第 58 卷 第 4 期/2021 年 2 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

基于改进残差网络的道口车辆分类方法

李宇昕，杨帆*，刘钊，司亚中

河北工业大学电子信息工程学院，天津 300401

摘要 为了提高模型在道口环境下的车辆图像的特征提取和识别能力，提出了一种基于改进残差网络的车辆分类方法。首先以残差网络为基础模型，改进了残差块中激活函数的位置，并将残差块中的一般卷积用分组卷积代替，引入注意力机制，用焦点损失函数替换交叉熵损失函数。实验部分先用公开数据集 Stanford Cars 进行预训练，再用自建的道口车辆数据集进行迁移学习。结果表明，改进模型在两个数据集中的准确率均优于几种经典的深度学习模型。

关键词 机器视觉；注意力机制；车型识别；残差网络；损失函数

中图分类号 TP391.4 **文献标志码** A **doi**：10.3788/LOP202158.0415009

Classification Method of Crossing Vehicle Based on Improved Residual Network

Li Yuxin, Yang Fan*, Liu Zhao, Si Yazhong

School of Electronic and Information Engineering, Hebei University of Technology, Tianjin 300401, *China*

Abstract To improve the feature extraction capability and recognition capability of models for vehicle images in crossing environments, a vehicle classification method based on an improved residual network is proposed. First, the residual network is used as the basic model, the position of the activation function on the residual block is improved, and the normal convolution in the residual block is replaced with a group convolution. An attention mechanism is then added in the residual block. Finally, the focal loss function replaces the cross-entropy loss function. In the experiment, the Stanford Cars public dataset is used for pretraining and a self-built crossing vehicle dataset is used for migration learning. The results show that the classification accuracy of the proposed model is better than several classical deep learning models in both datasets.

Key words machine vision; attention mechanism; vehicle type recognition; residual network; loss function

OCIS codes 150.0155; 150.1135; 100.3008

1 引 言

随着我国经济的快速发展，人民物质生活水平的不断提高，城市车辆数量与日俱增，各大城市的交通拥堵现象和交通事故也在不断增多，给城市交通管理系统造成了不小的压力，高效的车型识别逐渐成为智能交通领域的研究重点。

在当前的交通监控条件下，由于天气情况和道路环境复杂多变、摄像机角度不同、不同款式车辆之间的相似度小等因素，和一般的图像分类任务相比，车型分类难度更大[1]。能否完成车型识别任务的核心是如何找到好的特征。

在传统的车辆分类方法中，手工设计的特征描述子，如尺度不变特征转换(SIFT)[2]等，只能关注图像的浅层特征，对图像的质量要求较高，易受环境影响，鲁棒性差。在深度学习中，借助大量数据的卷积神经网络能够自动学习如何提取图像的深度特征，分类性能远远超过传统方法[3]。近年来，不断有

收稿日期：2020-09-02；**修回日期**：2020-09-28；**录用日期**：2020-11-05

基金项目：国家重点研发计划智能机器人专项(2019YFB1312102)、河北省自然科学基金(F2019202364)

* **E-mail**：commanderjy@163.com

学者将深度学习技术应用于车辆识别及分类领域。例如，Kang 等[4]提出一种轻量级卷积神经网络用于红外车型识别，大大减小了时间和资源成本。张洁等[5]将支持向量机(SVM)和深度卷积网络结合，设计了针对复杂背景的车型分类器。马永杰等[6]在传统卷积神经网络 AlexNet 的基础上结合 SVM，提出一种新的车辆识别方法，相较传统模型，该方法的速度和精度都有提高。张苗辉等[7]提出了一种多任务卷积神经网络，该网络有较好的泛化能力，对车辆图像的分类精度有明显的提升。

然而现有工作实验中使用的数据集都是从车辆正面拍摄采集的，没有其他车辆的干扰，但在实际的道口数据中，由于路况复杂，往往有很多车辆在同一张图像中，还有非机动车也进入机动车道，对识别造成干扰。因此，为了提高传统深度学习模型在真实道口环境下的车型识别准确率，本文提出一种改进残差网络车型识别模型(FA-ResNet)。主要改进方面：对残差块的激活函数在残差块中的相对位置进行替换；使用分组卷积替换传统卷积，在不明显增加参数量的前提下提升了特征图数量，强化了模型提取特征的能力；同时引入注意力机制，让模型可以自适应地对图像内的目标车辆进行训练；训练过程中用焦点损失替换交叉熵损失，这可以增加对难分类样本的权重、减少对易分类样本的权重，使得模型在训练时可以针对目标车辆进行特征提取，增强对相似度较高车型的分类能力。

