第 59 卷 第 2 期/2022 年 1 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于注意力机制的鞋型识别算法

张家钧，唐云祁*，杨智雄，耿鹏志

中国人民公安大学侦查学院，北京 100038

摘要 根据现场遗留鞋印推断出作案人所穿鞋型，再从周围监控视频中搜索嫌疑鞋型已成为公安机关侦破案件的重要技战法。针对该技战法完全依赖人工筛查、受主观影响大、易造成漏检等问题，提出了一种基于注意力机制的鞋型识别算法。首先，建立了贴近公安刑侦实战、样本容量为300的多背景监控鞋型数据集。然后，提出了一种注意力机制模型，以增强残差网络(ResNet50)对鞋子重要特征的提取能力。最后，对比了选取不同特征层输出作为鞋子特征及不同卷积特征聚合方法对识别精度的影响。为了增强模型的泛化能力，在损失函数中加入Label Smoothing。在多背景数据集上的实验结果表明，本算法的Rank-1、平均精度均值分别达到74.32%和56.97%。

关键词 机器视觉；深度学习；鞋型识别；注意力机制；特征聚合

中图分类号 TP391.4 **文献标志码** A **doi**：10.3788/LOP202259.0215004

Shoe Type Recognition Algorithm Based on Attention Mechanism

Zhang Jiajun, Tang Yunqi*, Yang Zhixiong, Geng Pengzhi

School of Investigation, People's Public Security University of China, Beijing 100038, *China*

Abstract It has become an important technique and tactics for the public security organs to infer the type of shoes worn by the perpetrators according to the shoe prints left at the scene, and then search the suspected type of shoes in the surrounding surveillance video. This technique is completely dependent on manual screening, which is greatly affected by subjective factors and easily leads to problems such as missed detection. To solve this problem, this paper proposes a shoe type recognition algorithm based on attention mechanism. First, close to the actual combat of public security criminal investigation, a multi background monitoring shoe data set with sample size of 300 is established. Then, an attention mechanism model is proposed to enhance the ability of the residual network (ResNet50) to extract important features of shoes. Finally, the effects of selecting the output of different feature layers as shoe features and different convolution feature aggregation methods on the recognition accuracy are compared. In order to enhance the generalization ability of the model, label smoothing is added to the loss function. The experimental results on the multi background data set show that the Rank-1 and mean average precision of the algorithm are 74.32% and 56.97%, respectively.

Key words machine vision; deep learning; shoe type recognition; attention mechanism; features aggregation

1 引 言

随着监控摄像头的普及，“足迹+监控”已成为公安侦查机关常用的技战法。该技战法依据作案人遗留在现场的鞋印足迹，利用“全国公安机关鞋样本数据库应用系统”[1]检索得到作案人所穿鞋型，进而在监控视频中追踪犯罪嫌疑人的影像。该技战法可实现由犯罪现场足迹到监控视频的跨模态

收稿日期：2021-01-12；**修回日期**：2021-02-10；**录用日期**：2021-03-16

基金项目：公安部技术研究计划(2020JSYJC21)、中央高校基本科研业务费(2021JKF203)

通信作者：*tangyunqi@ppsuc.edu.cn

追踪溯源，具有重要的实战应用价值[2]。2015年，在某市的案件现场中，技术人员将一枚作案人遗留在现场的鞋印输入“鞋样本数据库”中，得到该鞋印对应的鞋型，之后专案组利用该鞋型的外观信息在案发现场周围监控视频中成功锁定了犯罪嫌疑人[3]。2016年，技术人员将某案件现场提取的一枚残缺鞋印输入“鞋样本数据库”中，经过人工仔细复核后确定该残缺鞋印对应的鞋型，技术人员基于该鞋样的外观信息在案发现场周边的监控视频中进行检索比对，成功锁定了穿着该鞋在案发时间进出过现场的嫌疑人[3]。

