第 46 卷第 1 期
Vol. 46 No. 1 2025

青岛理工大学学报
Journal of Qingdao University of Technology

基于深度神经网络的钢结构焊接施工风险识别方法研究

王群力，张　凯*，刘丕养
（青岛理工大学 土木工程学院，青岛 266525）

摘　要：近年来随着社会经济和建筑行业的快速发展，安全问题日渐凸显，工人死亡和受伤事件屡见不鲜。钢结构焊接作为建筑行业施工中不可缺少的环节之一，需要一种高效快速的方法来检测工人在钢结构焊接中的施工风险。基于深度神经网络方法，构建图像文字描述模型，对钢结构焊接施工现场监控视频中提取的图像进行文字生成，并基于生成的文字识别现场工人的安全穿戴风险。通过网络爬虫和施工现场拍照收集图片数据，对图片数据进行数据增强和标注，制作成数据集；构建图像文字描述机器学习模型，使用建立的数据集对模型进行训练和验证，结果表明模型训练和验证识别准确度分别达到 88%和 85%；在文字识别结果的基础上采用关键词识别的方法进行风险结果判定。施工现场应用结果表明模型识别效果良好，并做出了准确直观的判定。

关键词：安全识别；深度学习；神经网络；图像文字生成；钢结构焊接
中图分类号：TU714　　**文献标志码**：A　　**文章编号**：1673-4602(2025)01-0001-07

Research on the risk identification method of steel structure welding based on deep neural network

WANG Qunli，ZHANG Kai*，LIU Piyang
（School of Civil Engineering，Qingdao University of Technology，Qingdao 266525，China）

Abstract：In recent years，with the rapid development of social economy and construction industry，safety problems have become increasingly prominent，and workers' deaths and injuries are not uncommon. Steel structure welding is one of the indispensable links in construction industry，and an efficient and rapid method is needed to detect the construction risk of workers in the welding process. Based on the deep neural network method，this study constructs an image text description model，which generates texts from the images extracted from the surveillance videos of the steel structure welding site and identifies the safety wearing risks of workers on the site based on the generated texts. Image data is collected through web crawler and construction site photos，and the data is enhanced and labeled to make a data set. The machine learning of the image text description model is constructed and the model is trained and verified with the established data set. The results show that the recognition accuracy of the model training and verification reaches 88% and 85% respectively. On the basis of the text recognition results，the method of keyword recognition is used to judge the

收稿日期：2023-11-23
基金项目：山东省自然科学杰出青年基金(JQ201808)
作者简介：王群力(1998—　)，男，浙江绍兴人。硕士，研究方向为施工风险识别与深度学习。E-mail：1727367478@qq.com。
*通信作者：张　凯(1980—　)，男，四川南充人。博士，教授，主要从事油气田开发工程等方面研究。E-mail：zhangkai@qut.edu.cn。

risk results. The application results on the construction site show that the recognition effect of the model is good, and accurate and intuitive judgment is made by the model.

Key words: security identification; deep learning; neural network; image text generation; steel structure welding

近年来随着我国建筑行业的迅速发展,施工现场的安全问题日渐凸显,据2022年住房和城乡建设部发布的通报,全国共发生房屋市政工程生产安全事故689起、死亡794人。因此安全是每个工程都需重点监测和控制的问题,如何有效减少安全事故和死亡率是建筑行业亟待解决的问题。钢结构焊接是建筑施工中重要的环节之一,施工过程存在对工人的眼睛、皮肤以及呼吸道的潜在危害,甚至会造成一些严重延伸疾病,若不穿戴电焊安全防护用具,将严重危害到工人的安全与健康。因此需要一种高效快速的方法来检测工人在焊接过程中的施工风险。