实验数据集包括公开数据集和基于道路卡口摄像机拍摄的车辆图像自建的车辆数据集。实验结果表明，所提模型在两个数据集上的准确率均优于经典模型。

2　改进的残差网络车型识别模型

2.1　整体结构

以 FA-ResNet 模型为核心，提出了一种针对实际道口图像的车型分类方法，整体框架如图 1 所示，主要由两个模块组成。在第一个模块中，在理想数据集下对改进的深度残差网络进行训练，通过学习获得车辆的特征表示；第二个模块中，针对实际道口摄像机拍摄的图片进行车型分类任务，将训练好的模型迁移学习到第二个模块，对处理好的图像进行分类，输出识别结果。

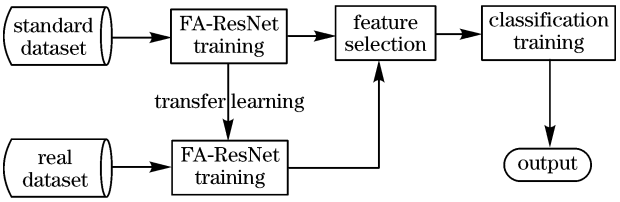

图 1　所提方法的流程图

Fig. 1　Flowchart of the proposed method

2.2　改进残差网络

随着深度神经网络的不断发展和完善，计算机的图像分类能力得到了令人瞩目的提升，例如 VGG 网络[8]和 GoogLeNet[9]。这些结构都是通过增加网络的层数深度来取得更好的训练结果的，但是神经网络的深度并不是越深越好。实验表明，网络深度达到 20 层以后，若继续堆加层数，分类的精度反而会降低。

为了解决这种退化问题，He 等[10]在 2016 年提出了残差网络。网络引入了恒等映射的设计概念，残差块模型如图 2(a)所示，缓解了深度增加带来的梯度爆炸、梯度消失或网络退化等问题，因此提升了信息传递路径的数量，使得网络可以在保证较高准确率的前提下，将深度增加到上千层时可以提取到图像更深层的特征。

改进的残差块如图 2(b)所示，使用分组卷积代替传统卷积，在不增加参数量和运算量的前提下增加特征图数量，且将传统残差块中的最后激活函数位置移动到特征融合之前，并在改进的残差块之后增加了注意力机制。

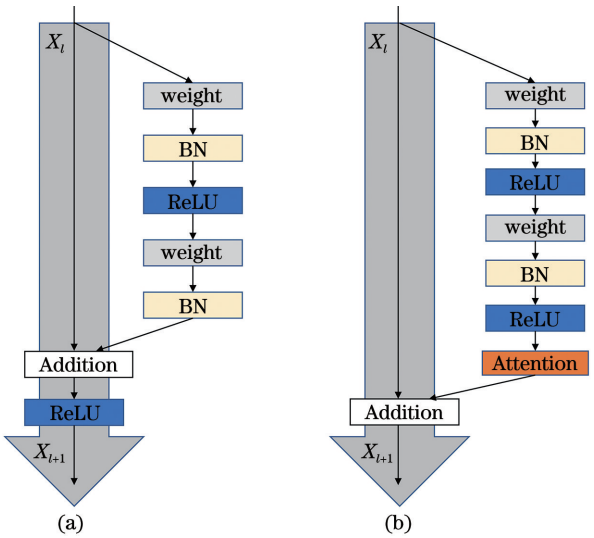

图 2　改进对比。(a)原始残差块；(b)改进残差块

Fig. 2　Improve comparison. (a) Original residual block; (b) improved residual block