“鞋印＋监控”技战法在公安刑侦破案中发挥着重要作用，但目前在监控中搜索嫌疑鞋型的工作全部由人工观看实现，极其消耗人力物力。我国公安机关刑侦办案时间紧、任务重，民警往往需要连续长期观看监控视频，易受主观因素的影响造成漏查。因此，亟需一种自动识别监控中鞋型的算法，以提高公安机关“鞋印＋监控”技战法的侦查效率。深度神经网络能学习到更深层次的特征[4]，具有较好的特征表达能力[5]。随着深度学习的发展，图像识别工作得到了很大的进展，针对鞋型识别的研究也逐渐得到了人们的关注。Huang等[6]收集整理了包含17151张鞋子图像的数据集，每张图像都标注了10个鞋子的部件属性，并提出了一种先检测鞋部件再进行鞋型检索的方法。Zhan等[7]提出了一个基于语义属性的跨域鞋型检索系统，基于鞋子的语义信息提出了鞋型检索卷积神经网络(SHOE-CNN)，可提取鞋子的三层特征并融合多层特征作为图像特征的表示向量，其检索精度比预训练的AlexNet模型高40%以上，在一定程度上解决了跨域鞋型检索中的问题。杨孟京等[8]制作了包含50双鞋共160231张低清晰度的鞋型数据集，并根据DeepID重新设计了网络结构。该网络由两层卷积层、两层池化层、两层全连接层组成，同时探究了全连接层的输出数量以及不同网络深度对实验结果的影响，最终的分类精度可达到96.06%。陈前等[9]针对以往鞋型检索方法只提取粗粒度特征、缺少关键语义特征、难以区分鞋型之间细小差异的问题，提出了一种结合部件检测和语义网络的细粒度鞋型检索方法，先训练出鞋型部件检测模型，再进行语义属性训练，以提取图像的特征向量，该方法的分类精度比已有检索精度较高的方法提升了约6%。目前，针对鞋型识别的研究较少，相比人脸识别等工作，监控视频中行人所穿鞋子的影像存在运动模糊、分辨率低、可利用特征少等问题，因此，如何在有限分辨率的鞋子图像中提取到更具判别性的特征是目前需要解决的一个难题。

针对上述问题，本文提出了一种基于注意力机制的监控视频鞋型识别算法。首先，搭建监控摄像头采集实验数据，贴近公安刑侦实战建立了样本容量为300的多背景监控鞋型数据集。然后，基于该数据集提出了一种注意力机制模型，在残差网络(ResNet50)[10]中融合注意力机制模块增强网络对鞋子重要特征的提取能力。最后，对比了选取全连接层和卷积层的输出作为鞋子特征[11]及不同卷积特征聚合方法对识别精度的影响[12]。为了增强模型的泛化能力，在损失函数中加入Label Smoothing[13]，以缩小正确标签和错误标签之间的差距，避免网络训练时出现过拟合等问题。

2　基于注意力机制的鞋型识别算法

2.1　网络框架

ResNet通过引入残差结构有效解决了深层网络普遍存在的梯度消失、性能退化等问题，简化了深层网络的训练难度。图1为ResNet50中的残差结构，其中，ReLU为修正线性单元，Conv为卷积层。假设残差结构的输出为$G(x)=F(x)+x$，其中，x为上一层的输出，残差结构通过学习得到$F(x)$。若网络发生梯度消失等问题，$F(x)$为0，此时$G(x)=x$，能有效避免网络性能退化等问题。

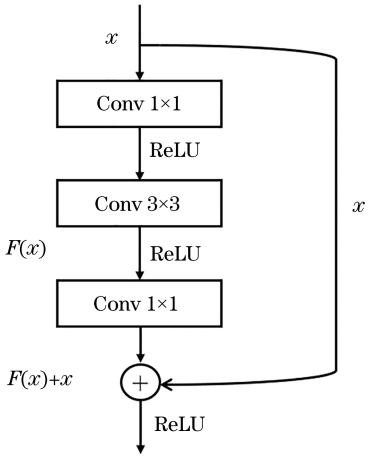

图1　ResNet50的残差结构

Fig. 1　Residual structure of the ResNet50

2.2　注意力机制模型

注意力机制的本质在于使计算机模仿人类观

察事物的方式，其核心目的是从众多信息中选择出对当前任务目标更有用的信息。注意力机制分为通道注意力机制和空间注意力机制。在神经网络中，输入的每个通道并不全部包含任务目标信息，每个通道对任务的贡献程度也不同，通道注意力机制则重点关注通道上的信息，其任务目标是学习获得每个通道的重要性程度。空间注意力机制重点关注重要特征在图像上的位置信息，即关注网络学习到的特征“在哪里”。

