激光与光电子学进展 先进成像

基于改进Transformer的小目标车辆精确检测算法

谢光达[1,2]，李洋[2*]，曲洪权[2]，孙再鸣[3]

[1]北方工业大学电气与控制工程学院，北京 100144；

[2]北方工业大学信息学院，北京 100144；

[3]华北电力大学控制与计算机工程学院，北京 102206

摘要 智能交通系统的建立离不开车辆检测技术。目前的主流方案是使用卷积神经网络(CNN)架构进行车辆检测，然而在复杂交通场景中，远距离小目标像素点少，CNN的下采样机制导致提取的特征缺乏充足的上下文信息，因而小目标检测面临极大挑战。针对这个问题，提出了一种基于视觉Transformer的小目标车辆检测算法。所提算法通过改进Transformer的线性嵌入模块，补充小目标的线性嵌入信息；对图像进行层级构建，每层仅对局部进行关系建模，同时扩大感受野，代替CNN提取出更强有力的小目标车辆特征，实现端到端的精确检测。在UA-DETRAC车辆数据集上进行验证，实验结果表明，改进后的车辆检测算法提高了对远距离及严重遮挡情况下小目标的检测性能，检测精度达到99.0%。

关键词 机器视觉；车辆检测；小目标；图像增强；视觉Transformer

中图分类号 O436 **文献标志码** A **DOI**：10.3788/LOP202259.1815016

Small Target Accurate Vehicle Detection Algorithm Based on Improved Transformer

Xie Guangda[1,2], Li Yang[2*], Qu Hongquan[2], Sun Zaiming[3]

[1]*School of Electrical and Control Engineering, North China University of Technology, Beijing* 100144, *China*;

[2]*Information College, North China University of Technology, Beijing* 100144, *China*;

[3]*School of Control and Computer Engineering, North China Electric Power University, Beijing* 102206, *China*

Abstract Intelligent transportation systems have been playing a major role in vehicle detection technology. Recently, the convolutional neural network (CNN) architecture is a popular method for vehicle detection. However, in complex traffic situations, only fewer pixels for long-distance small targets are available, and CNN's subsampling mechanism seems to be lacking sufficient context information of some extracted features, which gives small target detection great challenges. A small target vehicle detection algorithm based on a visual Transformer was introduced in this paper to solve the aforesaid problem. By improving the linear embedding module of the Transformer, information on the small targets was supplemented. Additionally, the image was constructed hierarchically, and each layer was only related to the part. Modeling, while expanding, the receptive field, instead of CNN, was conducted to extract more powerful features from small target vehicles to achieve accurate end-to-end detection. Data was verified using the UA-DETRAC vehicle dataset. The experimental results showed that the improved vehicle detection algorithm enhanced the detection performance of small targets at long distances and under severe occlusion conditions and that the detection accuracy reached 99.0%.

Key words machine vision; vehicle detection; small target; image enhancement; visual Transformer

1 引 言

随着城市智能交通的不断完善，车辆检测成为了自动驾驶中一个非常重要的工作[1]。通过分析交通摄像头采集到的图像数据，实现车辆目标检测是城市智能交通建设的重要组成部分。在实际数据采集过程中，受摄像头位置、角度等因素的影响，往往存在远距离车辆目标尺寸较小、遮挡严重等问题，导致车辆外观

收稿日期：2021-06-11；**修回日期**：2021-07-13；**录用日期**：2021-09-02

基金项目：国家自然科学基金面上项目(61971456)、毓优人才培养计划项目(218051360020XN115/014)

通信作者：*2442590777@qq.com

信息较少，现有算法难以精确检测。车辆检测的检测距离越远、精度越高，智能交通的安全性就会越好。因此，关于从复杂交通场景中高效、精确地检测出小目标车辆，提高检测距离的研究，对智能交通系统的管理与控制具有现实意义。

近年来，基于深度学习卷积神经网络（CNN）的图像目标检测算法发展迅速[2]，并且已经在智能交通领域的车辆检测任务中取得了良好的效果。CNN具有很强的图像特征学习能力，但它对目标的尺度变化比较敏感[3-4]。针对这个问题，Kampffmeyer等[5]提出了一种基于像素级、块区域及两者结合的深度卷积神经网络架构来实现航拍图像中单像素的分类，并构建土地覆盖图以实现小目标检测。刘峰等[6]提出了一种改进的YOLOV3算法，该算法解决了小目标特征容易丢失、分辨率低的问题，提升了小目标的检测性能。刘力荣等[7]采用Slim Net实现了全景图像上小目标的检测，平均正确率相比经典的VGG16有4.2个百分点的提升。Ren等[8]提出Recurrent Rolling Convolution（RRC）结构，改进了SSD算法对被遮挡物体或小物体的检测效果。Takeki等[9]针对大背景区域下的小目标检测问题，提出了一种基于深度卷积神经网络在大范围视场区域内检测小型鸟类目标的模型，并实现了高检测性能。Dosovitskiy等[10]直接将图像分成图块序列然后通过Transformer进行图像分类任务，其提出的视觉Transformer（ViT）与基于CNN的算法不同之处在于，Transformer通过注意力机制就可以获得每个图像块之间的语义信息，能够从一开始就获得一个全局感受野，充分利用了上下文语义信息，对小目标有更出色的识别能力，同时所需的计算资源也大大减少。

