DOI: 10.19595/j.cnki.1000-6753.tces.241036

# 基于扩散模型检测的高铁接触网绝缘子缺陷语义描述方法

陈 永[1,2] 安卓奥博[1] 周建宇[1]

（1. 兰州交通大学电子信息工程学院 兰州 730070
2. 兰州交通大学人工智能与图形图像处理工程研究中心 兰州 730070）

**摘要** 高铁接触网绝缘子作为高速铁路牵引供电的重要装置，可为接触网提供电气部件绝缘和腕臂结构支撑，其安全性对于高速铁路行车至关重要。针对绝缘子检测时易受复杂环境背景干扰，导致缺陷检测精度低以及无法提供缺陷语义描述的问题，该文提出一种基于扩散模型检测的绝缘子缺陷描述方法。首先，构建大核空间选择特征提取网络，加强绝缘子缺陷特征信息的提取能力；其次，基于扩散模型设计融合扩散机制的检测解码器，并对解码器生成的噪声框进行逆向贝叶斯扩散，还原绝缘子真值框的预测，提高模型的抗背景干扰能力；最后，设计交叉注意力机制的编码器和解码器，实现图像与文本的跨模态映射，并通过文本过滤机制驱动的多模态语言视觉预训练（BLIP）模型，完成绝缘子缺陷文本描述输出。实验结果表明，所提绝缘子缺陷检测模型的平均准确度达到 93.04%，相较于 DTER 和 Faster RCNN 的 $mAP_{0.5}$ 分别提升 4.63%和 5.78%，且 F1-score 高达 82.91%，平均双语评估替换评价指标（BLEU）和基于精确率的图像描述评价指标（CIDEr）分别达到 83.51%和 1.94。与其他方法相比，具有更高的检测精度和缺陷语义描述准确性，能够满足对高速铁路绝缘子缺陷的检测需求。

**关键词：**高铁接触网 绝缘子缺陷检测 缺陷语义描述 扩散模型 交叉注意力机制

**中图分类号：**TM755; TP389

## 0 引言

高速铁路是一种以电力作为牵引动力的铁路类型，而高铁接触网的主要用途是为高速列车提供牵引动力。其中，绝缘子在接触网中提供电气绝缘和机械支撑等作用。由于绝缘子长期暴露在室外，会受到外部天气和环境等自然因素的影响，易遭受极端天气、强电场、机械振动等不利因素的侵袭，使绝缘子表面出现破损、脏污、闪络等缺陷[1]。因此，高铁绝缘子缺陷会给高速铁路运行带来严重的安全隐患，危及行车安全。

随着计算机图像处理技术的发展，如何高效准确地检测绝缘子缺陷是目前研究的热点。传统基于计算机视觉的绝缘子缺陷检测主要以绝缘子轮廓、颜色等信息作为检测特征[2]。文献[3]设计绝缘子伞裙轮廓特征和灰度相似度匹配的融合算法，实现利用伞裙间距对缺陷绝缘子进行分类，但该算法容易受到感光敏感性的影响，导致绝缘子轮廓特征提取不准确。文献[4]改进遗传算法用于绝缘子表面破损检测，但该算法通过统计缺陷像素的阈值范围，只能完成绝缘子自爆和强破损的检测，导致此方法检测性能受限。

由于传统绝缘子检测方法在缺陷检测时局限性越发明显，而深度学习方法通过对大量绝缘子数据进行特征学习，可以有效地提高检测的准确性和效率，是目前的主流研究方向。文献[5]基于反卷积和多尺度特征融合构成的 MSD-Net（multi-scale discriminative network）解决了绝缘子缺陷区域像素信息少、形状尺寸不一造成的识别效果不佳的问题，但该方法中对于检测框过滤采用滑动窗口的方法，导致窗口边缘的特征信息被破坏。文献[6]通过掩膜区域卷积神经网络（Mask Region Convolutional Neural Network, Mask-RCNN）实现绝缘子检测后，

国家自然科学基金（62462043, 61963023）和兰州交通大学重点研发项目（ZDYF2304）资助。
收稿日期 2024-06-17 改稿日期 2024-08-10

又通过聚类算法分析绝缘子缺陷并进行判断，但聚类算法对初始中心点选择敏感，导致绝缘子缺陷判断精度较低。文献[7]采用基于 DETR（Detection Transformer）的解码器和编码器完成绝缘子的缺陷检测，但该方法对目标尺寸变化和遮挡较为敏感，导致检测准确度下降。文献[8]将 Transformer 特征提取网络和 Faster R-CNN 相结合，通过广域感受野完成复杂背景下绝缘子的检测，但 Transformer 不能完全适配传统卷积框架，导致有效特征不能被充分利用，因此检测精度下降。

图像描述是一种结合计算机视觉和自然语言处理的跨模态技术，目标是通过文本描述图像内容[9]。目前，图像描述方法已逐步应用于农业[10]、医学[11]等领域，能够为视觉检测结果生成相应的文字描述。例如，文献[10]设计了一种基于长短期记忆网络（Long Short-Term Memory, LSTM）的语义描述方法，通过 LSTM 模型对农业病害进行了描述；文献[11]通过大型语言和视觉助手（Large Language and Vision Assistant, LLaVA）语义描述模型对医学病情进行了描述。通过图像描述方法可以将视觉信息转换为文本信息，进一步提高视觉图像的信息量，以便更好地理解图像的语义信息[12-13]。对于高速铁路接触网绝缘子缺陷检测，相比于单模态图像检测方法，通过图像描述的方式，不仅可以提供丰富的绝缘子缺陷信息，如缺陷类型、位置等内容，而且可以自动生成缺陷检测文本日志，辅助技术人员提高检修效率。由于目前国内外高速铁路绝缘子缺陷描述相关研究较少，因此将图像描述方法应用于高铁接触网绝缘子缺陷检测领域中，对进一步提升接触网检测作业智能化具有重要意义。

综上所述，针对绝缘子检测时，易受环境复杂背景干扰，导致缺陷检测精度低及无法提供缺陷语义描述的问题，本文基于扩散检测模型，提出了一种基于扩散模型检测的高铁接触网绝缘子缺陷语义描述方法。首先，构建大核空间选择（Large Selective Kernel, LSK）特征提取网络，加强绝缘子的特征信息提取能力；其次，设计融合扩散过程的检测解码器，该解码器能够对预测框完成逆向贝叶斯扩散处理，还原绝缘子真实边界框预测，增强模型在复杂背景中的抗干扰能力；最后，由基于文本过滤机制的多模态视觉语言预训练（Bootstrapping Language-Image Pre-training, BLIP）模型完成绝缘子缺陷检测的描述输出。实验结果表明，所提方法对绝缘子缺陷具有更高的检测精度和缺陷语义描述准确性。

## 1　本文方法

### 1.1　网络整体结构

DiffusionDet 网络模型是基于贝叶斯推断定理提出的新型目标检测模型[14-15]，其将目标检测定义为一个去噪扩散过程，结合图像编码器和检测解码器，学习真值框到噪声框的去噪扩散过程，从而使解码器推理得到的随机生成框逐步细化为准确的绝缘子检测结果。本文在 DiffusionDet 的基础上提出一种绝缘子缺陷描述方法，其网络整体结构如图 1 所示。首先，输入绝缘子图像到图像编码器中，由大核空间选择机制的特征提取网络完成绝缘子缺陷特征提取；其次，通过逆向递推检测头完成对绝缘子的检测；然后，对检测得到的绝缘子图像进行图像编码（Vision Transformer, ViT）[16]，同时对绝缘子文本描述进行文本编码（Bidirectional Encoder Representations from Transformers, BERT）[17]，通过图文对比机制和图文匹配机制捕捉图像文本的映射