受距离、监控摄像机分辨率、天气等因素的影响，监控视频中的鞋子图像清晰度较低，一些重要的细节特征易被神经网络忽略。注意力机制可以帮助网络关注到这些细节特征，因此，基于建立的多背景监控鞋型数据集提出了一种新的注意力机制模型，该模型的结构如图 2 所示。首先，选用全局平均池化（AvgPool）和最大池化（MaxPool）将上层输出在空间维度上进行压缩。平均池化是对特征图上任意位置的反馈，最大池化是对传播过程中响应最大地方的反馈。实验数据集由模糊鞋子和复杂背景组成，其中，鞋子占整张图像的面积较大，若单独使用平均池化或最大池化易造成重要信息的丢失。因此，设上一层的输出特征为 $\boldsymbol{X}\in\boldsymbol{R}^{c\times h\times w}$，其中，$c$、$h$、$w$ 分别为上层输出特征的通道数、高度和宽度，经平均池化和最大池化得到特征 $\boldsymbol{X}_{\mathrm{avg}}\in\boldsymbol{R}^{c\times 1\times 1}$ 和 $\boldsymbol{X}_{\max}\in\boldsymbol{R}^{c\times 1\times 1}$。为了更好地建立通道之间的关联度并减少参数量和计算量，在池化层之后连接两层全连接（FC）层进行降维和升维操作，在全连接层之间选取适用于深度神经网络隐含层的 ReLU 激活函数。第二层全连接层经过 Sigmoid 激活函数，获得每一特征通道的归一化权重（0～1 之间）。特征通过第一层全连接层降维到 $\boldsymbol{W}_0\in\boldsymbol{R}^{\frac{c}{r}\times 1\times 1}$，之后经过第二层全连接层升维到 $\boldsymbol{W}_1\in\boldsymbol{R}^{c\times 1\times 1}$，其中，$r$ 为降维比例。为了减小模型的参数量并保留更多的细节特征，将 r 设置为 16。将输出的注意力权重系数相加并与输入特征相乘得到最终的输出。注意力权重 $\boldsymbol{W}$ 可表示为

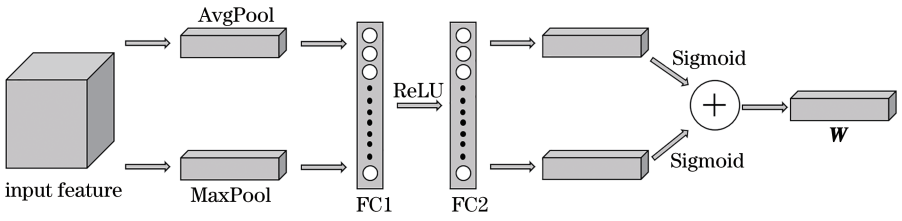

图 2　注意力机制模块的结构

Fig. 2　Structure of the attention mechanism module

$$\boldsymbol{W}=\sigma\left\{\boldsymbol{W}_1\left\{\delta\left\{\boldsymbol{W}_0\left[A(\boldsymbol{X})\right]\right\}\right\}\right\}+\sigma\left\{\boldsymbol{W}_1\left\{\delta\left\{\boldsymbol{W}_0\left[M(\boldsymbol{X})\right]\right\}\right\}\right\},\tag{1}$$

式中，$\boldsymbol{X}\in\boldsymbol{R}^{c\times h\times w}$ 为上一层的输出特征，A 为平均池化操作，M 为最大池化操作，$\boldsymbol{W}_0\in\boldsymbol{R}^{\frac{c}{r}\times 1\times 1}$、$\boldsymbol{W}_1\in\boldsymbol{R}^{c\times 1\times 1}$ 分别为经过第一层、第二层全连接层输出的特征，δ 为 ReLU 激活函数，σ 为 Sigmoid 激活函数。设通过注意力机制模型后的输出为 $\boldsymbol{X}_1$，可表示为

$$\boldsymbol{X}_1=\boldsymbol{W}\otimes\boldsymbol{X}。\tag{2}$$

图 3 为引入注意力机制模块后的残差结构，为提高算法的识别精度，使图像关键信息在网络中进行有效传递，帮助网络捕捉图像中的重要特征，将该注意力机制模块应用到 ResNet50 中的每一个残差结构中，具体算法流程如图 4 所示。