深度学习中的图像识别和自然语言处理功能已逐步应用在建筑行业中,检测人和物的不安全行为和不安全状态,且效果显著。PARK等[1]通过背景删减法和方向梯度直方图(HOG)对视频帧中的人和安全帽进行检测,识别出未佩戴安全帽的工人并发出警报。KIM等[2]提出基于卷积神经网络的裂缝自动检测方法取代人工目视检测。LUO等[3]基于YOLO2目标检测算法对项目建设过程中进入危险工作区的人员进行准确识别并及时预警。FANG等[4]基于Faster R-CNN目标检测算法对施工人员安全帽穿戴情况进行智能检测。石凌晨[5]结合注意力机制,构建基于改进ResNet主干网络和Mask R-CNN网络完成了对视频图像中钢筋排布的识别。梅杰等[6]将语义分割和目标检测算法应用于建筑施工行业来识别模板的覆盖率。杨静等[7]对检测的特征图进行2倍上采样,提高了对工人安全帽的检测精度。

以往的研究大都基于目标检测算法进行研究,目标检测算法是一种判别式模型,仅专注于特定目标的识别和定位,而增加了自然语言处理的图像文字生成模型可以提供更加全面的图像认知,对目标的识别包括描述目标图像的特征、结构、语义情景等信息,可以提供更加直观的图像信息。因此本文提出一种基于深度神经网络的钢结构焊接施工风险识别模型,对施工现场中的工人钢结构电焊场景进行风险识别。

1 数据集构建

1.1 数据获取

样本数据主要来源于3个途径:①通过爬虫代码从互联网中爬取工人电焊场景图片;②通过去工地现场实地拍摄正在从事焊接工作的工人;③网上和实际的图片结合,此途径更具有参考价值。图片主要包括佩戴安全帽和不配安全帽、佩戴电焊面罩和不佩戴电焊面罩、佩戴电焊手套和不佩戴电焊手套、佩戴护目镜和不佩戴护目镜4种情况。

1.2 数据增强和标注

对获取的数据进行多种方式的数据增强,增强方式包括对图片任意角度旋转、翻转,对图片色调、饱和度、亮度调整,对样本数量进行扩容。数据增强后的数据量和多样性可以使模型获得更好的性能,提高模型的鲁棒性,能够学习到更多的特征,以应对现场施工环境的复杂多样性。

数据增强完毕后对图像数据进行标注,总共为4处,分别为安全帽、电焊面罩、电焊手套和护目镜。未佩戴安全帽标注为头,未佩戴电焊手套标注为手,未佩戴电焊面罩标注为脸,如图1所示。用5个意义相近但表达方式不同的单词、句子对数据集中的每一张图片进行文字注释,并把生成的注释文件保存为.json格式文件。

图1 图像数据标注

2　CNN＋LSTM 框架模型构建

本文结合 CNN 与 LSTM 网络混合模型，并加入注意力机制，提取图像特征，根据不同图像区域的重要性生成图像描述。CNN＋LSTM 结构被广泛应用于多标签分类任务中[8]，但这些任务大都在通用领域中使用，而未在建筑施工行业中得到有效应用，而且在通用领域中只是用于识别而未能将结果进一步利用，本文将该结构模型应用到钢结构焊接施工中，并通过分析结果用来判定施工过程是否存在风险。CNN＋LSTM 模型结构如图 2 所示。