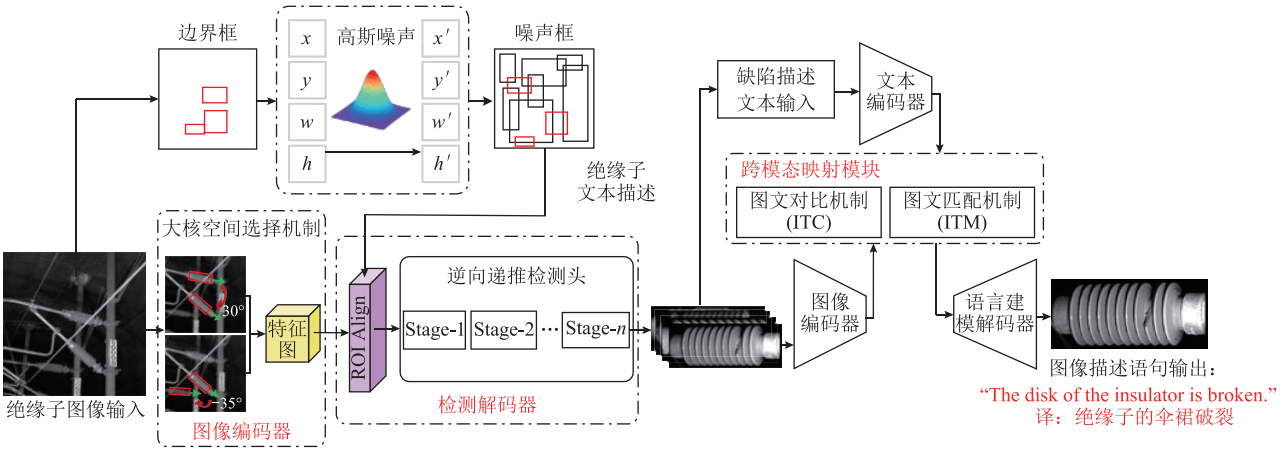



图 1　网络整体结构

Fig.1　Overall network architecture

关系；最后，由语言建模解码器输出生成的文本描述语句和绝缘子图像。

### 1.2 绝缘子目标检测阶段

在生成绝缘子缺陷描述之前，需要首先对绝缘子所在区域进行检测，获得每个绝缘子目标所在区域。该阶段主要由图像编码器和检测解码器构成。

#### 1.2.1 图像编码器

绝缘子在检测过程中，由于接触网拍摄时的前后遮挡，存在对不同绝缘子信息特征提取不准确的问题。针对此问题，采用对空间信息敏感的大核空间选择特征提取网络作为编码器，其结构如图 2 所示。由于传统卷积层的卷积核无法动态调整感受野的范围，导致普通卷积核在绝缘子特征提取时的感受野范围受限，无法充分提取其特征。因此，本文通过向层次块（Stage）中集成大核空间选择机制[18]，在空间维度上自适应地聚合空间信息，对不同绝缘子分配不同程度的感受野，动态调整特征提取网络得到的绝缘子缺陷特征，提升绝缘子缺陷特征提取性能。其中，基于大核空间选择机制的网络结构如图 3 所示。

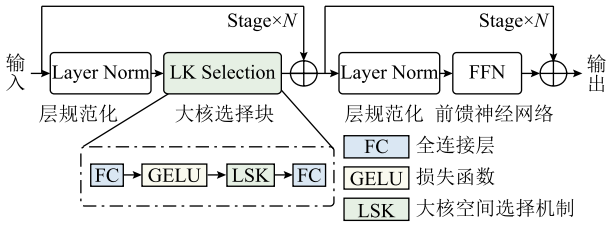



图 2 特征提取网络结构

Fig.2 Structure of the backbone

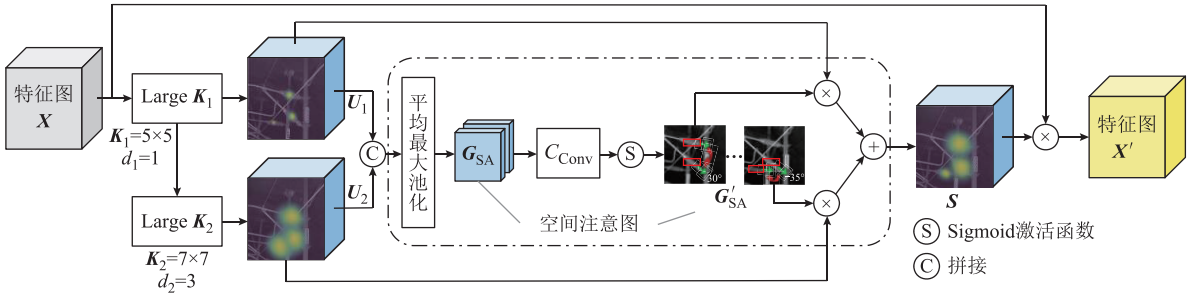



图 3 大核空间选择机制网络结构

Fig.3 Schematic of the LSK network

首先，图 3 中将层规范化后的特征图 $\boldsymbol{X}$ 输入串联大核卷积核 $\boldsymbol{K}_1$ 和 $\boldsymbol{K}_2$ 中，其中 $\boldsymbol{K}_1$ 卷积核由大小为 5×5、空洞率 $d_1$ 为 1 的膨胀卷积构成，$\boldsymbol{K}_2$ 卷积核由大小为 7×7、空洞率 $d_2$ 为 3 的膨胀卷积构成，达到理论卷积核大小为 29×29 的感受野范围 $\boldsymbol{R}_{\mathrm{RF}i}$，并将捕捉的 $\boldsymbol{R}_{\mathrm{RF}i}$ 由 1×1 卷积进一步处理，计算式为

$$\begin{cases}\boldsymbol{R}_{\mathrm{RF}i}=d_i(\boldsymbol{K}_i-1)+\boldsymbol{R}_{\mathrm{RF}i-1}\\ \boldsymbol{U}_i=C_{\mathrm{Conv}}^{1\times1}(\boldsymbol{R}_{\mathrm{RF}i})\end{cases} \tag{1}$$

式中，$\boldsymbol{R}_{\mathrm{RF}i}$ 为捕捉的第 $i$ 层感受野范围；$\boldsymbol{U}_i$ 为通过大核卷积核 $\boldsymbol{K}_i$ 生成的感受野范围；$d_i$ 为第 $i$ 个大核卷积核的空洞率；$C_{\mathrm{Conv}}^{1\times1}(\cdot)$为 1×1 卷积操作。

然后，将计算得到的感受野 $\boldsymbol{U}_1$ 和 $\boldsymbol{U}_2$ 进行拼接生成 $\boldsymbol{U}$，由平均池化和最大池化完成感受野空间信息的提取，得到空间注意图 $\boldsymbol{G}_{\mathrm{SA}}$，计算过程为

$$\begin{cases}\boldsymbol{G}_{\mathrm{SAmax}}=P_{\max}(\boldsymbol{U})\\ \boldsymbol{G}_{\mathrm{SAavg}}=P_{\mathrm{avg}}(\boldsymbol{U})\end{cases} \tag{2}$$

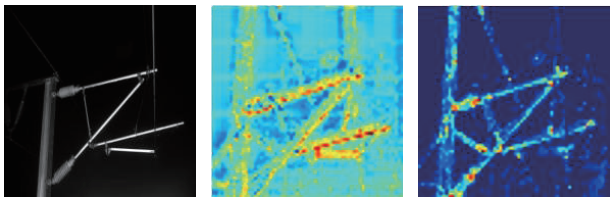
式中，$\boldsymbol{G}_{\mathrm{SAmax}}$ 为最大池化生成的空间注意力图；$\boldsymbol{G}_{\mathrm{SAavg}}$ 为平均池化生成的空间注意力图；$P_{\max}$ 为最大池化操作；$P_{\mathrm{avg}}$ 为平均池化操作。

最后，对空间注意图 $\boldsymbol{G}_{\mathrm{SA}}$ 进行卷积计算，将每张空间注意力图与其对应的空间感受野进行选择掩膜操作得到加权后的空间注意图 $\boldsymbol{G}'_{\mathrm{SA}}$，并与感受野范围 $\boldsymbol{U}_i$ 融合后得到空间注意力特征 $\boldsymbol{S}$，通过式（3）计算后，得到特征图 $\boldsymbol{X}'$。

$$\begin{cases}\boldsymbol{G}'_{\mathrm{SA}}=C_{\mathrm{Conv}}\left(\left[\boldsymbol{G}_{\mathrm{SAavg}};\boldsymbol{G}_{\mathrm{SAmax}}\right]\right)\\ \boldsymbol{S}=C_{\mathrm{Conv}}\left(\sum_{i=1}^{2}\boldsymbol{G}'_{\mathrm{SA}}\boldsymbol{U}_i\right)\\ \boldsymbol{X}'=\boldsymbol{X}\boldsymbol{S}\end{cases} \tag{3}$$

式中，$C_{\mathrm{Conv}}$ 为卷积操作；$[\boldsymbol{G}_{\mathrm{SAavg}};\boldsymbol{G}_{\mathrm{SAmax}}]$表示对空间注意图 $\boldsymbol{G}_{\mathrm{SAavg}}$ 和 $\boldsymbol{G}_{\mathrm{SAmax}}$ 的拼接操作。