2.2.1　分组卷积

一般卷积如图 3 所示。此时，输入特征图的尺寸为 $W\times H\times C$，分别对应特征图的宽、高、通道数；

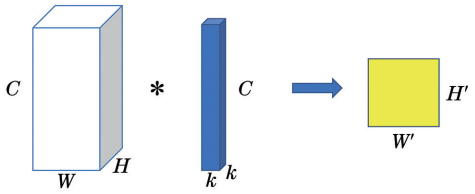

图 3　一般卷积

Fig. 3　Normal convolution

单个卷积核尺寸为 $k\times k\times C$，分别对应单个卷积核的宽、高、通道数；输出特征图的尺寸为 $W'\times H'$，输出通道数等于卷积核数量，输出的宽、高与卷积步长相关。

一般卷积的参数量和运算量分别为

$$p_a=k^2C,\tag{1}$$

$$F=k^2CW'H'。\tag{2}$$

分组卷积，是对输入的特征图进行分组，然后对每组分别进行卷积，如图 4 所示。

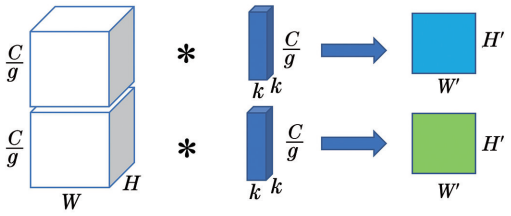

图 4　分组卷积

Fig. 4　Group convolution

假设输入特征图尺寸为 $W\times H\times\frac{C}{g}$，共有 g 组；单个卷积核每组的尺寸为 $k\times k\times\frac{C}{g}$，一个卷积核被分成 g 组；输出特征图的尺寸为 $W'\times H'\times g$，共生成 g 个特征图。因此分组卷积时的参数量和运算量分别为

$$p_a=k^2\times\frac{C}{g}\times g=k^2C,\tag{3}$$

$$F=k^2\times\frac{C}{g}\times W'\times H'\times g=k^2CW'H'。\tag{4}$$

由(3)、(4)式可知，尽管分组卷积被分成了 g 个特征图，但是它的参数量、运算量和普通的卷积是相同的。因此在同等条件下，使用分组卷积可以生成大量的特征图，即能够编码更多信息，强化模型的特征提取能力，让残差块可以提取到更多的车辆细节信息。

2.2.2　注意力机制

实际的卡口路况图像往往含有很多非目标车辆信息，可能会干扰车型的识别，给交通管理带来不必要的工作。在真实道路数据集的六分类任务中，会出现非机动车辆进入图片采集区、相邻车道的汽车也被采集等情况。在深度卷积网络中加入注意力机制后，网络能对特征进行自动选择，以此来获得更多具有关注性的信息，提高系统整体的识别准确率和速度。

在计算机视觉中引入注意力机制的目的在于使卷积神经网络更多关注具有较高信息量的区域或通道。很多学者以不同的方式将深度卷积网络和注意力机制结合。刘航等[11]提出一种基于注意力机制的遥感图像分割模型，该模型使用注意力机制进行加权处理，增强目标特征并抑制背景信息。席志红等[12]设计了一种基于残差注意力和多级特征融合的图像重建网络，该网络通过引入注意力机制来自适应地校正信道特征，提高网络表征力。Wang 等[13]提出一种注意力模块，该模块由传统卷积操作和两个下采样构成，并充当注意力图谱，扩大了底层特征的感受野，提高了分类的准确率。

注意力机制模型如图 5 所示。将输入分为两路，一路经过全局池化层(GP)和全连接层(FC)，压缩成 C 个一维的特征图权重，对每个通道的重要性进行预测，将权重与另一路输入中每个值相乘得到输出。注意力机制让模型可以更加关注信息量最大的通道特征，而抑制那些不重要的通道特征。

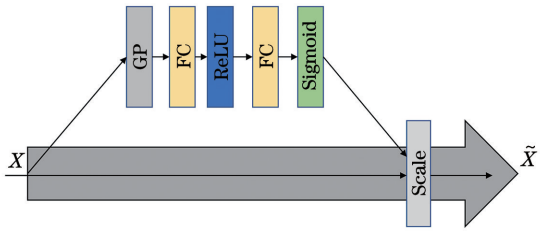

图 5　注意力模型

Fig. 5　Attention model

加入注意力机制后的效果可以通过图 6 中的热力图[14]来直观展示。实际城市道口中的每一台摄像机只针对一个车道的车辆进行拍摄，但是拍摄时往往会将其他车道的车辆拍摄进图像中，并且有些非机动车辆也会驶入机动车道，被摄像机拍摄到。本文中的目标车辆是指在被拍摄车道的机动车辆，非目标车辆是指被拍摄到的非机动车辆和处于其他车道的(非拍摄区域的)机动车辆。从图 6 可以看到：传统残差网络在对该图像进行处理时，对两个车道行驶的车辆均进行了特征提取；在加入注意力机制后，网络可以把特征的提取集中到左边的目标车辆微型面包车上，而对右边