本文提出了一种融合图像增强并通过层级视觉Transformer算法进行小目标车辆检测的算法。首先改进了Transformer的线性嵌入模块，融合边缘增强网络，对摄像头采集的道路图像增强后，图片边缘及小目标更加清晰，提高了道路两边密集的停车区域、远距离的小目标及遮挡情况下车辆目标的分辨率。然后通过层级视觉Transformer对图像进行特征向量编码，构建了类似于CNN中的特征金字塔，但完全摒弃了CNN进行特征提取的方式，这种结构对于交通场景下多尺度车辆检测模型的构建十分友好，提高了小目标车辆检测的精度。

2　所提算法原理与相关实验

车辆检测的目的是检测到图像中所有的车辆目标，目前最常见的方法是训练CNN模型提取图像高级特征，然后通过分类回归得到特定目标的边界框。与现有方法不同，为了充分利用图像上下文信息，提高对远距离小目标和遮挡情况下的检测效果，本实验组使用具有移位窗口的层级视觉Transformer架构来实现车辆检测。

2.1　基于视觉Transformer的图像增强模型

近年来，基于生成对抗网络（GAN）的超分辨率重建模型表现出了显著的图像增强性能，这有助于提高小目标分辨率，但重建后图像通常会丢失高频边缘信息，这会影响检测效果。为了解决这个问题，本实验组基于CNN的融合超分辨率重建和边缘增强模型[11]（EESRGAN）的思想，建立了一个基于Transformer的图像增强网络模型，模型结构如图1所示，由超分辨率生成器（G）、判别器（D_{Ra}）和边缘增强网络（EEN）3个部分组成。对于分辨率较低的原始交通图像（LR），经过生成器生成中间超分辨率图像（ISR），然后再经过一个EEN输出增强边缘的超分辨率图像（SR）。判别器分别接收高分辨率图像（HR）和生成器生成的中间超分辨率图像进行判别，判别器通过计算判别损失将梯度反向传播到生成器中，用来指导生成器训练。

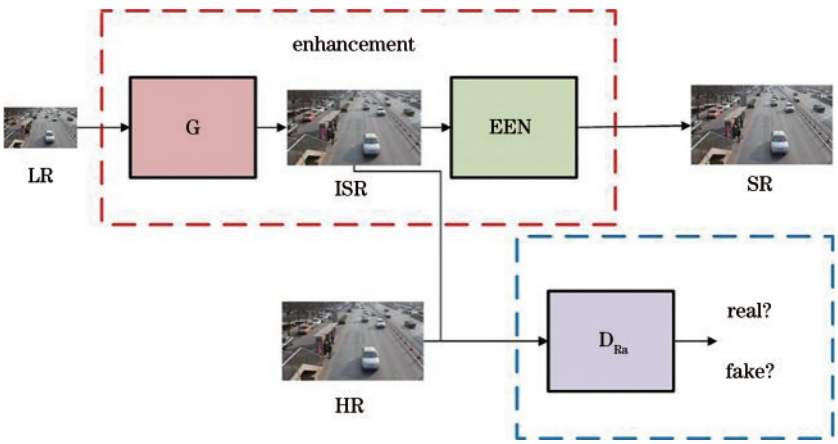

图1　图像增强模型整体架构

Fig. 1　Overall architecture of image enhancement model

对于生成器，采用ESRGAN[12]的生成器模型，该模型通过移除所有批标准化层(BN)在降低计算复杂度的同时提高了泛化能力，在此基础上，本实验组使用Transformer块代替CNN密集连接块(Dense Block)来提高模型的特征表示能力，具体结构如图2所示。生成器的输入是原始低分辨率图像，经过线性嵌入(linear embedding)处理后经过若干个Transformer块生成高级特征，然后通过上采样后得到高分辨率图像。