采用可视化方法直观对比改进后的有效性，如图 4 所示。从图 4 可以看出，改进后的特征提取网络相较于原始特征提取网络，加强了对绝缘子部件的特征提取能力，同时大幅提高了前景和背景的区分能力，如图 4c 中冷色和暖色的对比区域。由此说明特征提取网络可以准确地对目标完成聚焦，从而抑制黑色背景对绝缘子的噪声干扰。



（a）接触网图像　（b）初始特征提取网络　（c）本文方法

图 4　特征图可视化对比

Fig.4　Feature map visualization comparison

### 1.2.2　检测解码器

由于接触网背景混杂，部分绝缘子会被其他障碍物遮挡，导致绝缘子缺陷特征信息丢失[19]。针对该问题，本文设计的解码器由多个检测头组成多尺度级联结构[20-21]，其中每个检测头阶段负责提升边界框预测的准确度，逐步将噪声框细化为边界框预测结果。检测解码器结构如图 5 所示。

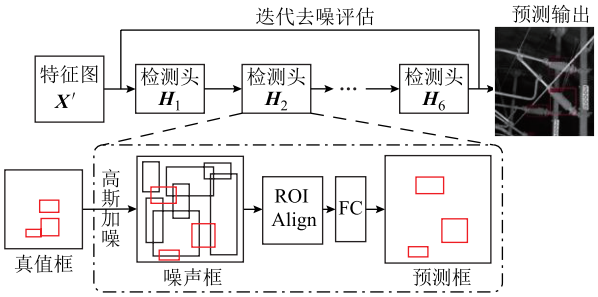



图 5　检测解码器结构

Fig.5　Detection decoder structure

首先，将一组绝缘子真值框（Ground Truth, GT）进行高斯加噪，从而得到一组带有噪声的预测框。加噪过程中，将绝缘子真值框坐标$(x_i, y_i, w_i, h_i)$与高斯噪声矩阵相加，得到多个噪声框坐标值$(x_i', y_i', w_i', h_i')$，前向噪声过程定义如式（4）和式（5）所示，加噪公式如式（6）所示。

$$q(z_t \mid z_0) \sim N\left(z_t; \sqrt{\bar{\alpha}_t} z_0, (1-\bar{\alpha}_t)\sigma^2\right) \quad (4)$$

正则化分布为

$$q(z_t \mid z_0) = \sqrt{\bar{\alpha}_t} z_0 + \varepsilon\sqrt{1-\bar{\alpha}_t} \quad \varepsilon \sim N(0, \sigma^2) \quad (5)$$

$$(x'_{i+1}, y'_{i+1}, w'_{i+1}, h'_{i+1}) = D_{\text{Diffuse}}(x_i, y_i, w_i, h_i) \quad (6)$$

式中，$q$ 为前向噪声过程的条件概率分布；$N(\cdot)$表示正态分布；$z_0$ 为初始绝缘子预测样本；$z_t$ 为 $t$ 时步的预测噪声样本；$\bar{\alpha}_t$ 为噪声方差矩阵；$\varepsilon$ 服从均值为 0、方差为 $\sigma^2$ 的正态分布；$D_{\text{Diffuse}}$ 表示对预测框的加噪操作。

然后，初始绝缘子特征图 $X$ 根据图像解码器得到的绝缘子特征图 $X'$ 和噪声框输入感兴趣区域对齐（Region of Interest Align, ROI Align）算法，通过特征剪裁和全连接层得到预测框。而后添加高斯噪声输入下一时刻的检测头 $H_2$，经过循环最终从检测头 $H_6$ 得到预测框。此时，根据式（7），预测上一时刻检测框的正态分布状态，经过叠加迭代次数，将噪声框坐标还原为真值框坐标，从而完成迭代去噪评估。计算过程为

$$p_\theta(z_{t-1} \mid z_t) \sim N\left(z_{t-1}; \mu_\theta(z_t, t), \sum_\theta(z_t, t)\right) \quad (7)$$

$$P_t = H_1(z_t) + H_2(z_t) + \cdots + H_6(z_t) \quad (8)$$

式中，$\theta$ 为权重矩阵；$\mu_\theta$ 为在权重矩阵 $\theta$ 下的正态分布均值；$P_t$ 为 $t$ 时步的预测框结果；$H(\cdot)$为检测头中对特征图中绝缘子预测框的逐步预测；$\sum_\theta(z_t, t)$ 为在权重矩阵 $\theta$ 下的正态分布的协方差矩阵，表示给定绝缘子预测样本 $z_t$ 的情况下，预测的下一时刻的绝缘子样本 $z_{t+1}$ 的不确定性或方差。

## 1.3　绝缘子缺陷描述阶段

在完成特征提取基础上，本文通过跨模态映射模块[22]，完成对绝缘子图像特征和文本特征的关系映射，语言建模编码器用于生成包含绝缘子缺陷特征信息的文本描述语句，其结构如图 6 所示。

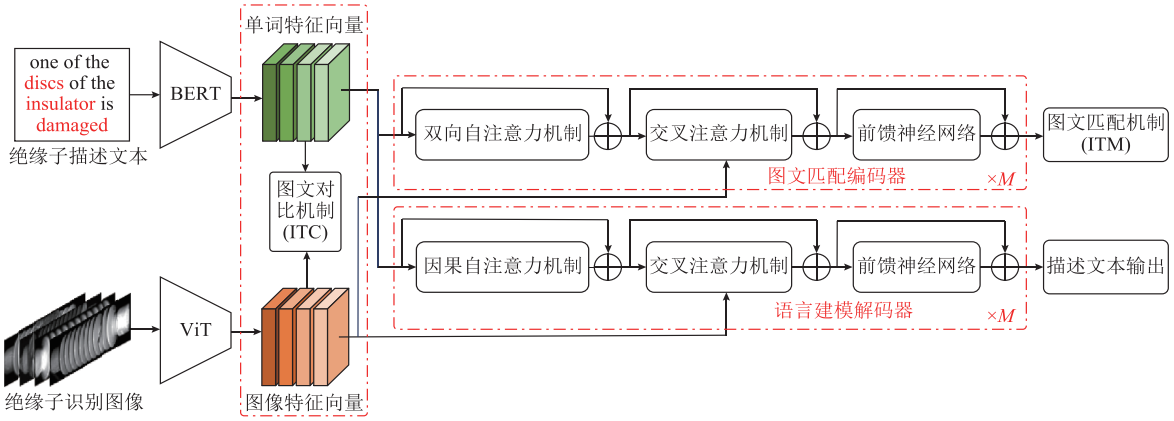



图 6　跨模态融合模块结构

Fig.6　Cross-modal fusion module structure

图 6 中，将生成的图像特征向量和单词特征向量初步进行图文对比损失计算，并将向量同步输入图文匹配编码器内。单词特征向量首先由双向自注意力机制完成建模，双向自注意力相比自注意力机制综合了 Token 序列前向和后向的文本描述信息，因此能够处理输入 Token 序列中的变化和噪声，提高了模型对不同 Token 排列方式的适应性，降低了模型对 Token 序列顺序的敏感性。双向自注意力计算式为

$$\mathrm{BackAtt}\left(\boldsymbol{Q}_i,\boldsymbol{K}_i,\boldsymbol{V}_i\right)=\mathrm{softmax}\left(\frac{\boldsymbol{Q}_{ib}\boldsymbol{K}_{ib}^{\mathrm{T}}}{\sqrt{k}}\right)\boldsymbol{V}_i \tag{9}$$

$$\mathrm{BiSelfAtt}(\boldsymbol{x}_i^{\mathrm{t}})=\mathrm{concat}\left[\mathrm{Att}\left(\boldsymbol{x}_i^{\mathrm{t}}\right),\mathrm{BackAtt}\left(\boldsymbol{x}_i^{\mathrm{t}}\right)\right] \tag{10}$$

式中，BackAtt 表示序列位置 $i$ 之前的自注意力计算；Att 表示序列位置 $i$ 之后的自注意力计算；BiSelfAtt 表示整体序列长度的自注意力计算；$\boldsymbol{x}_i^{\mathrm{t}}$ 为文本 Token 序列 $\boldsymbol{T}_{\mathrm{Token}}=\{\boldsymbol{x}_1^{\mathrm{t}},\boldsymbol{x}_2^{\mathrm{t}},\cdots,\boldsymbol{x}_n^{\mathrm{t}}\}$ 中的向量表示，$n$ 为序列长度；上角标 t 代表文本标识符；下角标中 b 表示后向注意力计算标识符；$\boldsymbol{Q}_i$、$\boldsymbol{K}_i$、$\boldsymbol{V}_i$ 分别为序列位置 $i$ 的查询集合、键集合、值集合的向量形式；$k$ 为每个 Token 中的词嵌入向量的维度；softmax 为激活函数；concat 表示拼接操作。