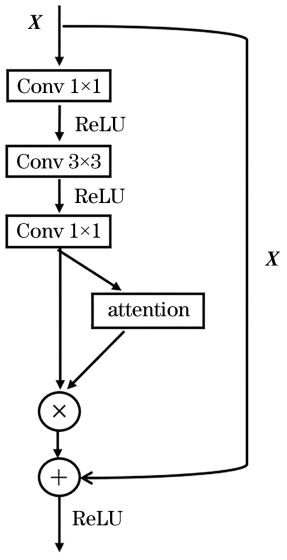

图 3　引入注意力模块后的残差结构

Fig. 3　Residual structure after introducing attention module

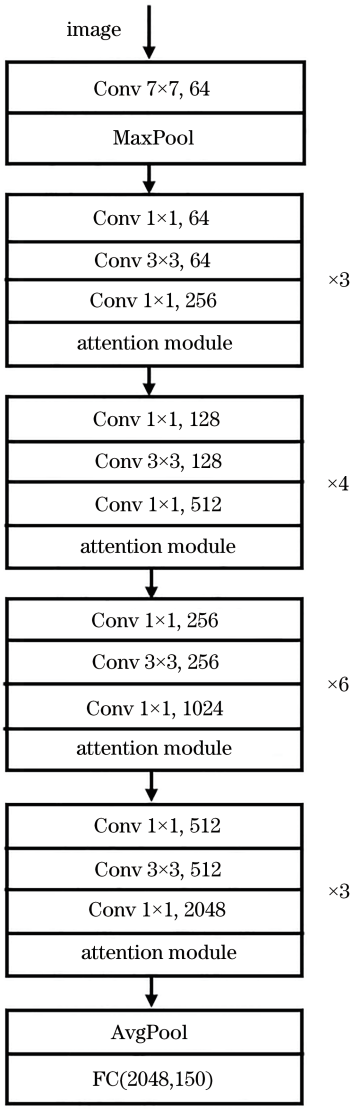

图 4　引入注意力模块后的算法流程图

Fig. 4　Algorithm flow chart after introducing attention module

2.3　特征层的选取及卷积特征聚合方法的对比

鞋型识别的关键在于如何更好地表达鞋型特征，在卷积神经网络中，卷积层的作用是提取特征，而全连接层综合了卷积层提取到的全部特征，是描述图像整体语义信息的深度全局特征，充当的是“分类器”的作用，缺少对细节信息的描述。因此，选择 ResNet50 中 Layer 4 和全连接层的输出作为鞋子特征，并探究 Layer 4 和全连接层对鞋型识别精度的影响。

用卷积层的输出作为鞋子特征时，要选择合适的卷积特征聚合方式处理卷积层的输出结果。最大池化是对传播过程中响应最大地方的反馈，可以更好地选择特征，保留更具辨识度的特征。平均池化是对特征图上任意位置的反馈，更加强调对整体特征进行下采样。针对上述两种卷积特征聚合方式的特点，尝试将最大池化和平均池化两种卷积特征聚合方法相结合，探究在保留更具辨识度特征的基础上保留更多细节特征的方法。结合最大池化和平均池化的方法可表示为

$$F_1=\left[M\left(\boldsymbol{X}_{\text{Layer4}}\right);A\left(\boldsymbol{X}_{\text{Layer4}}\right)\right]。\tag{3}$$

2.4　Label Smoothing

训练过拟合和模型缺乏泛化性是深度学习中常见的问题，为了进一步提升模型的泛化能力，在损失函数中加入了 Label Smoothing。分类网络最后一层输出是对类别预测的概率分数，传统的 one-hot 编码标签会激励网络对目标类别的预测概率趋近于 1，对非目标类别的预测概率趋近于 0，导致整个模型朝目标类别和非目标类别预测分数差值无限增大的方向学习，造成网络过拟合[14]。模型中的交叉熵损失函数可表示为

$$X_{\text{Loss}}=-\sum_{i=1}^{N}q_i\log\left(p_i\right),\tag{4}$$

$$q_i=\begin{cases}1,i\ \text{is target}\\0,i\ \text{is not target}\end{cases},\tag{5}$$

式中，N 为总类别数，i 为其中的某一类，q_i 为 i 的预测结果，p_i 为网络输出的置信度分数。

实验中，测试数据集的鞋子类别不会出现在训练数据集中，需要更加注意模型的泛化能力，避免发生训练过拟合等问题。因此，在交叉熵损失函数中加入 Label Smoothing，Label Smoothing 通过在输出中加入噪声，降低模型过拟合的风险，达到约束模型的目的[15]。加入 Label Smoothing 的交叉熵损失函数可表示为

$$X_{\text{Loss}}=-\sum_{i=1}^{N}y_i\log\left(p_i\right),\tag{6}$$

$$y_i=\begin{cases}1-\varepsilon,i\ \text{is target}\\\varepsilon/N,i\ \text{is not target}\end{cases},\tag{7}$$

式中，ε 为人为设置的超参数。

3　实验及结果分析

实验在 Linux 3.10.0 操作系统下进行，基于 PyTorch 深度学习框架平台展开研究。CPU 为 Intel(R) Xeon(R) CPU E5-2650 v4 2.20 GHz，显卡型号为 NVIDIA TITAN X (Pascal)，深度学习平台为 PyTorch 1.2.0，编译环境为 Python 3.5.6。

3.1　数据集的构建及预处理

为贴近公安刑侦实战，在足迹实验室模拟道路正侧和两侧尽头搭建三处海康威视监控摄像头，同时采用5种颜色深浅不一、复杂程度不同的背景铺设地面。对400双鞋进行监控视频的录制，视频的帧率和分辨率分别为50 frame/s和1920 pixel×1080 pixel。为使数据贴近公安刑侦实战，利用室外阳光及室内灯光控制数据集的整体亮暗程度，200名志愿者按正常步幅在模拟道路行走两个来回。搭建的设备及行走路线如图5所示。