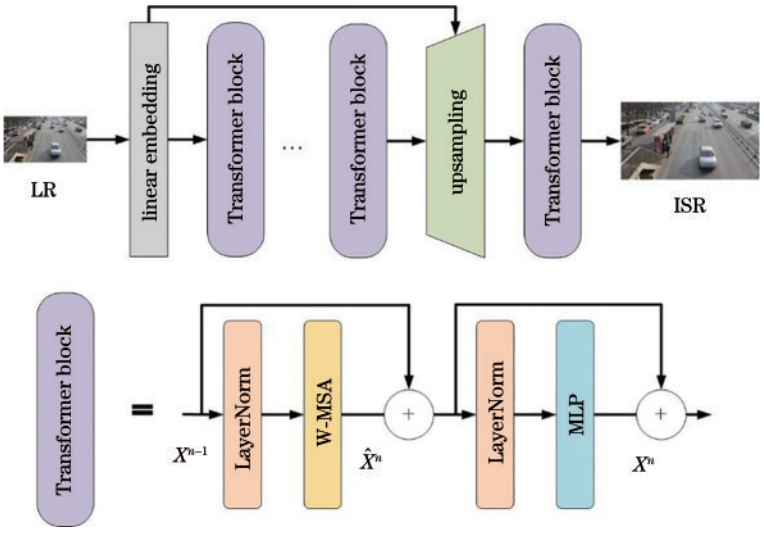

图2　生成器(G)(上)和Transformer块(下)

Fig. 2　Generator (G) (up) and Transformer block (down)

对于判别器，采用VGG16结构的网络模型，判别器是依赖生成器的，其预测了真实图像x_r比生成的中间伪图像x_f更加真实的概率。判别器的计算方式为

$$D_{Ra}(x_r,x_f)=\sigma\left\{C(x_r)-E_{x_f}\left[C(x_f)\right]\right\},\tag{1}$$

式中：σ表示Sigmoid函数；$C(x)$表示判别器的输出；$E_{x_f}[\cdot]$表示对所有生成器生成的伪图像数据取平均值。判别器和生成器的损失函数分别为

$$L_D^{Ra}=-E_{x_r}\left\{\log\left[D_{Ra}(x_r,x_f)\right]\right\}-E_{x_f}\left\{\log\left[1-D_{Ra}(x_f,x_r)\right]\right\},\tag{2}$$

$$L_G^{Ra}=-E_{x_r}\left\{\log\left[1-D_{Ra}(x_r,x_f)\right]\right\}-E_{x_f}\left\{\log\left[D_{Ra}(x_f,x_r)\right]\right\}。\tag{3}$$

生成器和判别器的损失是对称的，而且同时包含真实图像x_r和伪图像x_f。因此，生成器在对抗性训练中受益于生成数据和真实数据的梯度，这种设计方式有助于生成器学习生成更清晰的边缘和更详细的纹理。

边缘增强网络的结构如图3所示，经过生成器生成的中间超分辨率图像首先经过拉普拉斯算子提取边缘信息，然后经过线性嵌入操作后通过若干Transformer块提取特征，使用Sigmoid激活函数消除边缘噪声，最后将增强后的边缘信息添加到输入图像中同时减去由拉普拉斯算子提取的边缘信息，最终得到超分辨率图像。

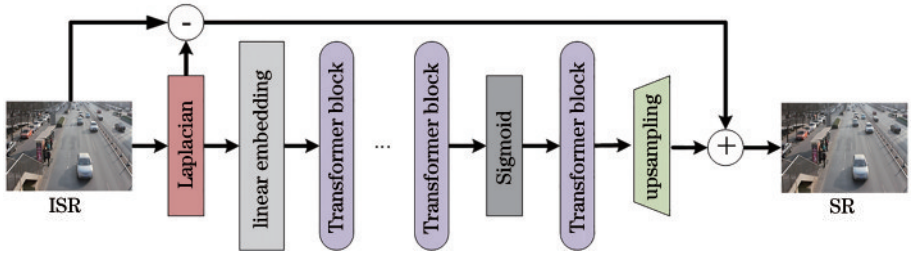

图3　边缘增强网络

Fig. 3　Edge-enhancement network

2.2　基于视觉Transformer的车辆检测模型

目前，基于视觉Transformer的目标检测算法表现出了强大的性能，原因在于Transformer并不像CNN一样通过不断地堆叠卷积层来完成对图像从局部信息到全局信息的提取，而是通过特有的长距离依赖使得模型从浅层到深层都能较好利用全局有效信息，同时多头注意力机制(MSA)保证了网络可以关注到多个图像像素间的关联信息，从而达到较好的检测效果。