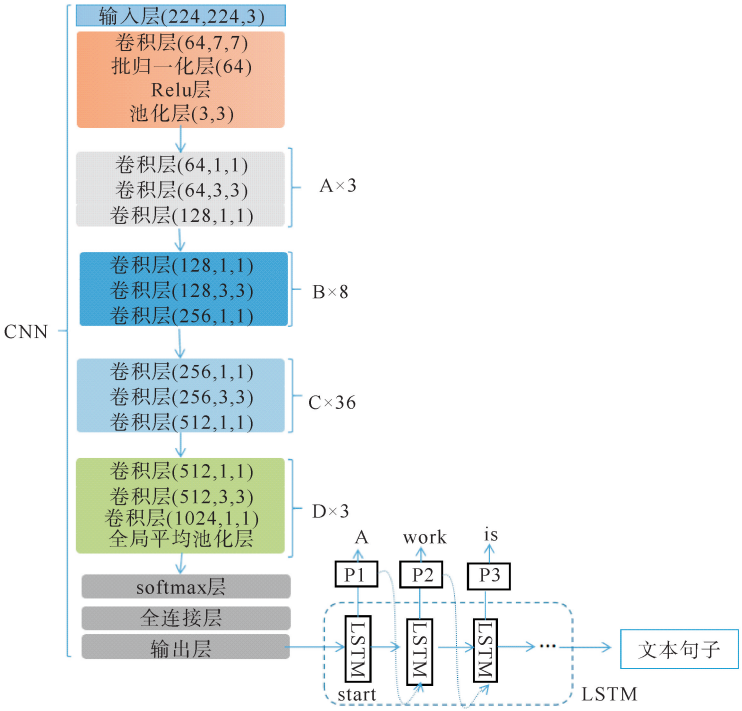


图 2　CNN＋LSTM 模型结构

2.1　卷积神经网络(CNN)

翻译模型主要由“编码器—解码器”两部分构成，编码器对输入的图像提取特征并输出，将编码器输出的特征输入到解码器中，最后生成文字描述内容。实现编码器功能的是卷积神经网络(CNN)，本文选用 Resnet152[9]网络架构代替普遍使用的 VGG-16 网络，Resnet152 拥有 152 层卷积层，特征提取的准确性和鲁棒性更好。Resnet152 核心包括输入层、卷积层、池化层和全连接层。输入层接收输入数据，其中 A、B、C、D 为 4 个残差块，A、B 和 C 3 个残差块中的卷积层之间都通过批归一化层和 Relu 层连接，D 残差块中的卷积层之间直接连接；最后在 D 层后连接一层全局平均池化层。卷积层中卷积核(滤波器)对输入图像进行滑动计算，可以提取图像中的局部特征，每个卷积核均有不同的权重参数，计算如式(1)所示。池化层连接在卷积层后，对输入特征图进行下采样，通过对特征图降维处理，可减少模型的计算量和参数量，提高模型训练的效率，减少训练时间。全连接层位于平均池化层之后，在全连接层中，神经元与前一层的所有神经元都有连接，每个输入特征与每个神经元权重相乘并进行累加，也就是将之前学到的特征值组合到一起，然后经过激活函数得到输出，其主要作用是将池化后的特征转换成一维的特征向量。

$$Y=a\left(\sum_{i=1}^{E}X*W_i+F\right) \tag{1}$$

式中：Y 为卷积层的输出；a 表示激活函数；$*$ 表示卷积操作；X 为输入的特征图；E 为输入特征图的通道数；W_i 为第 i 个卷积核；F 为偏执项。

2.2　长短期记忆网络(LSTM)

实现解码器功能的为长短期记忆网络。长短期记忆网络中的门控单元拥有遗忘机制，设有遗忘门 f_t、输入门 i_t 和输出门 h_t，可以得到输入序列中的时序关系，并生成相应的图像描述。遗忘门决定单元状态需要遗忘的信息，遗忘门的值在 0～1，接近 0 表示遗忘，接近 1 表示继续传递；输入门决定单元状态需要记忆的信息，对数据信息进行强调，通过遗忘门和输入门来更新当前时刻的单元状态。当前时刻的单元状态、当前时刻输入和上一时刻输出共同决定当前时刻的输出。LSTM 中的内存单元和门的具体计算见式(2)—(4)。

$$f_t=\sigma(\boldsymbol{W}_f\cdot x_t+\boldsymbol{U}_f\cdot h_{t-1}+\boldsymbol{b}_f) \tag{2}$$

$$i_t=\sigma(\boldsymbol{W}_i x_t+\boldsymbol{U}_i h_{t-1}+\boldsymbol{b}_i) \tag{3}$$

$$h_t=o_t\circ\tanh(C_t) \tag{4}$$

式中：$\boldsymbol{W}$、$\boldsymbol{U}$ 为各个门控的权重矩阵；$\boldsymbol{b}_f$、$\boldsymbol{b}_i$ 分别为遗忘门和输入门的偏执项；x_t 为当前时刻的输入；h_{t-1} 为上一时刻的输出；C_t 为当前时刻的单元状态；∘为 Hadamard 乘积；σ 为 sigmoid 激活函数。