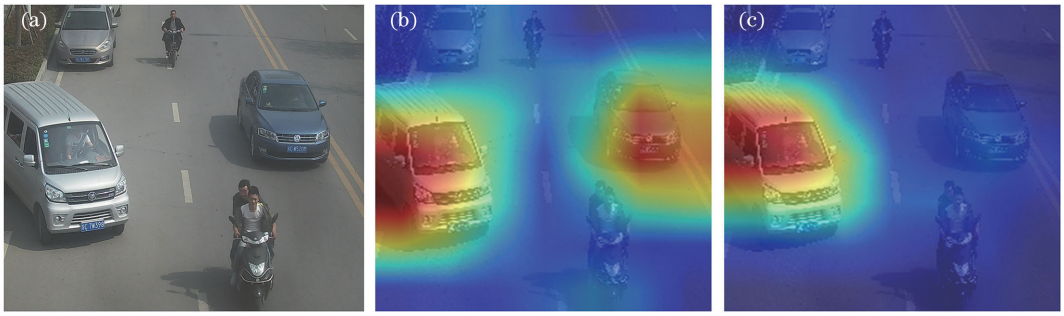

图 6　不同模型处理的热力图。(a)原图;(b)原始模型 ResNet;(c)增加注意力机制后的模型

Fig. 6　Heat maps processed by different models. (a) Original map; (b) original model ResNet; (c) model with attention mechanism

非目标车辆和非机动车不进行处理,可以较大地提高网络的分类效率。

2.3　损失函数

由于训练数据集中的车型种类较多,为了降低相似车型之间的影响,使用焦点损失代替交叉熵损失。

传统交叉熵损失函数为

$$C_{\mathrm{E}}(p,y)=\begin{cases}-\log(p), & y=1\\ -\log(1-p), & y\neq 1\end{cases},\tag{5}$$

式中:y 为数据标签;p 为概率。为了表示方便,用 p_{t} 代替 p,则表达式为

$$p_{\mathrm{t}}=\begin{cases}p, & y=1\\ 1-p, & y\neq 1\end{cases}。\tag{6}$$

将(6)式代入(5)式中,得到

$$C_{\mathrm{E}}(p,y)=C_{\mathrm{E}}(p_{\mathrm{t}})=-\log(p_{\mathrm{t}})。\tag{7}$$

为了控制正负样本对总损失的权重,增加一个参数 α_{t},通过对 α_{t} 取一个较小值来降低负样本的权重。

$$C_{\mathrm{E}}(p_{\mathrm{t}})=-\alpha_{\mathrm{t}}\log(p_{\mathrm{t}})。\tag{8}$$

在(8)式的基础上,再增加一个控制容易分类和难分类样本的权重,减少易分类样本的权重,使模型在训练时更专注于难分类的样本。于是焦点损失的公式为

$$L_{\mathrm{fl}}=\begin{cases}-(1-p_{\mathrm{t}})^{\gamma}\log\ p_{\mathrm{t}}, & y=1\\ -p_{\mathrm{t}}^{\gamma}\log(1-p_{\mathrm{t}}), & y=0\end{cases},\tag{9}$$

式中:焦点参数 $\gamma\geqslant 0$;$(1-p_{\mathrm{t}})^{\gamma}$ 为调制系数。

多分类任务下的焦点损失为

$$L_{\mathrm{fl}}=-(1-p_{\mathrm{prediction}})^{\gamma}\log\ p_{\mathrm{prediction}},\tag{10}$$

式中:$p_{\mathrm{prediction}}$ 为目标的预测值。

3　实验结果分析

实验包含两部分:第一部分为 FA-ResNet 与现有模型在 Stanford Cars 数据集[15]上的实验与分类结果对比;第二部分为 FA-ResNet 在自建道口车辆数据集上的消融实验结果与分析。