然后通过交叉注意力机制，拼接文本特征查询矩阵 $\boldsymbol{W}_{\mathrm{Q}}^{\mathrm{t}}$ 和视觉特征键矩阵 $\boldsymbol{W}_{\mathrm{K}}^{\mathrm{i}}$，并与视觉特征值矩阵 $\boldsymbol{W}_{\mathrm{V}}^{\mathrm{i}}$ 进行点乘运算。具体过程为：首先将文本序列 $\boldsymbol{T}_{\mathrm{Token}}$ 中 $\boldsymbol{x}_1^{\mathrm{t}}\in\mathbf{R}^{n\times k}$ 对应的文本特征查询矩阵 $\boldsymbol{W}_{\mathrm{Q1}}^{\mathrm{t}}$ 和图像序列 $\boldsymbol{I}_{\mathrm{Token}}$ 中 $\boldsymbol{x}_2^{\mathrm{i}}\in\mathbf{R}^{n\times k}$ 对应的视觉特征键矩阵 $\boldsymbol{W}_{\mathrm{K2}}^{\mathrm{i}}$，以及图像序列 $\boldsymbol{I}_{\mathrm{Token}}$ 中 $\boldsymbol{x}_2^{\mathrm{i}}\in\mathbf{R}^{n\times k}$ 对应的视觉特征值矩阵 $\boldsymbol{W}_{\mathrm{V2}}^{\mathrm{i}}$ 进行交叉注意力计算，最终生成文本和图像的混合特征标识向量，完成图文对比损失。$\boldsymbol{x}_1^{\mathrm{t}}$ 和 $\boldsymbol{x}_2^{\mathrm{i}}$ 的交叉注意力计算式为

$$\mathrm{CrossAtt}(\boldsymbol{x}_1^{\mathrm{t}},\boldsymbol{x}_2^{\mathrm{i}})=\mathrm{softmax}\left(\frac{\left(\boldsymbol{x}_1^{\mathrm{t}}\boldsymbol{W}_{\mathrm{Q1}}^{\mathrm{t}}\right)\left(\boldsymbol{x}_2^{\mathrm{i}}\boldsymbol{W}_{\mathrm{K2}}^{\mathrm{i}}\right)^{\mathrm{T}}}{\sqrt{k}}\right)(\boldsymbol{x}_2^{\mathrm{i}}\boldsymbol{W}_{\mathrm{V2}}^{\mathrm{i}}) \tag{11}$$

式中，CrossAtt 表示对文本 Token 序列和图像 Token 序列进行交叉注意力计算，从而建立文本特征与视觉特征的联系；上角标 i 代表图像标识符。

语言建模解码器中，因果自注意力是一种在生成文本序列过程时，能够生成符合文本序列规范[23]的注意力机制。相比自注意力机制添加了掩码，通过屏蔽当前 Token 序列所在位置之后的序列 Tokens，有助于模型利用当前序列位置之前的信息，从而提高在文本生成任务中的性能，因果自注意力计算公式为

$$\mathrm{CausalSelfAtt}\left(\boldsymbol{Q}_i\right)=\sum_{j=1}^{i}\mathrm{Att}\left(\boldsymbol{Q}_i,\boldsymbol{K}_j\right)\boldsymbol{V}_j \tag{12}$$

式中，CausalSelfAtt 表示进行注意力计算时，每个 Token 只能计算当前及之前的 Token，而不能计算之后的 Token，从而限制注意力范围。式（12）表明，$\boldsymbol{Q}_i$ 只能与序列中位置 $i$ 之前的 $\boldsymbol{K}_j$ 和 $\boldsymbol{V}_j$ 进行注意力计算，不受位置 $i$ 之后 $\boldsymbol{K}_j$ 和 $\boldsymbol{V}_j$ 的影响，防止生成描述被后续 Token 干扰。之后文本和图像再次由交叉注意力进行对齐操作，通过前馈神经网络后进行语言建模损失计算。

### 1.4 损失函数

总损失函数由接触网绝缘子目标检测和缺陷描述两部分损失函数构成，其中目标检测阶段采用多任务损失函数 $L_{\mathrm{det}}$，主要由分类损失 $L_{\mathrm{cls}}$、预测框回归损失 $L_{\mathrm{L1}}$ 和预测框优化 $L_{\mathrm{giou}}$ 损失函数组成，表示为

$$L_{\mathrm{det}}=\lambda_{\mathrm{cls}}L_{\mathrm{cls}}+\lambda_{\mathrm{L1}}L_{\mathrm{L1}}+\lambda_{\mathrm{giou}}L_{\mathrm{giou}} \tag{13}$$

式中，$\lambda_{\mathrm{cls}}$、$\lambda_{\mathrm{L1}}$、$\lambda_{\mathrm{giou}}$ 分别为损失函数分量 $L_{\mathrm{cls}}$、$L_{\mathrm{L1}}$、$L_{\mathrm{giou}}$ 的权重。绝缘子缺陷描述阶段的总损失 $L_{\mathrm{cap}}$ 主要由图文对比损失 $L_{\mathrm{ITC}}$、图文匹配损失 $L_{\mathrm{ITM}}$、语言建模损失 $L_{\mathrm{LM}}$ 三部分组成，表示为

$$L_{\mathrm{cap}}=L_{\mathrm{ITC}}+L_{\mathrm{ITM}}+L_{\mathrm{LM}} \tag{14}$$

图文对比损失利用单个模态的输出进行对比学习，将图像编码器和文本编码器获得的图像特征向量和单词特征向量，分别计算图片向量与单词向量队列中所有样本的相似度 $\boldsymbol{p}^{\mathrm{i2t}}$ 和单词向量与图像向量队列中所有样本的相似度 $\boldsymbol{p}^{\mathrm{t2i}}$。图文对比损失公式为

$$\begin{aligned}L_{\mathrm{ITC}}=-\frac{1}{2}\Big\{&E_{\left(\boldsymbol{x}^{\mathrm{i}},\boldsymbol{x}^{\mathrm{t}}\right)\in\boldsymbol{D}}\left[\sum\boldsymbol{y}^{\mathrm{i2t}}\left(\boldsymbol{x}^{\mathrm{i}}\right)\ln\left(\boldsymbol{p}^{\mathrm{i2t}}(\boldsymbol{x}^{\mathrm{i}})\right)\right]+\\&E_{\left(\boldsymbol{x}^{\mathrm{i}},\boldsymbol{x}^{\mathrm{t}}\right)\in\boldsymbol{D}}\left[\sum\boldsymbol{y}^{\mathrm{t2i}}\left(\boldsymbol{x}^{\mathrm{t}}\right)\ln\left(\boldsymbol{p}^{\mathrm{t2i}}(\boldsymbol{x}^{\mathrm{t}})\right)\right]\Big\}\end{aligned} \tag{15}$$

式中，$E$ 为图像和文本的相似期望；$\boldsymbol{y}^{\mathrm{i2t}}(\cdot)$为图片向量与单词向量队列中所有样本的相似度；$\boldsymbol{y}^{\mathrm{t2i}}(\cdot)$为单词向量与图像向量队列中所有样本的相似度；$\boldsymbol{x}^{\mathrm{i}}$ 为图像样本矩阵；$\boldsymbol{x}^{\mathrm{t}}$ 为文本样本矩阵；$\boldsymbol{D}$ 为训练数据集。

图文匹配损失通过学习图像和文本的多模态表示，根据多模态特征预测一个图像文本对是否匹配，从而进行二分类任务，图文匹配损失公式为

$$L_{\mathrm{ITM}}=-E_{\left(\boldsymbol{x}^{\mathrm{i}},\boldsymbol{x}^{\mathrm{t}}\right)\in\boldsymbol{D}}\left[\sum\boldsymbol{y}^{\mathrm{itm}}(\boldsymbol{x}^{\mathrm{i}},\boldsymbol{x}^{\mathrm{t}})\ln\left(\boldsymbol{p}^{\mathrm{itm}}(\boldsymbol{x}^{\mathrm{i}},\boldsymbol{x}^{\mathrm{t}})\right)\right] \tag{16}$$

式中，$\boldsymbol{y}^{\mathrm{itm}}(\cdot)$为真值文本对匹配值；$\boldsymbol{p}^{\mathrm{itm}}(\cdot)$为预测文本对匹配值。

