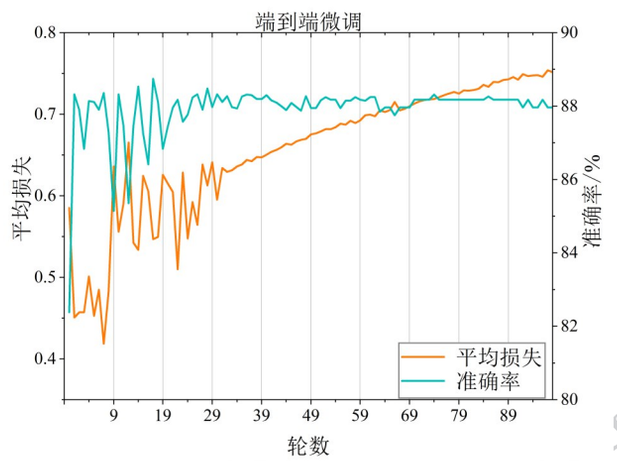
主干	分类头	准确率/%
微调后的图像编码器	已解耦文本特征分类头	92.2
	可训练的线性层	92.2

3)为确认PFT过程的最佳解冻层数，本文对微调图像编码器过程的损失和验证集准确率数据绘制曲线，如图5(a)所示。损失曲线(蓝色)可以观察到，因渐进式图像编码器微调策略每10轮解冻1层Transformer层，每层刚解冻后损失激增，曲线出现“尖刺”，但在10轮的微调周期内损失稳定收敛，符合预期表现。观察准确率曲线(绿色)，在解冻Transformer11层时，准确率快速上升。随着更多层的逐步解冻，准确率曲线在每10轮周期内出现“先升后降”的波动趋势。分析波动原因，“先升”是由于刚解冻新层时模型尚未收敛，导致性能下降，随着训练进行，准确率随之回升；“后降”则说明在每10轮的微调后期，虽然损失值保持下降，但模型出现过拟合问题，导致泛化能力下降。综合观察可知，在仅解冻Transformer5~11层的阶段，模型达到了最高验证集准确率，继续解冻低层时，准确率出现了不同程度的降低。

为验证PFT策略在抑制过拟合方面的有效性，本文绘制了端到端微调的损失和准确率曲线与PFT进行对比，如图5(b)所示。观察损失曲线趋势可以



(a)　渐进式微调



(b)　端到端微调

((a) progressive fine-tuning; (b)end-to-end fine-tuning)

图5　损失和准确率曲线

Fig. 5　Loss and accuracy curves

发现，随着训练轮数增加，端到端微调的损失出现震荡并逐渐升高，而PFT的损失趋于收敛，说明PFT策略能够有效稳定训练过程。

3.6　可视化实验

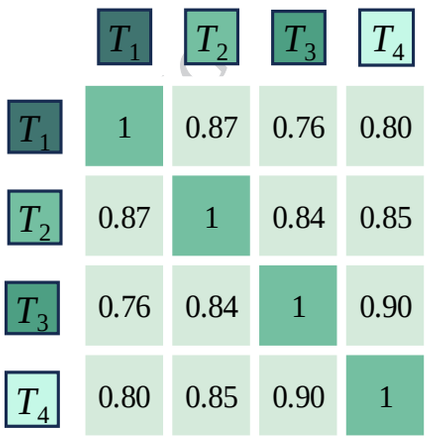
3.6.1　文本解耦效果

本文采用文本特征解耦损失函数微调文本编码器，目的是降低不同缺陷文本之间的相似度。如图6(a)所示，解耦前缺陷文本间的相似度极高，无法用作分类，体现了CLIP模型在细粒度分类任务上的局限性。如图6(b)所示，微调后输出的4类文本特征之间接近正交，实现有效解耦。