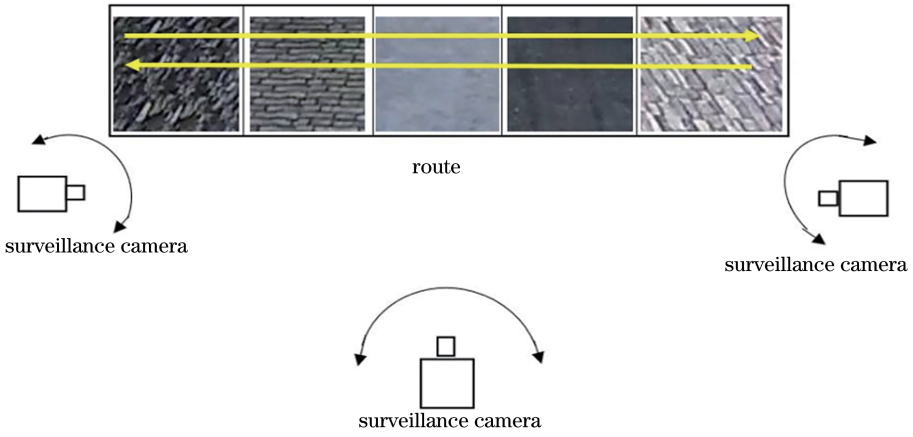

图5　搭建的设备及行走路线

Fig. 5　Equipment construction and walking route

为获得监控视频中行人所穿鞋子的图像，用HMSTranscoder视频转换器对录制视频进行解码并对解码视频进行分帧处理；为了防止图像重复度过高，每隔2帧截取1张图像并模拟目标检测算法[16]手动截取图像中的鞋子，截取的鞋图分辨率为120 pixel×120 pixel。截取示意图如图6所示。最后，去除100双重复的鞋子，选取150双鞋（每双鞋包含200张清晰度较低的侧面和正面鞋子图像），共计30000张鞋图作为训练数据集；选取150双鞋图（每双测试样本包含1张侧面和正面鞋子图像）作为测试数据集；选取5000张鞋图作为混淆样本与400张样本鞋构成鞋样数据库。

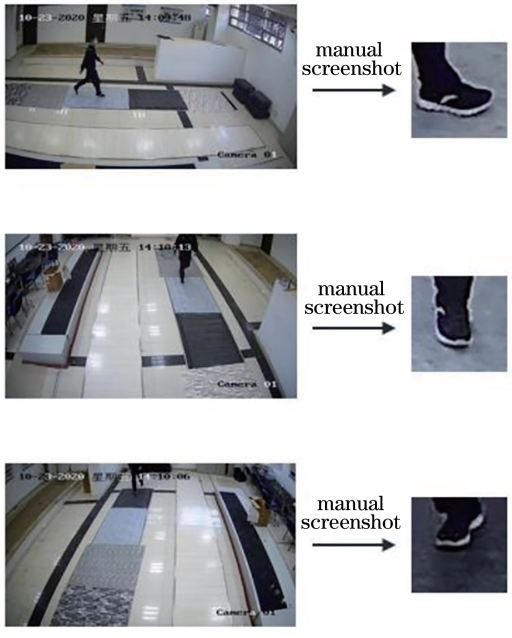

图6　鞋子图像的截取示意图

Fig. 6　Schematic diagram of capturing shoe images

刑侦实战中，监控视频中的鞋子会以各种角度出现，进而对鞋型识别工作造成较大干扰。针对该问题，通过手动标注鞋尖和鞋跟两处关键点，利用两点间斜率沿两点之间的中点旋转对鞋子进行校正。对测试数据集和鞋样数据库中的数据进行预处理，校正示意图如图7所示。

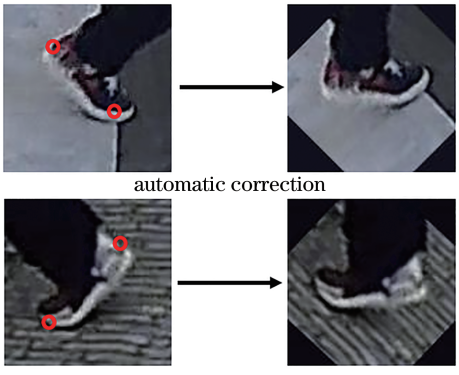

图7　鞋子图像的校正示意图

Fig. 7　Schematic diagram of correcting shoe images

3.2　评价指标

为了评估本算法的性能，选取的评价指标有平

均精度均值(mAP)、累计匹配特性(CMC)曲线和Rank-1指标。mAP可以反映识别出的鞋子在数据库中所有正确图像排在列表前的程度,能全面反映模型性能的好坏,mAP指对平均精度(AP)求和之后再求平均。CMC曲线是计算Rank-k的击中率,形成Rank-accuracy曲线。Rank-k表示算法返回的排序中,正确匹配出现在前k位的概率,由于实验需要准确识别出某一特定嫌疑鞋子,因此用Rank-1作为算法的评价指标。相关评价指标可表示为