3.1　实验环境与参数设置

实验所使用的计算机为 T640 图形工作站,Ubuntu 操作系统,64 GB 内存,使用 Pytorch 深度学习框架,GPU 配置为 GeForce GTX 1080Ti 12 GB,CPU 处理器配置为 Interl Xeon(R) Silver 4114 2.20 GHz。

由于分组卷积的分组数必须要能整除输入通道数,模型中的分组数必须是 2^n,所以分组数选择 2。焦点损失中的焦点参数设置为 $\gamma=2$,是根据文献[16]进行取值的。训练参数设置如下:使用 SGD 算法更新参数,动量参数设置为 0.89,训练批次为 4,初始学习率为 0.001,并且每训练 20 个 epoch 学习率降低 10%,当训练的损失值不再明显下降时停止训练。

3.2　Stanford Cars 数据集结果与分析

理想实验使用斯坦福大学的 Stanford Cars 数据集,该数据集共有 16185 张车辆图片,包含 196 种车辆型号,训练集共 8144 张图像,测试集共 8041 张图像。数据集中部分图像如图 7 所示。

实验采用的评价指标为分类准确率,公式为

$$A=\frac{\sum_{i=1}^{m}f(x_i)=y_i}{m},\tag{11}$$

式中:i 为样本序号;m 为样本数;$f(x_i)$为模型预测输出;y_i 为真实标签。

为了对 FA-ResNet 进行验证和分析,选取文献[17-19]的分类结果作为对比,对比结果如表 1 所示。

表 1　不同模型在 Stanford Cars 数据集中的准确率

Table 1　Accuracy of different models on Stanford Cars dataset

Model	Accuracy /%
Three-scale Attention[17]	81.50
B-CNN[18]	86.50
Kernel-Pooling[19]	85.70
FA-ResNet	86.97

图 7　Stanford Cars 数据集中的部分图像

Fig. 7　Partial images in Stanford Cars dataset

文献[17]提出一种多样化视觉注意力网络来解决细粒度分类问题，获得 81.50%的准确率；文献[18]使用一种将 VGG 模型作为骨架模型进行双线性特征融合的编码方式，得到 86.50%的准确率；文献[19]在数据集没有额外标注的情况下，在 ResNet-50 的基础上使用池化核改进残差块，最终得到 85.70%的准确率；由于 FA-ResNet 不仅改进残差块、引入注意力机制，同时使用焦点损失，这可以让网络更加关注于损失大的难训练样本，提高了模型对数据集整体的识别准确率，达 86.97%。

3.3　自建道口车辆数据集实验结果与分析

实际道口数据集手工分为 6 类：小轿车、微型面包车（以下简称微面）、SUV、货车、大客车（大巴和公交）、其他（自行车和电动车等），共 15988 张图片。随机抽取 80%作为训练集，20%作为测试集。图像中包含了复杂的真实道口交通情况，以验证 FA-ResNet 的准确性。道口数据集部分图像如图 8 所示。