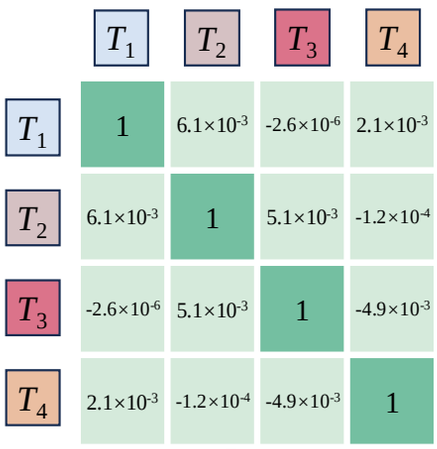
3.6.2　图像特征空间可视化

本小节利用t-SNE算法(Van等，2008)对图像-

文本对比学习、CLIP ViT-B/32 和 VL-Bolt 的螺栓缺陷图像特征空间进行可视化，分析 VL-Bolt 在螺栓缺陷分类上的适配性。如图 7 所示，图像-文本对比学习未能学习到有效的类间区分，导致特征空间类别间边界模糊，可视化结果进一步验证了图像-文本对比学习在螺栓缺陷数据集上的不适配现象；端到端微调的 CLIP ViT-B/32 模型在特征空间中呈现出类别内聚集高度紧凑、类别间边界异常清晰的特征分布，然而过于锐利的决策边界导致了泛化能力不足；VL-Bolt 以文本特征作为锚点，利用渐进式图像编码器微调策略将图像特征与文本特征对齐，在保持类别间分离性的同时，边界分布更加自然、不过度紧凑，证明其在特征表示上取得了更合理的平衡。



（a） 解耦前

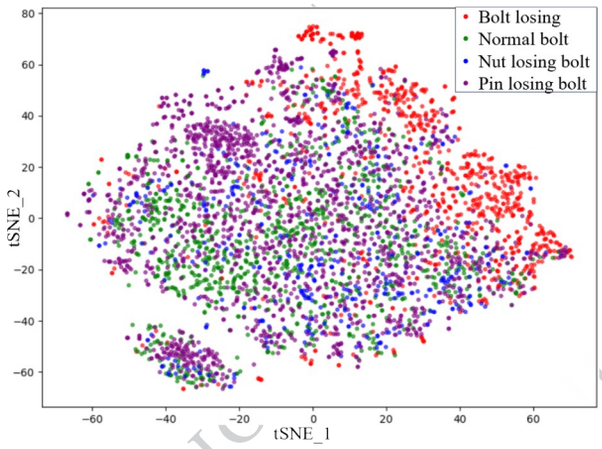


（b） 解耦后

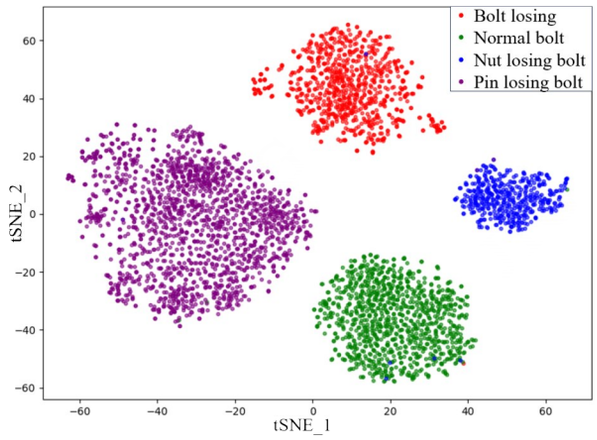
((a) before; (b) after)