$$X_{\text{Precision}}=\frac{X_{\text{TP}}}{X_{\text{TP}}+X_{\text{FP}}},\tag{8}$$

$$X_{\text{AP}}=\frac{\sum X_{\text{Precision}}}{X_{\text{images}}},\tag{9}$$

$$X_{\text{mAP}}=\frac{\sum_{i=0}^{N}X_i^{\text{AP}}}{N},\tag{10}$$

式中,X_{TP}为预测正确的正样本数,X_{FP}为预测错误的正样本数,X_{image}为图像数量。

3.3　实验结果及分析

3.3.1　注意力机制模型的对比分析

为了验证本算法中注意力模型的有效性,将目前使用较多的挤压与激励网络(SENet)[17]、卷积模块注意力机制(CBAM)[18]和高效通道注意力网络(ECA-Net)[19]注意力机制与ResNet50结合后与本算法进行对比实验,结果如表1和图8所示。可以发现,本算法中的注意力机制模型识别精度优于ResNet50,且相比常用的SENet、CBAM和ECA-Net注意力机制的识别精度也有明显提升,这表明本算法中的注意力机制模型更适用于实验数据集,可提升网络对监控视频中鞋子特征的提取能力。

表1　不同注意力机制模型的识别精度

Table 1　Recognition accuracy of different attention mechanism models　　unit:%

Model	Rank-1	mAP
ResNet50	61.09	46.34
ResNet50+SENet	59.14	43.79
ResNet50+CBAM	62.65	45.86
ResNet50+ECA-Net	59.14	45.71
Ours	**65.37**	**47.99**

3.3.2　热力图可视化

通过可视化分析,可以明确了解网络重点学习的是哪些特征,并针对可视化结果对网络模型进行调整。为了更直观地显示出网络模型关注的信息,引入了热力图可视化(Grad-CAM)[20]和Guided Grad-CAM对鞋子区域进行可视化分析,结果如图9所示。对比人工标注的目标信息和可视化中网络学习到的信息,可以明显发现,本算法中的注意力机制模型能准确学习到鞋子的重要特征并排除复杂背景的干扰。

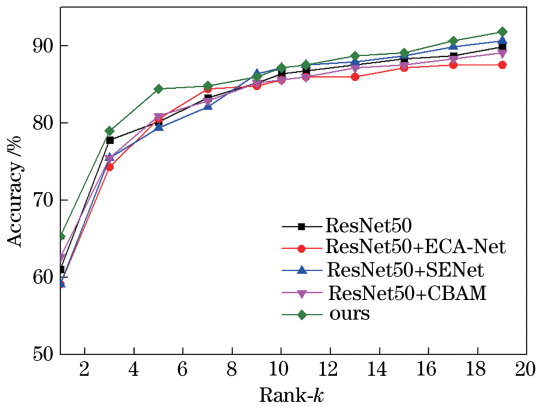

图8　不同注意力机制模型的识别精度曲线

Fig. 8　Recognition accuracy curves of different attention mechanism models

3.3.3　不同特征层及卷积特征聚合方法的对比分析

为了探究神经网络中不同特征层对识别精度的影响,对比了选取全连接层和Layer 4的输出作为鞋子特征对识别精度的影响。针对实验建立的数据集,在Layer 4上对比了最大池化、平局池化以及结合最大池化和平均池化的三种卷积特征聚合方法对识别精度的影响,结果如表2和图10所示。可以发现,选取Layer 4的输出作为鞋子特征的识别精度明显高于选取全连接层输出作为鞋子的特征,原因是全连接层综合了卷积层提取到的全部特征,是对语义信息进行描述的深度特征,缺少细节特征信息。而卷积层经过卷积特征聚合后,只保留了鞋子的细节特征,有效减少了背景等无用特征的影响。此外,最大池化作为卷积特征聚合方法的识别精度最高,原因是实验使用的数据集中,识别鞋型依靠的最重要特征是鞋子的logo,最大池化的作用是对传播过程中响应最大地方的反馈,减少了背景等无用特征的干扰,而平局池化以及两种卷积方式相结合的方法都保留了部分无用特征,对识别精度造成了一定的影响。因此,最大池化是最适用于本数据集的卷积特征聚合方法。