2.3 注意力机制

采用注意力机制[10]对图像特征和 LSTM 中的隐藏状态进行注意力权重计算，然后对图像特征值加权求和，生成与当前时刻词语相关的图像区域，使模型只关注此时刻目标区域的内容，这样可以提供更准确和有关联性的描述，软注意力机制更加便捷，且更符合人类视觉特性，因此使用软注意力机制与解码器配合，对编码器输出特征图和当前时刻解码器输出门 h_t 进行比较计算。当需要处理特征矩阵 $\boldsymbol{A}$ 时，该注意力机制对矩阵 $\boldsymbol{A}$ 中每个元素赋不同分值，其计算见式(5)、式(6)。

$$S(p,k_i)=\frac{\exp(\mathrm{sim}(q,k_i))}{\sum_{m=1}^{n}\exp(\mathrm{sim}(p,k_i))} \tag{5}$$

$$A(p,(k_i,v_i))=\sum_{i=1}^{n}S(p,k_i)\cdot v_i \tag{6}$$

式中：S 为每个元素的权重分数；A 为注意力机制计算得到的最终注意力值；p 为隐藏状态；k_i 为提取的图像特征；v_i 为与图像特征对应的区域特征文本信息；q 为索引变量，用于注意力得分求和和归一化操作，其中每一个元素的得分传入 Softmax 层归一化获得相应权重；sim 为注意力机制的打分函数。

2.4 模型评价指标

2.4.1 交叉熵损失函数

交叉熵损失函数[11]是模型预测结果和真实值之间差异的一种损失函数，损失函数值越小，说明结果越接近真实值，模型预测效果越好，损失函数计算见式(7)。

$$L=-\sum(T\times\log\xi) \tag{7}$$

式中：L 为损失函数；T 为真实标签；ξ 为模型预测结果。

2.4.2 准确度

准确度[12]用于衡量模型预测过程中预测正确的样本比例，也反映了模型的性能，具体计算见式(8)。

$$P=\frac{T}{T+F} \tag{8}$$

式中：P 为准确度；T 为模型预测正确的样本数；F 为模型预测错误的样本数。

2.4.3 字幕评价指标

本文采用 Bleu[13]指标对字幕进行评价，该指标的设计就是用来评价原翻译文本和模型翻译结果之间的相似度的。Bleu 评价指标通过对需要翻译和参考翻译之间的 N-gram[12]匹配度进行计算，得到的一个综合的分数，以此来判断输出句子结果的好坏，Bleu 分数的取值范围为 0～1，其值越接近于 1 表示机器翻译越准确。根据 N 的阶数，Bleu 有 4 种评价指标，本文取 $N=4$，故采用 Bleu4 作为评价指标，Bleu4 计算见式(9)。

$$U=V\cdot\exp(\sum_{n=1}^{N}h_n\log(p_n)) \tag{9}$$

式中：U 为模型输出的评价指标值；V 为惩罚因子，用来调整候选句子与参考句子之间的长度差异；N 为 N-gram[14]的最大阶数，其值取 4；h_n 为 N-gram 的权重，取值为 $1/N$；p_n 为候选句子中 N-gram 精确匹配数与参考句子的数量之比。

其中，

$$V=\begin{cases}1 & c>r\\ \mathrm{e}^{1-r/c} & c\leqslant r\end{cases} \tag{10}$$

$$p_n=\frac{\sum_i\sum_k\min(d_k(e_i),\max_{i\in n}d_k(v_i))}{\sum_i\sum_k d_k(e_i)} \tag{11}$$

式中：c 为候选句子的长度；r 为参考句子中与候选句子相对最短的，如此可以避免模型翻译的句子过短；d_k 为长度为 k 的 N-gram 在参考句子中出现的次数；e_i 为模型生成的第 i 个文本；v_i 为参考句子的第 i 个文本。