图 6 文本特征解耦前后相似度

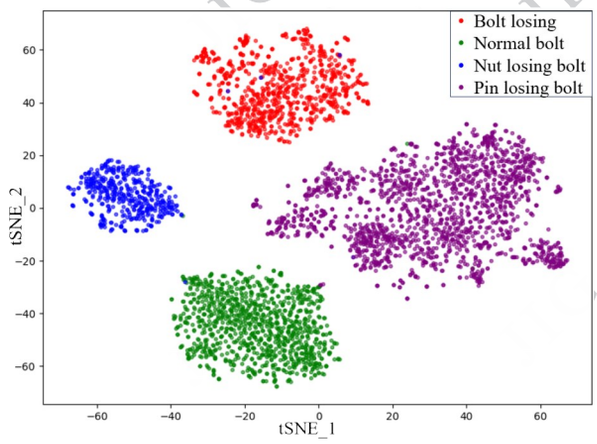
Fig. 6 Text feature similarity before and after decoupling



（a） 图像-文本对比学习



（b） CLIP ViT-B/32



（c） VL-Bolt

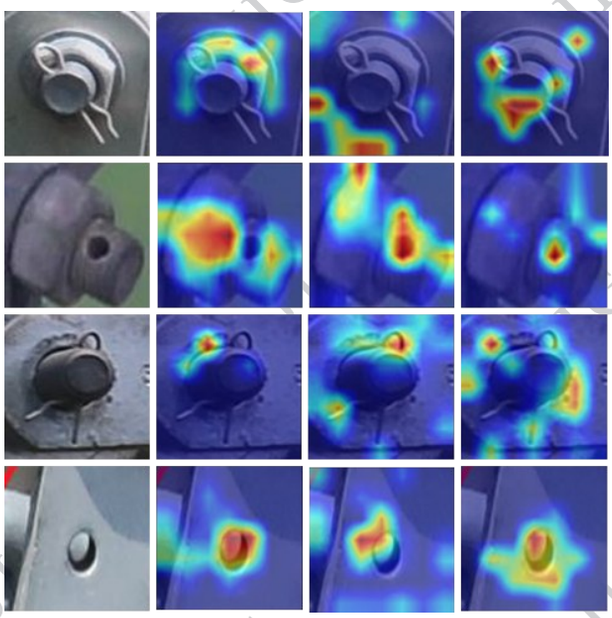
((a) image-text contrastive learning; (b) CLIP ViT-B/32; (c) VL-Bolt)

图 7 螺栓缺陷在图像特征空间中的 t-SNE 可视化

Fig. 7 The t-SNE visualization of bolt defects in the image feature space

3.6.3　注意力图可视化

为了能直观地观察VL-Bolt模型在螺栓缺陷分类任务上的有效性，本文利用Grad-CAM技术（Selvaraju 等，2017）对 ViT-B/32、CLIP ViT-B/32 和 VL-Bolt三种模型在螺栓缺陷分类过程中的判别性区域进行可视化分析。如图8所示，本文选取4张螺栓图片，可视化不同模型在螺栓缺陷分类时的判别性区域。暖色区域代表模型更加关注的区域，而冷色区域则表示模型对该区域的关注较少。对比发现，VL-Bolt模型在各类别上充分关注到了缺陷特征且聚焦于缺陷本身，如缺销螺栓重点关注到了销孔所在位置，缺螺母螺栓关注了螺母应在区域；而另外两种模型都出现了不同程度的关注背景和非缺陷区域，如ViT模型在缺销螺栓上未能准确关注销孔，CLIP-ViT模型关注到了正常螺栓的背景而非螺栓本体。注意力图可视化实验验证了本文方法使模型学习到了不同螺栓缺陷的特征。



(a)原始图片 (b)ViT (c)CLIP-ViT (d)VL-Bolt
((a) original image; (b) ViT; (c) CLIP-ViT; (d)VL-Bolt)

图8　利用Grad-CAM技术可视化分析

Fig. 8　Visualization analysis with Grad-CAM

4　结　论

本文阐述了螺栓缺陷数据集进行图像-文本对比学习时存在的类内语义重叠问题；为提升视觉-语言模型在螺栓缺陷分类上的适配度，提出一种融合图文特征对齐和文本解耦的输电线路螺栓缺陷分类方法。该方法通过文本特征解耦和文本锚点引导的图像特征对齐架构，提高了视觉-语言对比预训练模型（CLIP）的螺栓缺陷分类准确率；同时，本文设计的渐进式图像编码器微调策略，在保留底层结构信息基础上，有效抑制了过拟合问题。实验结果表明，本文方法实现了不同缺陷类别螺栓的图像和文本特征在特征空间中有效分离，分类准确率可达92.2%，显著优于基线模型，为视觉-语言模型在电力巡检领域的应用提供了新思路。