3.3.4　Label Smoothing

为了增强模型的泛化能力,避免训练过拟合风险,在交叉熵损失函数中加入了Label Smoothing。

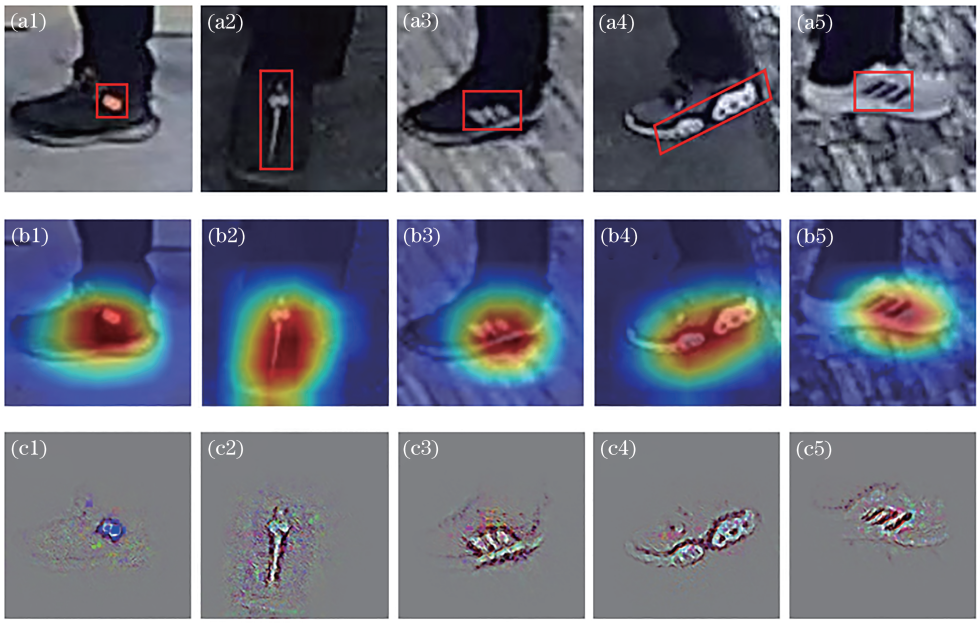

图 9　不同模型的可视化结果。(a)原始图像;(b) Grad-CAM;(c) Guided Grad-CAM

Fig. 9　Visualization results of different models. (a) Original image; (b) Grad-CAM; (c) Guided Grad-CAM

表 2　不同特征层和卷积特征融合方法对比

Table 2　Comparison of different feature layers and convolution feature aggregation methods　　unit:%

Layer and feature aggregation method	Rank-1	mAP
FC	65.37	47.99
Layer 4+AvgPool	69.65	52.59
Layer 4+AvgPool+MaxPool	73.15	54.44
Layer 4+MaxPool	**73.93**	**54.87**

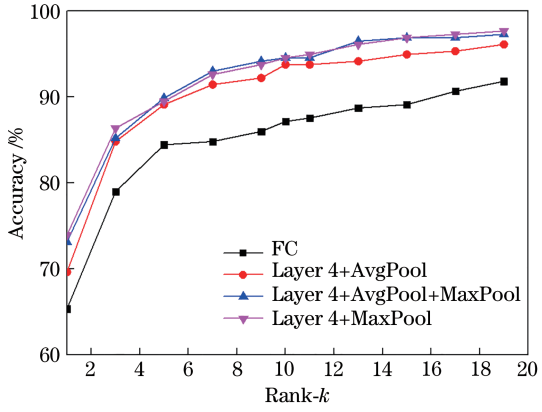

图 10　不同特征层和卷积特征融合方法的对比

Fig. 10　Comparison of different feature layers and convolution feature aggregation methods

Label Smoothing 中超参数 ε 对识别精度的影响如表 3 所示,可以发现,当 ε=0.14 时,模型的泛化能力最好,识别精度最高。

综上所述,本算法在实验数据集上得到的最终 Rank-1、mAP 分别为 74.32% 和 56.97%,相比基础网络 ResNet50,本算法的 Rank-1、mAP 精度分别提升了 13.23 和 10.63 个百分点。

表 3　超参数 ε 对识别精度的影响

Table 3　Influence of hyper parameter ε on recognition accuracy　　unit:%

ε	Rank-1	mAP
Without Label Smoothing	73.93	54.87
0.12	70.04	55.80
0.13	71.21	56.65
0.14	**74.32**	**56.97**
0.15	69.65	55.84

4　结　　论

针对贴近公安刑侦实战的多背景监控鞋型数据集,提出了一种新的注意力机制模型,将其引入 ResNet50 中并进行可视化,进一步增强了 ResNet50 对重要特征的提取能力。同时,探究了用不同特征层的输出作为鞋子特征和不同卷积特征聚合方法对识别精度的影响。为了增强模型的泛化能力,降低模型过拟合的风险,在损失函数中引入了 Label Smoothing 并探究了超参数对识别精度的影响,进一步提升了算法的识别精度,增强了模型的泛化能力。实验结果表明,本算法具有较强的实用性。将深度学习方法应用到公安刑侦实战鞋型自动识别中取得了不错的识别效果,避免了人工查看监控视