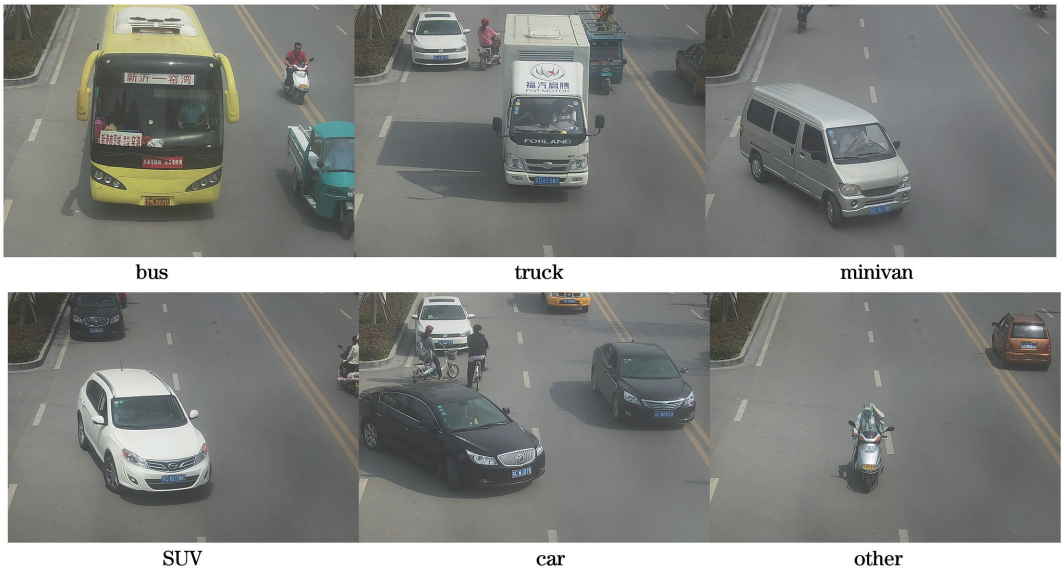

图 8　实际道口数据集中的部分图像

Fig. 8　Partial images in real crossing dataset

为了提升模型对道口环境下车辆图像的特征提取能力，采用分组卷积（GC）、增加注意力机制（AT）、使用焦点损失（FL）3 种改进措施。为了表明各项方法的有效性，在道口车辆数据集中，控制一项作为变量进行消融实验，实验结果如表 2 所示，并选择实验 1，2，7，8 的结果绘制曲线图，如图 9、10 所示。

表 2　消融实验结果

Table 2　Results of ablation experiment

Experiment No.	Group convolution	Attention model	Focal loss	Accuracy /%
1				80.44
2	√			81.12
3		√		88.43
4			√	90.15
5	√	√		88.91
6	√		√	92.13
7		√	√	94.19
8	√	√	√	94.96

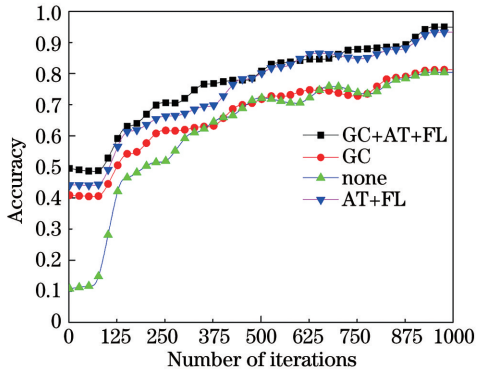

图 9　消融实验的准确率

Fig. 9　Accuracy of ablation experiment

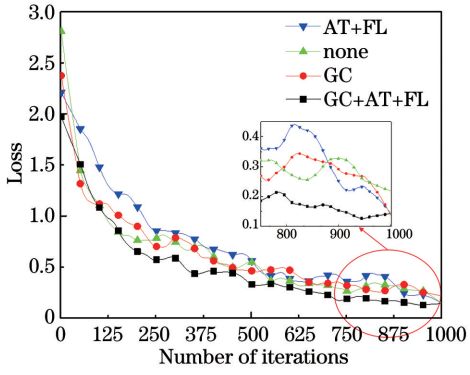

图 10　消融实验的损失

Fig. 10　Loss of ablation experiment

从实验 1 和实验 2、实验 7 和实验 8 的对比中可以看出，分组卷积可以小范围提升模型的分类性能，准确率可以提升 0.5 个百分点到 0.8 个百分点；从实验 1 和 3、4 与实验 7 和 6、8 的两两对比中可以发现，注意力机制对模型分类准确率的提升有较大帮助，但是提升的幅度不稳定；通过实验 2 和实验 6、实验 5 和实验 8 的对比可以知道，焦点损失可以有效且稳健地提升模型的分类准确率。

综上可以知道，改进损失函数对模型训练准确率的提升有很大影响，改进卷积方式可以在一定范围内有效提升模型提取特征的能力。注意力机制在面对两个车道有相同标签的机动车时，虽然可以正确分类，但是识别的车道会有错误，使得模型对以后的目标车辆判断出现误差，导致准确率不稳定。

4　结　　论

在真实的道路图像中，往往有很多因素干扰对目标车辆的识别。为了增加模型对图片整体信息的把握，提出了一种基于残差网络的道口车辆分类模型。所提方法在传统深度残差网络的基础上进行改进：重新设计激活函数在残差块中的相对位置，并用分组卷积代替了传统卷积，同时加入注意力机制，进一步提升了对车型特征的提取准确率。在训练过程中，使用焦点损失代替传统的交叉熵损失，使得模型在训练时更专注于难分类样本，实验结果表明，焦点损失的使用可以更好地增强网络对车型的识别能力。Stanford Cars 数据集上的实验结果表明，所提改进模型有较高的准确率。为了应对真实复杂的道口情况，进一步增加消融实验，将在 Stanford Cars 数据集上训练好的模型迁移学习到自建的道口图像数据集中，结果表明所提模型在自建数据集中依然有较好的识别效果。