语言建模损失是以图像为基础的文本解码器，通过交叉熵损失函数完成对下一个 Token 的预测，以自回归方式对应文本概率，计算公式为

$$L_{\mathrm{LM}}=-E_{\boldsymbol{x}^{\mathrm{t}}\in\boldsymbol{D}}\left[\sum_{l=l_{\min}}^{l_{\max}}\ln\delta\left(\boldsymbol{x}_l^{\mathrm{t}}\middle|\boldsymbol{x}_{<l}^{\mathrm{t}}\right)\right] \tag{17}$$

式中，$\delta\left(\boldsymbol{x}_l^{\mathrm{t}}\middle|\boldsymbol{x}_{<l}^{\mathrm{t}}\right)$表示数据集中，预测为正确单词的概率；$l$ 为当前语句序列长度；$l_{\max}$ 为语句序列最大长度；$l_{\min}$ 为语句序列最小长度；$\boldsymbol{x}_{<l}^{\mathrm{t}}$ 为历史词；$\boldsymbol{x}_l^{\mathrm{t}}$ 为当前词。

## 2　实验验证

### 2.1　接触网绝缘子缺陷数据集

本文使用的接触网数据集由兰新高铁和青藏铁路格尔木段现场拍摄的接触网图像，以及其他相关接触网图像资源构成。采集接触网图像时，在距离铁路沿线安全距离外进行现场拍摄，以及通过接触网检测车在高速铁路不同维修天窗期内进行拍摄。并且由 Labelme 工具对上述接触网图像进行标注，形成了 2 624 幅分辨率为 512×512 的接触网数据集，其中 70%为训练集，30%为验证集和测试集。同时制作了绝缘子缺陷描述数据集用于描述生成任务，通过对绝缘子缺陷图像进行图像镜像、图像反转等操作，扩充绝缘子缺陷图像样本集。经过扩充后，生成共 800 幅绝缘子描述图像，其中主要缺陷包括 200 幅异物侵入图像、162 幅绝缘子破损图像、124 幅伞裙裂缝图像、124 幅闪络烧伤图像及 190 幅正常绝缘子图像。此外，为了提高绝缘子缺陷文本描述数据集的质量，本文又采用同义词描述的方式，并通过 COCO_Caption[24]的标注方法对数据集进行标注，对每幅绝缘子缺陷图像采用 5 句 COCO_Caption 风格的标注语句进行描述，进一步扩充了缺陷文本描述的数量，最终生成绝缘子缺陷描述文本语句 4 000 条。

为了验证所提模型训练的有效性，对本文采用的绝缘子缺陷描述阶段的总损失 $L_{\mathrm{cap}}$ 训练曲线进行可视化输出，如图 7 所示，可以看出，所提模型较好地实现了收敛并趋于稳定。

为了进一步验证本文所提模型的可靠性，选择模型预测值与真实值之间的方差作为评价指标，对模型不同训练程度下的预测性能进行可视化输出显示，方差计算曲线如图 8 所示。方差描述了在不同迭代阶段的训练模型中，训练数据预测值的变化波动情况。方差越大，说明预测值与均值的差异越大；

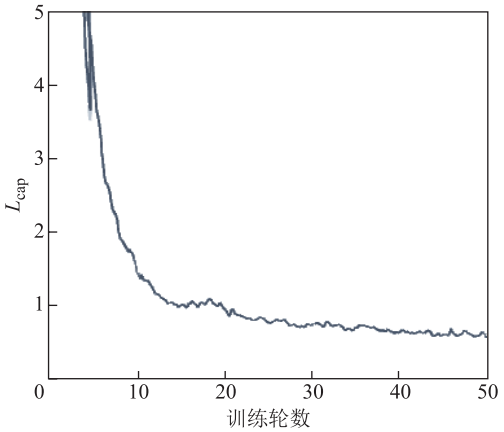



图 7　损失函数曲线

Fig.7　Loss function curve

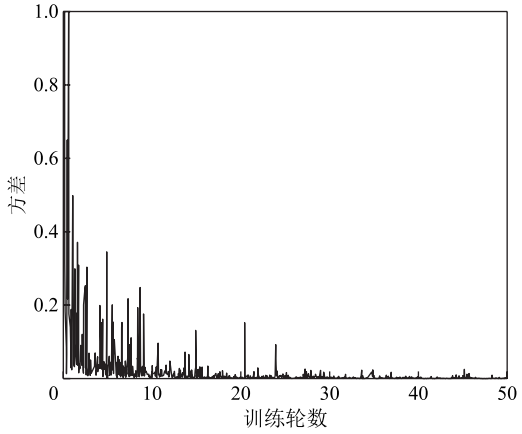



图 8　方差计算曲线

Fig.8　Variance calculation curve

方差越小，说明预测值与均值的差异越小，模型稳定性越高。此外，高方差表明模型对训练数据的敏感程度高，易导致过拟合。从图 8 可以看出，本文所提方法在绝缘子缺陷文本描述时，方差整体逐渐趋于平稳，说明所提方法在预测数据上的稳定性及泛化能力逐步提高，具有较好的可靠性。

### 2.2　绝缘子目标检测和缺陷描述实验

训练环境使用 PyTorch 深度学习框架，搭载 10vCPU Intel Xeon Gold 6248R 处理器，配备 NVIDIA A100 PCIE(40GB) GPU 和 72 GB 内存。模型预处理参数中，语句最小长度设置为 5，最大长度设置为 30，学习率设置为 0.000 01，训练轮数 epoch 为 50。

#### 2.2.1　绝缘子目标检测评估指标

为了定量评价检测性能，采用平均准确率（mean Average Precision, mAP）、平均召回率（Average Recall, AR）、调和平均值（F1-Score）指标对本文所提方法以及其他文献中的方法进行定量分析，评价值越大表明模型检测性能越好，评估指标对比结果见表 1。

**表 1 评估指标对比实验**

Tab.1 Evaluation metric comparison experiment

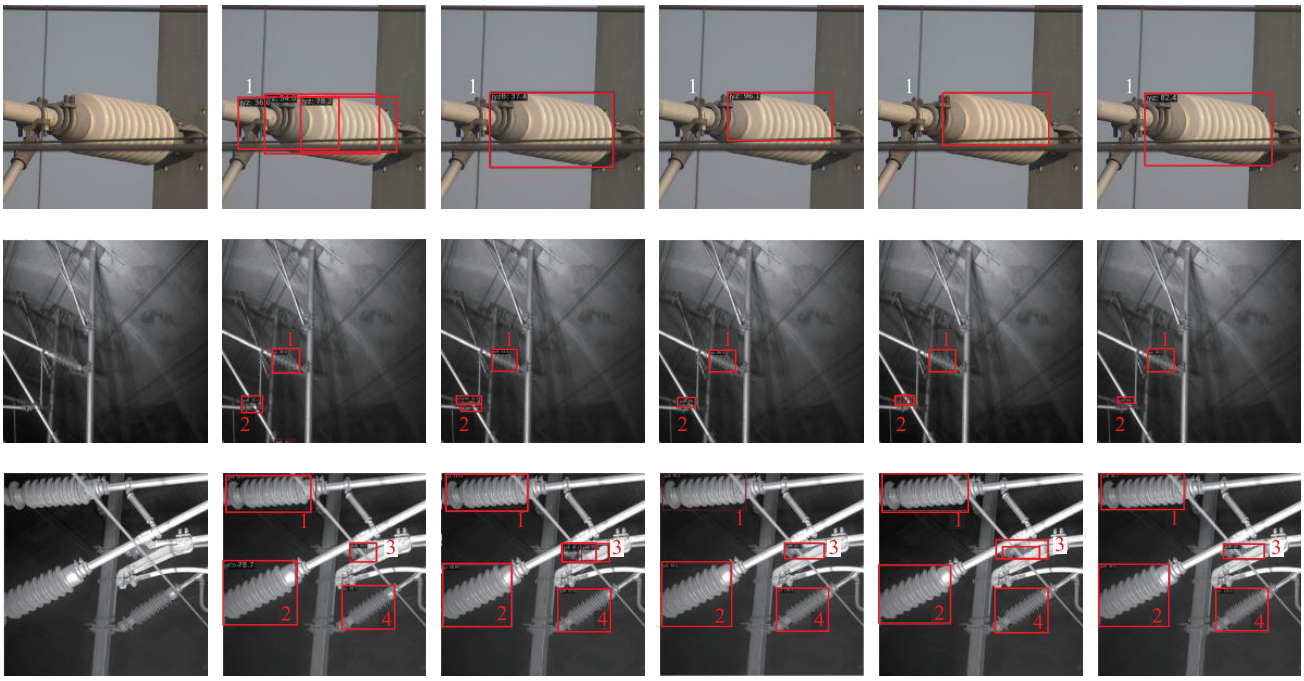
（%）

| 方法 | $mAP_{0.5}$ | $mAP_{0.5:0.95}$ | AR | F1-Score |
|---|---|---|---|---|
| YOLOv8[25] | 78.91 | 54.21 | 72.27 | 69.30 |
| GLIP[26] | 87.73 | 69.25 | 81.47 | 79.61 |
| DETR[7] | 88.92 | 72.25 | 84.47 | 81.61 |
| Faster R-CNN[8] | 87.96 | 68.41 | 73.77 | 75.91 |
| 本文方法 | 93.04 | 75.41 | 83.22 | 82.91 |