3　模型训练与应用

3.1　模型识别的主要步骤

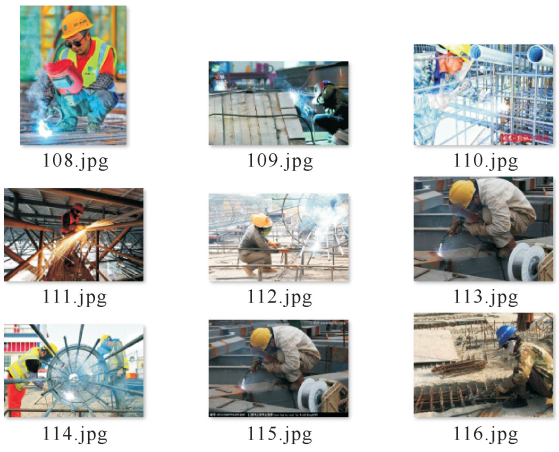
1）通过施工现场拍摄和网络爬虫获取图片数据，如图 3(a)所示。

2）对图片数据进行预处理和图片标注，并制作对应图片的注释文件，如图 3(b)所示。

3）将制作完毕的数据集按照 8∶2 的比例随机分成训练集 train 和验证集 val。

4）构建 CNN＋LSTM 图像文字生成模型，将数据集输入模型，然后根据结果进行模型参数调优，并使用 Adam 优化器[15]在训练过程中自动调整模型的学习率，提高训练效果。

5）采用关键词过滤的方法，进一步判定模型输出的结果是否存在安全风险。



(a) 图片集

(b) .json注释文件

图 3　图片与注释文件

3.2　训练和验证结果分析

将前期准备完毕的数据输入到模型中，通过卷积神经网络提取特征，再将特征向量输入到长短期记忆网络中，最后生成文字描述结果。如图 4 所示，模型收敛速度较快，最终训练的准确度为 88%，验证的准确度为 85%左右，两者的拟合效果较好。如图 5 所示，模型的损失值前期快速下降，最终训练的损失值稳定在 2.25，验证的损失值为 2.3，效果良好，可以准确识别出图片中的钢结构焊接工人的安全穿戴情况，识别良好。如图 6 所示，Bleu4 分数最终稳定值接近 0.5，效果较好，但仍有可以提升的空间。

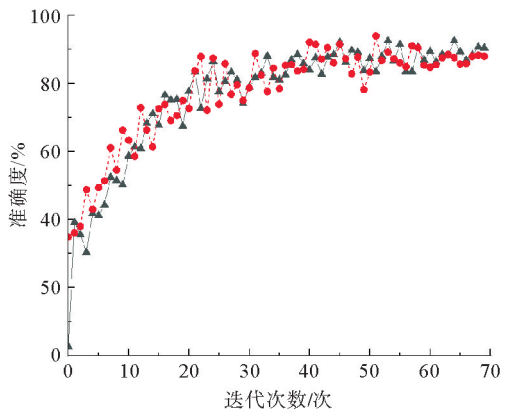


图 4　训练阶段和验证阶段准确度

训练准确度；验证准确度

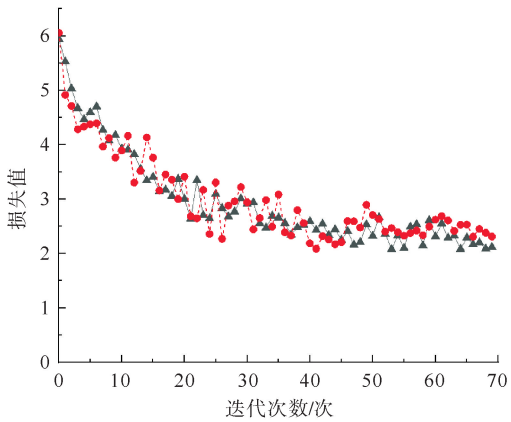


图 5　训练阶段和验证阶段损失值

训练损失值；验证损失值

3.3 结果判定

文字识别完毕后，对识别的文字内容做出有无风险的判定，识别出任何一种安全防护用具未佩戴均判定为有风险，否则安全。判定结果能够更加准确直观地理解和判断图像的内容并可做出相应的响应。

具体判定过程：首先构建一个敏感词的列表，然后通过对生成的句子文本分词，最后进行敏感词匹配。模型输出的文本中包含有关“未穿戴”或“未佩戴”等字样的敏感词，就识别为存在安全风险。