但是由于道口的交通情况过于复杂，所提模型无法对少部分比较复杂的图片进行更高精度的识别。如何更高效地确定目标车辆是下一步研究的方向，同时所提模型仍有一定的可优化空间。

参　考　文　献

[1] Zhao D B, Chen Y R, Lv L. Deep reinforcement learning with visual attention for vehicle classification[J]. IEEE Transactions on Cognitive and Developmental Systems, 2017, 9(4): 356-367.

[2] Lowe D G. Distinctive image features from scale-invariant keypoints [J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[3] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks

[C]// Proceedings of the 25th International Conference on Neural Information Processing Systems, December 3-6, 2012, Lake Tahoe, Nevada. New York: Curran Associates, 2012: 1097-1105.

[4] Kang Q, Zhao H D, Yang D X, et al. Lightweight convolutional neural network for vehicle recognition in thermal infrared images[J]. Infrared Physics & Technology, 2020, 104: 103120.

[5] Zhang J, Zhao H D, Li Y H, et al. Classifier forrecognition of fine-grained vehicle models under complex background[J]. Laser & Optoelectronics Progress, 2019, 56(4): 041501.
张洁, 赵红东, 李宇海, 等. 复杂背景下车型识别分类器[J]. 激光与光电子学进展, 2019, 56(4): 041501.

[6] Ma Y J, Ma Y T, Chen J H. Vehicle recognition based on multi-layer features of convolutional neural network and support vector machine[J]. Laser & Optoelectronics Progress, 2019, 56(14): 141001.
马永杰, 马芸婷, 陈佳辉. 结合卷积神经网络多层特征和支持向量机的车辆识别[J]. 激光与光电子学进展, 2019, 56(14): 141001.

[7] Zhang M H, Zhang B, Gao C C. Object classification based on multitask convolutional neural network[J]. Laser & Optoelectronics Progress, 2019, 56(23): 231502.
张苗辉, 张博, 高诚诚. 一种多任务的卷积神经网络目标分类算法[J]. 激光与光电子学进展, 2019, 56(23): 231502.

[8] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10)[2020-09-01]. https://arxiv.org/abs/1409.1556.

[9] Szegedy C, Liu W, Jia Y Q, et al. Going deeper with convolutions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition, June 7-12, 2015, Boston, MA. New York: IEEE Press, 2015.

[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, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[11] Liu H, Wang X L. Remote sensing image segmentation model based on attention mechanism[J]. Laser & Optoelectronics Progress, 2020, 57(4): 041015.
刘航, 汪西莉. 基于注意力机制的遥感图像分割模型[J]. 激光与光电子学进展, 2020, 57(4): 041015.

[12] Xi Z H, Yuan K P. Super-resolution image reconstruction based on residual channel attention and multilevel feature fusion[J]. Laser & Optoelectronics Progress, 2020, 57(4): 041504.
席志红, 袁昆鹏. 基于残差通道注意力和多级特征融合的图像超分辨率重建[J]. 激光与光电子学进展, 2020, 57(4): 041504.

[13] Wang F, Jiang M Q, Qian C, et al. Residual attention network for image classification[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6450-6458.

[14] 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.

[15] Krause J, Stark M, Jia D, et al. 3D object representations for fine-grained categorization[C]// 2013 IEEE International Conference on Computer Vision Workshops, December 2-8, 2013, Sydney, NSW, Australia. New York: IEEE Press, 2013: 554-561.

[16] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 318-327.

[17] Zhao B, Wu X, Feng J S, et al. Diversified visual attention networks for fine-grained object classification[J]. IEEE Transactions on Multimedia, 2017, 19(6): 1245-1256.

[18] Lin T Y, RoyChowdhury A, Maji S. Bilinear convolutional neural networks for fine-grained visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(6): 1309-1322.

[19] Wang Y M, Morariu V I, Davis L S. Learning a discriminative filter bank within a CNN for fine-grained recognition[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 4148-4157.