从表 1 可知，与 DETR 和 Faster R-CNN 方法相比，所提方法在交并比（Intersection over Union, IOU）阈值为 0.5 的范围内，$mAP_{0.5}$ 分别提升 4.63%和 5.78%。在 IOU 阈值为 0.5～0.95 的范围内，$mAP_{0.5:0.95}$ 达到 75.41%，较 DETR 和 Faster RCNN 分别提升 4.37%和 10.23%，同时 F1-Score 达到最高值 82.91%，表示所提模型的检测准确性更好。

### 2.2.2 绝缘子目标检测实验

首先，对接触网图像中的陶瓷绝缘子、悬式绝缘子和柱式绝缘子进行检测；同时，与表 1 中方法进行对比；最后，为了解释绝缘子检测的可靠性，将绝缘子标注号码添加到下列检测图像中，实验结果如图 9 所示。



（a）接触网原图 （b）YOLOv8[25] （c）GLIP[26] （d）DETR[7] （e）Faster R-CNN[8] （f）本文模型

图 9 接触网绝缘子检测实验结果

Fig.9 Test results of catenary insulators

由实验结果可以看出，YOLOv8 模型检测结果存在边框重叠问题，如图 9b 第一行的 1 号绝缘子、第二行的 2 号绝缘子。GLIP 模型同样存在检测框重叠问题，如图 9c 第三行的 3 号绝缘子。文献[7]中 DETR 模型和文献[8]中 Faster R-CNN 模型均存在检测不完整的问题，如图 9d 和图 9e 第一行中，线缆以下绝缘子部分均未被成功检测。而本文模型对不同距离和不同方向的绝缘子均可以进行准确识别。

### 2.2.3 绝缘子缺陷描述实验

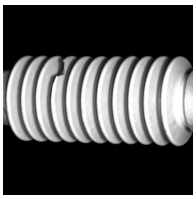
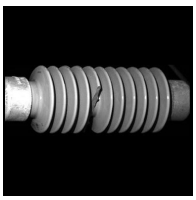
在完成绝缘子检测实验后，进一步完成绝缘子图像的缺陷描述实验。绝缘子缺陷图像主要包含异物侵入、破损缺陷、闪络烧伤、裂缝缺陷以及完好的绝缘子图像，与基于 LSTM 和 LLaVA 的图像描述生成模型进行绝缘子描述对比实验，结果见表 2。

首先在绝缘子异物侵入描述部分，从表 2 可以看出，基于 LSTM 模型的描述准确性较差，对于绝缘子缺陷，该方法无法正确完成缺陷文本描述；而基于 LLaVA 的模型虽然识别出图①的塑料袋，但未能识别图②中的石子，并将其误判为裂缝缺陷，用“a crack in it”描述；而本文方法描述准确性更高。例如，对图①描述结果为“a white plastic bag is wrapped around the insulator”，表明存在一个白色塑料袋包裹在绝缘子表面；对图②用“a stone is in the middle of the insulator”描述，翻译为绝缘子的中间部分有一块石头，同时“middle”突出了该缺陷在图像中的相对位置。对于表 2 中第二行图像③

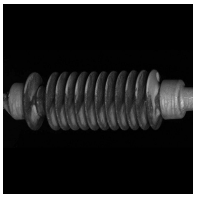
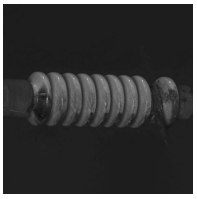
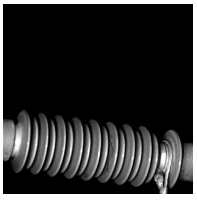
和④，基于 LLaVA 的模型基本可以描述绝缘子的缺陷类型，但存在误识别的问题，如将破损识别为裂缝；相比之下，本文模型对缺陷部位的描述更加详尽，如对图④描述为“one of the discs of the insulator is damaged”，翻译为绝缘子的一个伞片是破损的，可见所提方法能够准确描述绝缘子的

**表 2　绝缘子缺陷图像描述对比**

Tab.2　Comparison of insulator defect description effect

| 缺陷图像 | 图像描述 |
|---|---|
|  ①异物侵入 | 1. LSTM<br>caption: “an insulator can be used normally”<br>译：一个正常的绝缘子<br>2. LLaVA<br>caption: “the insulator in the image has a piece of plastic, wrapped around it”<br>译：图中的绝缘体被一块塑料包裹着<br>3. 本文模型<br>caption: “a white plastic bag is wrapped around the insulator”<br>译：一个白色塑料袋包裹在绝缘子表面 |
|  ②异物侵入 | 1. LSTM<br>caption: “a good insulator without damage”<br>译：一个正常使用的绝缘子会导致问题<br>2. LLaVA<br>caption: “the problem with the insulator in the image is that it has a crack in it”<br>译：图中绝缘体的问题是存在一条裂缝<br>3. 本文模型<br>caption: “a stone is in the middle of the insulator”<br>译：绝缘子的中间部分有一块石头 |
|  ③破损缺陷 | 1. LSTM<br>caption: “an insulator can be used normally”<br>译：一个正常的绝缘子<br>2. LLaVA<br>caption: “there is a crack in the insulator”<br>译：绝缘子存在一条裂缝<br>3. 本文模型<br>caption: “the damaged area is in the middle of the insulator”<br>译：绝缘子中间有一片破损区域 |
|  ④破损缺陷 | 1. LSTM<br>caption: “a broken insulator”<br>译：一个故障的绝缘子<br>2. LLaVA<br>caption: “the insulator in the image has a crack in it, which is a sign of damage or wear”<br>译：图中的绝缘子有裂缝，这是损坏或磨损的迹象<br>3. 本文模型<br>caption: “one of the discs of the insulator is damaged”<br>译：绝缘子的一个伞片是破损的 |

（续）

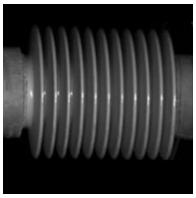
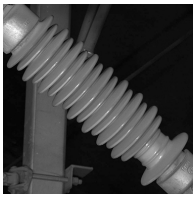
| 缺陷图像 | 图像描述 |
|---|---|
|  ⑤闪络烧伤 | 1. LSTM<br>caption: “a disc of the insulator is intact”<br>译：绝缘子的伞片是完整的<br>2. LLaVA<br>caption: “the insulator in the image is damaged”<br>译：图中的绝缘子损坏<br>3. 本文模型<br>caption: “the insulator was burned”<br>译：绝缘子被烧伤 |
|  ⑥闪络烧伤 | 1. LSTM<br>caption: “an insulator can be used normally”<br>译：一个正常的绝缘子<br>2. LLaVA<br>caption: “the insulator in the image has a damage in it, which is a defect that could potentially cause electrical issues or even a short circuit”<br>译：图中的绝缘体有损伤，这是一个可能导致电气问题甚至短路的缺陷<br>3. 本文模型<br>caption: “an insulator with a black burn mark”<br>译：绝缘子有一片黑色烧伤痕迹 |
|  ⑦裂缝缺陷 | 1. LSTM<br>caption: “a insulator can cause problems”<br>译：一个会导致问题的绝缘子<br>2. LLaVA<br>caption: “the insulator in the image has a crack in it”<br>译：图中的绝缘体有一条裂缝<br>3. 本文模型<br>caption: “a crack in the insulator”<br>译：一个存在裂缝的绝缘子 |
|  ⑧裂缝缺陷 | 1. LSTM<br>caption: “a disc of the insulator is intact”<br>译：绝缘子的伞片是完整的<br>2. LLaVA<br>caption: “there is damage to the insulator. It has a crack in it, which indicates that it is not in good condition and may not be functioning properly”<br>译：绝缘子存在损坏，有一个裂缝在绝缘子上，因此表明绝缘子状态不好，可能不会正常工作<br>3. 本文模型<br>caption: “the crack is in the middle disk of the insulator”<br>译：裂缝在绝缘子的中部伞片 |