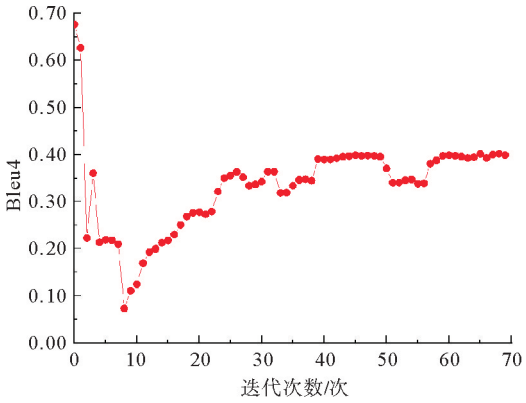


图 6 Bleu4 分数

3.4 模型应用实例

将本模型应用于青岛市黄岛区某一建筑工地施工现场。选取该工地监控中钢结构焊接工程的一段视频，从该视频中截取若干视频帧作为模型应用对象进行分析。将视频帧输入模型，输出结果如图 7、图 8 所示，该模型在施工现场应用中，识别判断准确，可以有效识别出是否存在风险，表明模型识别效果良好。

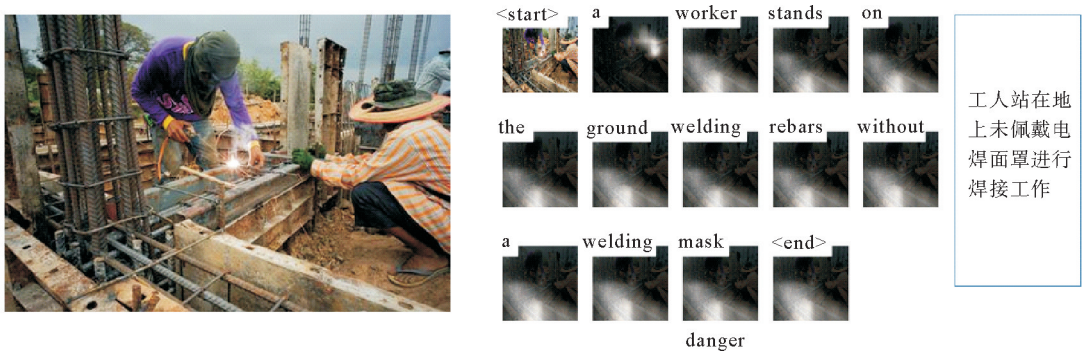


图 7 未佩戴面罩识别“存在风险”

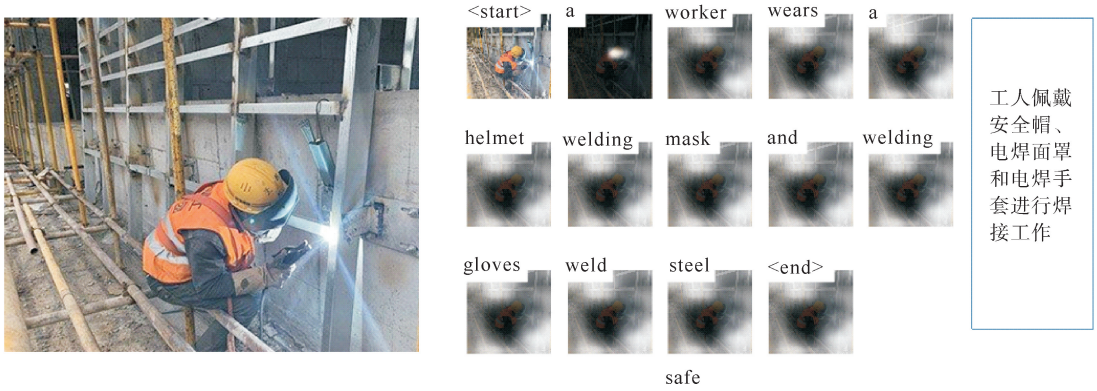


图 8 正确佩戴护具识别“安全”