频易受到主观影响的问题。但在实际案件中，监控视频中的鞋子可能没有明显标志，且易受光照的影响，只能大致看出鞋子的轮廓，而本算法对没有明显标志、细节信息丢失较多的数据识别效果较差，下一步将重点针对该问题进行改进，使算法更贴近于实战，进一步提升鞋型的识别精度。

参 考 文 献

[1] Jin Y F, Bai Y P, Shi F, et al. Application and optimization of national database of shoes' patterns [J]. Forensic Science and Technology, 2018, 43(6): 511-513.
金益锋, 白艳平, 石峰, 等. 全国公安机关鞋样本数据库应用系统的应用与优化[J]. 刑事技术, 2018, 43(6): 511-513.

[2] Sun Y H. Insight on comprehensive application of various detection means into video tracking[J]. Forensic Science and Technology, 2019, 44(3): 257-260.
孙熠赫. 论视频追踪中多种侦查手段的综合运用[J]. 刑事技术, 2019, 44(3): 257-260.

[3] Jin Y F, Cui J, Shi F, et al. Shoes' information combined into video utilization for case solving[J]. Forensic Science and Technology, 2019, 44(5): 463-465.
金益锋, 崔佳, 石峰, 等. 鞋样本信息与视频在案件侦破中的关联应用[J]. 刑事技术, 2019, 44(5): 463-465.

[4] Razavian A S, Azizpour H, Sullivan J, et al. CNN features off-the-shelf: an astounding baseline for recognition[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops, June 23-28, 2014, Columbus, OH, USA. New York: IEEE Press, 2014: 512-519.

[5] Oquab M, Bottou L, Laptev I, et al. Learning and transferring mid-level image representations using convolutional neural networks[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE Press, 2014: 1717-1724.

[6] Huang J S, Liu S, Xing J L, et al. Circle & search [J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2014, 11(1): 1-21.

[7] Zhan H J, Shi B X, Kot A C. Cross-domain shoe retrieval with a semantic hierarchy of attribute classification network[J]. IEEE Transactions on Image Processing, 2017, 26(12): 5867-5881.

[8] Yang M J, Tang Y Q, Jiang X J. Novel shoe type recognition method based on convolutional neural network[J]. Laser & Optoelectronics Progress, 2019, 56(19): 191505.
杨孟京, 唐云祁, 姜晓佳. 基于卷积神经网络的鞋型识别方法[J]. 激光与光电子学进展, 2019, 56(19): 191505.

[9] Chen Q, Liu L, Fu X D, et al. Fine-grained shoe image retrieval by part detection and semantic network[J]. Journal of Image and Graphics, 2020, 25(8): 1578-1590.
陈前, 刘骊, 付晓东, 等. 部件检测和语义网络的细粒度鞋类图像检索[J]. 中国图象图形学报, 2020, 25(8): 1578-1590.

[10] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[11] Cui J J, Zhao X R, Li D X. An automatic shoeprint retrieval method using neural codes for commercial shoeprint scanners[M]//Yang J F, Hu Q H, Cheng M M, et al. Computer vision. CCCV 2017. Communications in computer and information science. Singapore: Springer, 2017: 158-169.

[12] Zhang H, Wu J X. A survey on unsupervised image retrieval using deep features[J]. Journal of Computer Research and Development, 2018, 55(9): 1829-1842.
张皓, 吴建鑫. 基于深度特征的无监督图像检索研究综述[J]. 计算机研究与发展, 2018, 55(9): 1829-1842.

[13] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 2818-2826.

[14] Müller R, Kornblith S, Hinton G. When does label smoothing help?[EB/OL]. (2020-06-10) [2021-01-04]. https://arxiv.org/abs/1906.02629.

[15] Luo H, Jiang W, Gu Y Z, et al. A strong baseline and batch normalization neck for deep person re-identification[J]. IEEE Transactions on Multimedia, 2020, 22(10): 2597-2609.

[16] Geng P Z, Yang Z X, Zhang J J, et al. Pedestrian shoes detection algorithm based on SSD[J]. Laser & Optoelectronics Progress, 2021, 58(6): 0610009.
耿鹏志, 杨智雄, 张家钧, 等. 基于SSD的行人鞋子检测算法[J]. 激光与光电子学进展, 2021, 58(6):

0610009.

[17] Hu J, Shen L, Albanie S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011-2023.

[18] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11211: 3-19.

[19] Wang Q L, Wu B G, Zhu P F, et al. ECA-Net: efficient channel attention for deep convolutional neural networks[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 11531-11539.

[20] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[J]. International Journal of Computer Vision, 2020, 128(2): 336-359.