在缺陷类别文本描述方面，本文采取的单一文本描述有效降低了数据准备阶段的工作量，但也面临对螺栓细节特征学习不充分的问题。后续工作考虑将不同缺陷特有的特征文本进行解耦，进一步提高模型的泛化性。此外，本文方法仍然面临因数据集规模而无法充分发挥该架构潜力的问题。

参考文献（References）

Gu C Y, Li Z, Shi J T, Zhao H H, Jiang Y and Jiang X C. 2020. Detection for pin defects of overhead lines by UAV patrol image based on improved Faster-RCNN. High Voltage Engineering, 46(9): 3089-3096 (顾超越, 李喆, 史晋涛, 赵航航, 江一, 江秀臣. 2020. 基于改进Faster-RCNN的无人机巡检架空线路销钉缺陷检测. 高电压技术, 46(9): 3089-3096) [DOI: 10.13336/j.1003-6520.hve.2019074]

Gao P, Geng S, Zhang R, Ma T, Fang R, Zhang Y, Li H and Qiao Y. 2024. CLIP-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision, 132(2): 581 - 595 [DOI: 10.1007/s11263-023-01891-x]

Jiang R X, Liu L B and Chen C W. 2023. Clip-count: Towards text-guided zero-shot object counting// Proceedings of the 31st sACM International Conference on Multimedia. Ottawa, Canada: ACM: 4535 - 4545 [DOI: 10.1145/3581783.3611789]

Khan Z and Tapaswi M. 2024. FigCLIP: Fine-grained clip adaptation via densely annotated videos [EB/OL]. [2024-06-16]. https://arxiv.org/abs/2401.07669

Kim W, Son B and Kim I. 2021. ViLT: Vision-and-language transformer without convolution or region supervision// Proceedings of the 38th International Conference on Machine Learning. Vienna, Austria: PMLR: 5583 - 5594.

Lezama J, Qiu Q, Musé P and Sapiro G. 2018. Ole: Orthogonal low-rank embedding—a plug and play geometric loss for deep learning// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 8109 - 8118. [DOI: 10.1109/CVPR.2018.00847]

Li J, Jie Z, Ricci E, Ma L and Sebe N. 2024. Enhancing Robustness of Vision-Language Models through Orthogonality Learning and Self-Regularization[EB/OL].[2024-06-16]. https://arxiv.org/abs/2407.08374

Li X Y. 2022. Research on bolt state classification method of overhead power line based on transfer learning. Electric Engineering,(20): 96 - 99 (李学渊. 2022. 基于迁移学习的架空电力线路螺栓状态分类方法研究. 电工技术,(20):96 - 99)[DOI:10.19768/j.cnki.dgjs.2022.20.030]

Liao R J, Wang Y Y, Liu H, Liu H B and Ma Z P. 2018. Research status of condition assessment method for power transmission and transformation equipment. High Voltage Engineering,44(11):3454 - 3464 (廖瑞金, 王有元, 刘航, 刘宏波, 马志鹏. 2018. 输变电设备状态评估方法的研究现状. 高电压技术, 44(11):3454 - 3464)[DOI:10.13336/j.1003-6520.hve.20181031002]

Lin Z J, Liang Y J and Jiang Q N. 2021. A bolt defect recognition algorithm based on attention model// TALLÓN-BALLESTEROS A J, ed. Fuzzy Systems and Data Mining VII. [S. l.]: IOS Press: 86 - 93 [DOI:10.3233/FAIA210179]