破损缺陷。在绝缘子闪络烧伤描述方面，实验结果见表 2 第三行图像⑤和⑥。基于 LSTM 的模型无法识别烧伤病斑；LLaVA 模型将烧伤识别为“damage”，无法描述准确的缺陷类型；而本文模型通过描述语句可以直观地反映烧伤类型，如“a black burn mark”表示一片黑色的烧伤印记。最后，在绝缘子裂缝缺陷描述方面，在表 2 第四行图像⑦和⑧中，基于 LLaVA 的模型描述准确性较 LSTM 模型有所提高，但仍存在描述歧义的问题。综合上述实验结果可以得出，所提方法能够准确地完成对不同类型缺陷的描述。

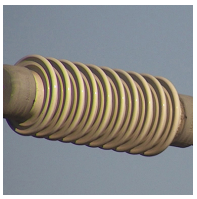
在完成对绝缘子缺陷描述的实验后，为了进一步验证绝缘子描述的有效性，对不同场景下的正常绝缘子进行描述实验，结果见表 3。其中，基于 LSTM 的图像描述模型将四张正常绝缘子图像描述为“a good insulator without damage”，翻译为没有损伤的绝缘子，该模型描述较为单一。此外，基于 LLaVA 的模型以介绍图像内容和描述绝缘子形状为主，如“image”“show”和“large”等词虽然反映了绝缘子的形状，但忽略了绝缘子的状态。而本文模型对绝缘子的正常状态文本描述更加准确，并且可以提供更丰富的语句文本描述。

**表 3 绝缘子正常图像描述对比**

Tab.3 Comparison of normal image description effect of insulator

| 正常图像 | 图像描述 |
|---|---|
|  | 1. LSTM<br>caption: “a good insulator without damage”<br>译：一个没有损伤的完整绝缘子<br>2. LLaVA<br>caption: "the problem with the insulator in the image is that it is broken"<br>译：图中的绝缘子损坏了<br>3. 本文模型<br>caption: “a good insulator without damage”<br>译：一个没有损伤的绝缘子 |
|  | 1. LSTM<br>caption: “a good insulator without damage”<br>译：一个没有损伤的完整绝缘子<br>2. LLaVA<br>caption: “the image shows a normal insulator”<br>译：图中展示了一个普通的绝缘子<br>3. 本文模型<br>caption: “an insulator can be used normally”<br>译：一个正常使用的绝缘子 |

（续）

| 正常图像 | 图像描述 |
|---|---|
|  | 1. LSTM<br>caption: “a good insulator without damage”<br>译：一个没有损伤的完整绝缘子<br>2. LLaVA<br>caption: “the insulator in the image is large”<br>译：一个大绝缘子在图中<br>3. 本文模型<br>caption: “a good insulator without damage”<br>译：一个没有损伤的绝缘子 |
|  | 1. LSTM<br>caption: “a good insulator without damage”<br>译：一个没有损伤的完整绝缘子<br>2. LLaVA<br>caption: “the image shows a large insulator”<br>译：图中展示的是一个大绝缘子<br>3. 本文模型<br>caption: “an insulator can be used normally”<br>译：一个正常使用的绝缘子 |

### 2.2.4 绝缘子缺陷描述评估指标

最后进行文本描述定量评价，采用双语评估替补评价指标（Bilingual Evaluation Understudy, BLEU）、基于共识的图像描述评价指标（Consensus-based Image Description Evaluation, CIDEr）、基于召回率的评估指标（Recall-Oriented Understudy for Gisting Evaluation, ROUGE-L）、显式排序的翻译评估指标（Metric for Evaluation of Translation with Explicit Ordering, METEOR）以及语义命题的图像描述指标（Semantic Propositional Image Caption Evaluation, SPICE）进行性能测试，结果见表 4。由表 4 可以看出，本文方法的平均双语评估替换值 BLEU 达到 83.51%，相比 LSTM 模型提升 22.52%，相比 LLaVA 模型提升 8.93%，同时在 CIDEr、ROUGE-L、METEOR

**表 4 图像描述模型指标对比**

Tab.4 Comparison of image caption model

| 评价指标 | LSTM | LLaVA | 本文方法 |
|---|---|---|---|
| BLEU-1(%) | 72.29 | 81.77 | 88.00 |
| BLEU-2(%) | 69.68 | 80.69 | 84.18 |
| BLEU-3(%) | 66.46 | 73.65 | 81.86 |
| BLEU-4(%) | 64.22 | 70.54 | 79.99 |
| CIDEr | | 1.85 | 1.94 |
| ROUGE-L(%) | | 78.48 | 81.59 |
| METEOR(%) | | 43.44 | 51.50 |
| SPICE(%) | | 31.25 | 37.88 |

和 SPICE 指标上均达到最高，分别为 1.94、81.59%、51.50%、37.88%，由此可证明本文方法对不同绝缘子具有检测精度高、缺陷语义描述正确和语句多样性丰富的优点。

## 3　结论

针对绝缘子缺陷检测时易受复杂环境背景干扰，导致缺陷检测精度低以及无法提供缺陷语义描述的问题，提出一种绝缘子缺陷图像描述生成模型。通过实验对比分析，可以得出以下结论：

1）采用大核空间选择特征提取网络作为编码器，增强了绝缘子缺陷检测网络对关键特征的提取能力。

2）设计融合扩散机制的检测解码器，对解码器生成的噪声框进行逆向贝叶斯扩散，还原对绝缘子真值框的预测，克服了复杂环境背景干扰的问题，提高了预测框匹配的准确性。

3）设计跨模态映射模块，完成对绝缘子图像缺陷特征和文本特征的关系映射，并通过语言建模编码器输出绝缘子缺陷文本描述，完成检测任务。

4）实验结果表明所提绝缘子检测模型的平均准确率、平均召回率和 F1-Score 分别达到 93.04%、83.22%和 82.91%，平均双语评估替换值 BLEU 达到 83.51%，具有更高的检测精度及缺陷描述性能，能够满足对绝缘子缺陷检测的需求。

## 参考文献


[1] 张血琴, 周志鹏, 郭裕钧, 等. 不同材质绝缘子污秽等级高光谱检测方法研究[J]. 电工技术学报, 2023, 38(7): 1946-1955.
Zhang Xueqin, Zhou Zhipeng, Guo Yujun, et al. Detection method of contamination grades of insulators with different materials based on hyperspectral technique[J]. Transactions of China Electrotechnical Society, 2023, 38(7): 1946-1955.

[2] 余颖, 刘亚东, 李维, 等. 配电线路针式绝缘子早期故障动态特性研究[J]. 电工技术学报, 2023, 38(1): 71-82.
Yu Ying, Liu Yadong, Li Wei, et al. Simulation and experimental research on pin insulator incipient fault dynamic characteristic in the distribution network[J]. Transactions of China Electrotechnical Society, 2023, 38(1): 71-82.

[3] Tan Ping, Li Xufeng, Xu Jinmei, et al. Catenary insulator defect detection based on contour features and gray similarity matching[J]. Journal of Zhejiang University: Science A, 2020, 21(1): 64-73.

[4] 顾桂梅, 陈国翠. 改进GA-BP算法的棒式绝缘子表面缺陷识别[J]. 铁道科学与工程学报, 2022, 19(2): 546-553.
Gu Guimei, Chen Guocui. Surface defect recognition of bar insulator based on improved GA-BP algorithm [J]. Journal of Railway Science and Engineering, 2022, 19(2): 546-553.

[5] 李斌, 屈璐瑶, 朱新山, 等. 基于多尺度特征融合的绝缘子缺陷检测[J]. 电工技术学报, 2023, 38(1): 60-70.
Li Bin, Qu Luyao, Zhu Xinshan, et al. Insulator defect detection based on multi-scale feature fusion[J]. Transactions of China Electrotechnical Society, 2023, 38(1): 60-70.

[6] Tan Ping, Li Xufeng, Ding Jin, et al. Mask R-CNN and multifeature clustering model for catenary insulator recognition and defect detection[J]. Journal of Zhejiang University: Science A, 2022, 23(9): 745-756.

[7] Wen Feng, Wang Mei, Hu Xiaojie. DFAM-DETR: deformable feature based attention mechanism DETR on slender object detection[J]. IEICE Transactions on Information and Systems, 2023, E106.D(3): 401-409.