4 结束语

本文通过网络爬虫和现场拍摄的图片数据，经过数据处理工具制作成数据集，构建图像文字描述模型，将数据集输入模型中训练，得到较高的准确率和较低的损失值，对钢结构焊接施工场景的安全穿戴风险识别效果良好。对模型识别出的结果内容进行判定，在生成字幕的同时对图像内容进行有无风险的判定，对现场实例的应用识别准确，证明模型效果良好。未来需通过优化网络结构和增加使用场景对所使用

的方法进一步改进和验证。

本文使用深度学习方法，对钢结构焊接施工场景中的防护用具穿戴风险识别进行了方法研究，是对当前建筑施工行业安全问题的一次积极探索和研究，对建筑行业的健康安全发展具有积极意义，对当下建筑业中存在的工人安全问题具有重要意义。

参考文献(References)：

[1] PARK M W，ELSAFTY N，ZHU Z H. Hardhat-wearing detection for enhancing on-site safety of construction workers[J]. Journal of Construction Engineering and Management，2015，141(9)：04015024.

[2] KIM B，CHO S. Automated vision-based detection of cracks on concrete surfaces using a deep learning technique[J]. Sensors，2018，18(10)：3452.

[3] LUO H B，LIU J J，FANG W L，et al. Real-time smart video surveillance to manage safety：A case study of a transport mega-project[J]. Advanced Engineering Informatics，2020，45：101100.

[4] FANG Q，LI H，LUO X C，et al. Detecting non-hardhat-use by a deep learning method from far-field surveillance videos[J]. Automation in Construction，2018，85：1-9.

[5] 石凌晨. 基于计算机视觉的钢筋排布检测[D]. 重庆：重庆大学，2022.
SHI Lingchen. Rebar arrangement detection based on computer vision[D]. Chongqing：Chongqing University，2022.

[6] 梅杰，李庆斌，陈文夫，等. 基于目标检测模型的混凝土坯层覆盖间歇时间超时预警[J]. 清华大学学报(自然科学版)，2021，61(7)：688-693.
MEI Jie，LI Qingbin，CHEN Wenfu，et al. Over warning of concrete pouring interval based on object detection model[J]. Journal of Tsinghua University(Science and Technology)，2021，61(7)：688-693.

[7] 杨静，张育飞，毛晓琦. 施工作业面安全帽的深度学习检测方法[J]. 计算机应用，2020，40(S2)：178-182.
YANG Jing，ZHANG Yufei，MAO Xiaoqi. Deep learning detection method of safety helmet on construction working surface[J]. Journal of Computer Applications，2020，40(S2)：178-182.

[8] BHALEKAR M. D-CNN：A new model for generating image captions with text extraction using deep learning for visually challenged individuals[J]. Engineering，Technology & Applied Science Research，2022，12(2)：8366-8373.

[9] HE K M，ZHANG X Y，REN S Q，et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2016：770-778.

[10] XU K，BA J，KIROS R，et al. Show，attend and tell：Neural image caption generation with visual attention[J]. arXiv Preprint arXiv，2015：1502.03044. DOI：10.48550/arXiv.1502.03044.

[11] QU Z，MEI J，LIU L，et al. Crack detection of concrete pavement with cross-entropy loss function and improved VGG16 network model[J]. IEEE Access，2020，8：54564-54573.

[12] CHEN Y D，ZHAO C，YU Z，et al. On the relation between sensitivity and accuracy in in-context learning[J]. arXiv Preprint arXiv，2022：2209.07661. DOI：10.48550/arXiv.2209.07661.

[13] PAPINENI K，ROUKOS S，WARD T，et al. Bleu：A method for automatic evaluation of machine translation[C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)，Philadelphia，2002：311-318.

[14] BROWN P F，PIETRA V J D，SOUZA P V D，et al. Class-based *N*-gram models of natural language[J]. Computational Linguistics，1992，18(4)：467-479.

[15] KINGMA D P，BA J. Adam：A method for stochastic optimization[J]. arXiv Preprint arXiv，2014：1412.6980. DOI：10.48550/arXiv.1412.6980.

（责任编辑　赵金环；英文校审　程文华）