Liu L S, Zhao J L, Chen Z and Zhao B, Ji Y. 2022. A new bolt defect identification method incorporating attention mechanism and wide residual networks. Sensors, 22 (19): 7416 [DOI: 10.3390/s22197416]

Momeni L, Caron M, Nagrani A, Zisserman A and Schmid C. 2023. Verbs in action: Improving verb understanding in video-language models// Proceedings of the IEEE/CVF International Conference on Computer Vision. [S. l.]: IEEE: 15579 - 15591[DOI: 10.1109/ICCV51070.2023.01428]

Qi Y C, Geng S F, Zhao Z B, Lv X C and Sun M. 2023. A method for super resolution processing of bolt image based on feature transfer. CAAI Transactions on Intelligent Systems, 18(4):858 - 866 (戚银城, 耿劭锋, 赵振兵, 吕雪纯,孙梦. 2023. 基于特征迁移的螺栓图像超分辨率处理方法. 智能系统学报, 18(4):858 - 866)[DOI:10.11992/tis.202201009]

Radford A, Kim J W, Hallacy C, Ramesh A and Goh G, Agarwal S, Sutskever I. 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]

Raghu M, Unterthiner T, Kornblith S and Zhang C. 2021. Do vision transformers see like convolutional neural networks? [EB/OL]. [2024-06-16]. https://arxiv.org/abs/2108.08810

Ramachandram D and Taylor G W. 2017. Deep multimodal learning: a survey on recent advances and trends. IEEE Signal Processing Magazine, 34(6):96 - 108[DOI:10.1109/MSP.2017.2738401]

Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D and Batra D. 2017. Grad-CAM: visual explanations from deep networks via gradient-based localization// Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE:618 - 626 [DOI: 10.1109/ICCV.2017.74]

Maaten L V D and Hinton G. 2008. Visualizing data using t-SNE. Journal of Machine Learning Research, 9(11):2579 - 2605[DOI: -]

Yang J, Li C, Zhang P, Xiao B, Liu C, Yuan, L and Gao J. 2022. Unified contrastive learning in image-text-label space// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, LA, USA: IEEE: 19163 - 19173[DOI:10.1109/CVPR52688.2022.01857]

Wortsman M, Ilharco G, Kim J W, Li M, Kornblith S, Roelofs R and Schmidt L. 2022. Robust fine-tuning of zero-shot models // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, LA, USA: IEEE:7959 - 7971[DOI: 10.1109/CVPR52688.2022.00780]

Zheng X, Ji R, Sun X, Zhang B, Wu Y and Huang F. 2019. Towards optimal fine-grained retrieval via decorrelated centralized loss with normalize-scale layer[J]. IEEE Transactions on Image Processing, 28(10): 5022 - 5035. [DOI:10.1109/TIP.2019.2911704]

Zhang K, He Y X, Zhao K, Feng X H, Zhao Z B and Ma Z Y. 2021. Multi-label classification method of bolt attributes based on deformable NTS-Net. Journal of Image and Graphics, 26(11): 2582 - 2593 (张珂, 何颖宣, 赵凯, 冯晓晗, 赵振兵, 马占宇. 2021. 可变形 NTS-Net 的螺栓属性多标签分类. 中国图象图形学报, 26(11):2582 - 2593)[DOI:10.11834/jig.200703]

Zhang K, Zheng Z Y, Shi C J, Zhao Z B and Xiao Y J. 2025. Transmission line bolt defects classification based on multi-modal contrastive learning. High Voltage Engineering, 51(2): 630 - 641 (张珂, 郑朝烨, 石超君, 赵振兵,肖扬杰. 2025. 基于多模态对比学习的输电线路螺栓缺陷分类. 高电压技术, 51(2):630 - 641)[DOI:10.13336/j.1003-6520.hve.20232123]