[8] Chen Yanping, Deng Chong, Sun Qiang, et al. Lightweight detection methods for insulator self-explosion defects[J]. Sensors, 2024, 24(1): 290.

[9] Yang Zuopeng, Wang Pengbo, Chu Tianshu, et al. Human-centric image captioning[J]. Pattern Recognition, 2022, 126: 108545.

[10] 谢州益, 冯亚枝, 胡彦蓉, 等. 基于 ResNet18 特征编码器的水稻病虫害图像描述生成[J]. 农业工程学报, 2022, 38(12): 197-206.
Xie Zhouyi, Feng Yazhi, Hu Yanrong, et al. Generating image description of rice pests and diseases using a ResNet18 feature encoder[J]. Transactions of the Chinese Society of Agricultural Engineering, 2022, 38(12): 197-206.

[11] Li Chunyuan, Wong C, Zhang Sheng, et al. LLaVA-Med: training a large language-and-vision assistant for biomedicine in one day[J/OL]. ArXiv, 2023: 2306.00890. https://arxiv.org/abs/2306.00890v1.

[12] Ghandi T, Pourreza H, Mahyar H. Deep learning approaches on image captioning: a review[J]. ACM Computing Surveys, 2023, 56(3): 1-39.

[13] Sun Wei, Wang Chunshan, Gu Jingqiu, et al. Veg-DenseCap: dense captioning model for vegetable leaf disease images[J]. Agronomy, 2023, 13(7): 1700.

[14] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J/OL]. ArXiv, 2020: 2006.11239. https://arxiv.org/abs/2006.11239v2.

[15] 袁志祥, 高永奇. InternDiffuseDet: 结合可变形卷积和扩散模型的目标检测方法[J]. 计算机工程与应用, 2024, 60(12): 203-215.
Yuan Zhixiang, Gao Yongqi. Intern diffuse det: object detection method combining deformable convolution and diffusion model[J]. Computer Engineering and Applications, 2024, 60(12): 203-215.

[16] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale[C]//2021 IEEE/CVF International Conference on Learning Representations (ICLR), Onlie, 2021: 11926.

[17] Devlin J, Chang Mingwei, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Minnesota, USA, 2019: 4171-4186.

[18] Li Yuxuan, Hou Qibin, Zheng Zhaohui, et al. Large selective kernel network for remote sensing object detection[C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2023: 16748-16759.

[19] 苟军年, 杜愫愫, 刘力. 基于改进掩膜区域卷积神经网络的输电线路绝缘子自爆检测[J]. 电工技术学报, 2023, 38(1): 47-59.
Gou Junnian, Du Susu, Liu Li. Transmission line insulator self-explosion detection based on improved mask region-convolutional neural network[J]. Transactions of China Electrotechnical Society, 2023, 38(1): 47-59.

[20] Chen Shoufa, Sun Peize, Song Yibing, et al. Diffusion Det: diffusion model for object detection[C]// 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2023: 19773-19786.

[21] 张烨, 李博涛, 尚景浩, 等. 基于多尺度卷积注意力机制的输电线路防振锤缺陷检测[J]. 电工技术学报, 2024, 39(11): 3522-3537.
Zhang Ye, Li Botao, Shang Jinghao, et al. Defect detection of transmission line damper based on multi-scale convolutional attention mechanism[J]. Transactions of China Electrotechnical Society, 2024, 39(11): 3522-3537.

[22] Li Junnan, Li Dongxu, Xiong Caiming, et al. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation[C]// Proceedings of the 39th International Conference on Machine Learning, Baltimore, USA, 2022: 12888-12900.

[23] 张中文, 吐松江·卡日, 张紫薇, 等. 基于双分支特征融合的电力设备缺陷文本挖掘方法[J]. 高压电器, 2024, 60(6): 188-196.
Zhang Zhongwen, Tusongjiang K, Zhang Ziwei, et al. Text mining method for power equipment defects based on two-branch feature fusion[J]. High Voltage Apparatus, 2024, 60(6): 188-196.

[24] Chen Xinlei, Fang Hao, Lin T Y, et al. Microsoft COCO captions: data collection and evaluation server[J/OL]. ArXiv, 2015: 1504.00325. https://arxiv.org/abs/1504.00325v2.

[25] Zhang Lin, Li Boqun, Cui Yang, et al. Research on improved YOLOv8 algorithm for insulator defect detection[J]. Journal of Real-Time Image Processing, 2024, 21(1): 22.

[26] Li L H, Zhang Pengchuan, Zhang Haotian, et al. Grounded language-image pre-training[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022: 10955-10965.

**作者简介**

陈 永 男，1979 年生，教授，博士生导师，研究方向为轨道交通异常检测。

E-mail：edukeylab@126.com（通信作者）

安卓奥博 男，1999 年生，硕士研究生，研究方向为计算机视觉。

E-mail：123028557@qq.com

# Semantic Description Method of High-Speed Railway Contact Net Insulator Defects Based on Diffusion Model Detection

*Chen Yong*[1,2]　*An Zhuoaobo*[1]　*Zhou Jianyu*[1]

（1. School of Electronic and Information Engineering　Lanzhou Jiaotong University　Lanzhou　730070　China
2. Engineering Research Center for Artificial Intelligence and Graphics & Image Processing
Lanzhou Jiaotong University　Lanzhou　730070　China）

**Abstract**　The catenary insulator is a critical component of the traction power supply system for high-speed railways. It not only provides electrical control insulation but also plays an essential role in supporting the catenary arm structure. Therefore, the operational safety of the insulator is directly related to the stability of the entire high-speed railway system. However, the detection of insulator defects is often subject to various interferences due to the complex and dynamic railway environment, resulting in low detection accuracy. Moreover, traditional detection methods generally only identify the presence of defects but fail to provide specific semantic descriptions of these defects. This limitation significantly hampers the efficiency of fault diagnosis and maintenance operations. To address these challenges, this paper proposes a defect description method for insulators based on a diffusion model. This method optimizes existing detection technologies in several ways, enabling the model to not only detect insulator defects more accurately but also generate detailed textual descriptions of these defects.

Firstly, we designed a large-kernel spatial selection feature extraction network. Compared to traditional feature extraction networks, this network captures the feature information of insulator defects through larger spatial convolution kernels, significantly enhancing the model's ability to extract insulator defect features. The model can accurately identify potential defects in the insulator, even in complex backgrounds. Secondly, we proposed a detection decoder with a fusion diffusion mechanism based on the diffusion model. This decoder generates noise boxes and uses inverse Bayesian diffusion to restore predictions of the insulator's true bounding box, significantly improving the model's resistance to background interference. This innovation allows the model to more effectively isolate background noise in complex environments, thereby improving the accuracy of defect detection. Finally, to address the limitations of traditional detection models in semantic description, we designed an encoder and decoder based on a cross-attention mechanism to achieve cross-modal mapping between images and text. By using the BLIP model driven by a text filtering mechanism, the model can generate corresponding textual descriptions of the defects based on the detection results. The functionality not only provides maintenance personnel with more intuitive references but also greatly enhances the efficiency of fault handling. Experimental results validate the effectiveness of our method. The proposed insulator defect detection model achieved the $mAP_{0.5}$ of 93.04% and the AR and F1-score of up to 83.22% and 82.91%. The BLEU achieved 83.51%, with CIDEr of 1.94, ROUGE-L of 81.59%, METEOR of 51.50%, and SPICE of 37.88%.

The experimental results lead to the following conclusions: (1) Utilizing a large-kernel spatial selection feature extraction network as the image encoder enhances the insulator defect detection network's ability to focus on key features, thereby improving the model's detection accuracy. (2) To address the issue of insulator defect detection being easily disturbed by complex background environments, a detection decoder with a fusion diffusion mechanism was designed. This decoder performs inverse Bayesian diffusion on the noise boxes generated by the decoder, restoring the prediction of the insulator's true bounding box. The model's ability to resist background interference reduces the loss of semantic information related to insulator defects, and enhances the accuracy of the predicted bounding boxes. (3) A cross-modal mapping module was designed to map the relationship between insulator image defect features and text features. The language modeling encoder outputs a textual description of the insulator defects, completing the detection task. Thus, the proposed model not only offers higher detection accuracy but also generates accurate and detailed semantic descriptions of the defects, meeting the actual needs for insulator defect detection and description.

**Keywords**：High speed railway catenary, insulator defect detection, defect image caption, diffusion model, cross-attention

（编辑　李　冰）