Zhang S, Wang H T, Dong X C, Li Y R, Li Y, Wang X Y and Sun Y Y. 2021. Bolt detection technology of transmission lines based on deep learning. Power System Technology, 45(7):2821 - 2828 (张姝, 王昊天, 董骁翀, 李玉荣, 李烨, 王新迎, 孙英云. 2021. 基于深度学习的输电线路螺栓检测技术. 电网技术, 45(7):2821 - 2828)[DOI:10.13335/j.1000-3673.pst.2020.1336]

Zhao Z B, Duan J K, Kong Y H and Zhang D X. 2021. Construction and application of bolt and nut pair knowledge graph based on GGNN. Power System Technology, 45(1): 98 - 106 (赵振兵, 段记坤, 孔英会, 张东霞. 2021. 基于门控图神经网络的栓母对知识图谱构建与应用. 电网技术, 45(1):98 - 106)[DOI:10.13335/j.1000-3673.pst.2020.0063]

Zhao Z B, Jiang Z G, Li Y X, Qi Y C, Zhai Y J, Zhao W Q and Zhang K. 2021. Overview of visual defect detection of transmission line components. Journal of Image and Graphics, 26(11):2545 - 2560 (赵振兵, 蒋志钢, 李延旭, 戚银城, 翟永杰, 赵文清, 张珂. 2021. 输电线路部件视觉缺陷检测综述. 中国图象图形学报, 26(11):2545 - 2560)[DOI:10.11834/jig.200689]

Zhao Z B, Zhang W, Zhai Y J, Zhai Y J, Zhao W Q, Zhang K, Kong Y

H and Qi Y C. 2020. Concept, research status and prospect of electric power vision technology. Electric Power Science and Engineering, 36(1):1 - 8(赵振兵, 张薇, 翟永杰, 赵文清, 张珂, 孔英会, 戚银城. 2020. 电力视觉技术的概念▮研究现状与展望. 电力科学与工程, 36(1):1 - 8)[DOI:10.3969/j.ISSN.1672-0792.2020.01.001]

Zhang K, Xiao Y J, Wang J, Du M, Guo X, Zhou R and Zhao Z. 2024. DP-GAN: A transmission line bolt defects generation network based on dual discriminator architecture and pseudo-enhancement strategy. IEEE Transactions on Power Delivery, 39(3):1622-1633.[DOI:10.1109/TPWRD.2024.3373130]

Zhang J, Huang J, Jin S and Lu S. 2024. Vision-language models for vision tasks: A survey. IEEE transactions on pattern analysis and machine intelligence, 46(8):5625-5644.[DOI:10.1109/TPAMI.2024.3369699]

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]

Zhou K, Yang J, Loy C C and Liu Z. 2022. Conditional prompt learning for vision-language models // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, LA, USA: IEEE:16816 - 16825[DOI:10.1109/CVPR52688.2022.01631]

Zhang R, Zhang W, Fang R, Gao P, Li K, Dai J and Li H. 2022. Tip-adapter: Training-free adaption of CLIP for few-shot classification // European conference on computer vision. Cham: Springer Nature Switzerlan:493 - 510[DOI:10.1007/978-3-031-19833-5_29]

作者简介

张珂,1980年生,男,教授,通信作者,主要研究方向为深度学习、计算机视觉、电力视觉、生物特征识别等。E-mail: zhangkeit@ncepu.edu.cn

高明伟,男,硕士研究生,主要研究方向为电力视觉和多模态学习。 E-mail: 220242215118@ncepu.edu.cn

郑朝烨,男,博士研究生,主要研究方向为电力视觉和多模态学习。E-mail: zyezheng@163.com

李硕士,男,讲师,主要研究方向为深度学习、图像增强和电力视觉。 E-mail: lishuoshi@ncepu.edu.cn

聂鼎,男,高级工程师,主要研究方向为智能配电网、输电设备缺陷图像分析。 E-mail: 183013028@qq.com

周帅,男,工程师,主要研究方向为电力设备缺陷智能检测。E-mail: zhoushuailijinmei@163.com