相比目前常用的基于CNN的车辆检测算法，本实验组所采用的算法是具有移位窗口的层级视觉

Transformer 目标检测算法，该算法将目前最新的 Swin Transformer[13]框架应用到车辆检测任务当中，其检测流程如图 4 所示。首先，通过补丁分区（patch partition）将输入图片 $H\times W\times 3$ 划分为不重合的补丁（patch）集合，其中每个补丁尺寸为 4×4，特征维度为 $4\times 4\times 3$，数量为 $H/4\times W/4$；然后通过线性嵌入（linear embedding）将划分后的补丁特征维度改为 $4\times 4\times C$，并送入多个 Swin Transformer 块中；之后通过补丁合并（patch merging）将输入按照 2×2 的相邻补丁合并，这样补丁块的数量就变为了 $H/8\times W/8$，特征维度就变为 $4C$，重复 n 次此步骤，直到补丁块的数量变为 $H/32\times W/32$，特征维度变为 $8C$；最后送入回归头（regression head）进行目标分类和定位回归。

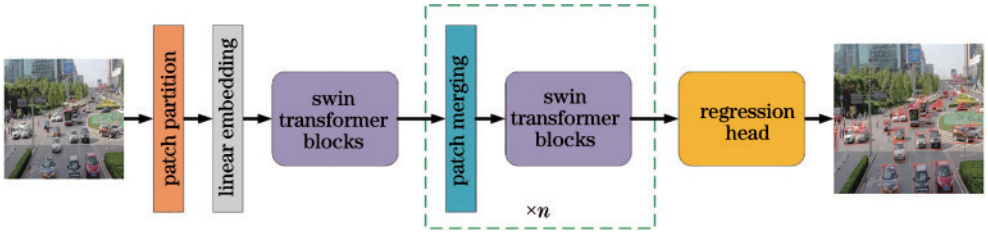

图 4　Swin Transformer 检测流程

Fig. 4　Swin Transformer detection process

[图 5(a)]为两个连续 Swin Transformer 块的示意图，一个 Swin Transformer 块由一个带两层多层感知机（MLP）的多头自注意力模块组成，在每个多头自注意力模块和每个多层感知机之前使用层归一化（LN）操作，之后使用残差连接。当两个连续的 Transformer 模块串联时，多头注意力模块分别为常规窗口多头自注意力（W-MSA）和移位窗口多头自注意力（SW-MSA）。常规窗口划分方案如[图 5(b)]所示，常规划分将 8×8 尺寸的特征图划分 2×2 个窗口，每个窗口有 4×4 个补丁块，然后在每个窗口内计算自注意力。移位窗口划分的分区可以移动，产生新窗口，如[图 5(c)]所示。新窗口跨越了常规窗口的边界，其目的是提供相邻窗口间信息的连接，基于引入局部的思想，每层仅对局部进行关系建模，同时不断缩小特征图的宽高，进而扩大感受野，保持非重叠窗口的高效计算。

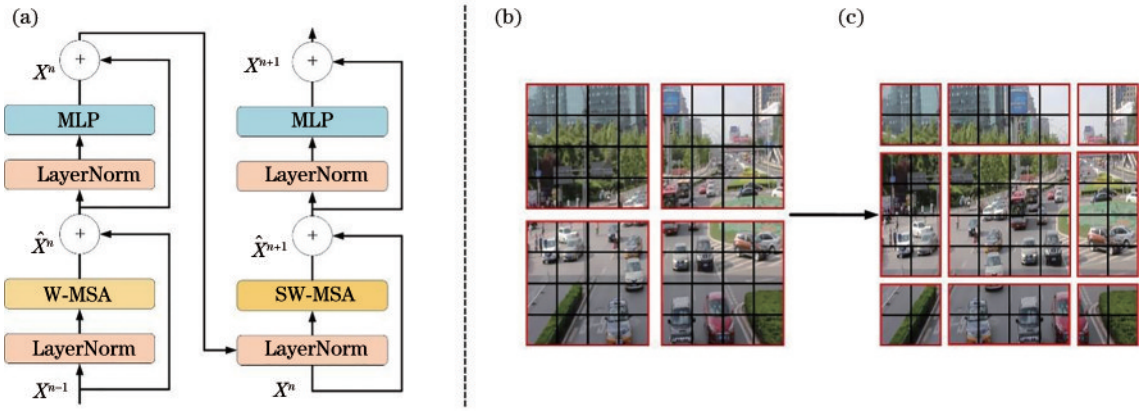

图 5　Swin Transformer 块和具有移位窗口的多头自注意力。(a) Swin Transformer 块；(b) W-MSA；(c) SW-MSA

Fig. 5　Swin Transformer block and SW-MSA. (a) Swin Transformer block; (b) W-MSA; (c) SW-MSA

2.3　端到端的车辆检测总体框架

尽管拥有层级结构和具有移位窗口的 Transformer 在检测领域已经具有良好的性能，但是这种策略并不是 Transformer 在车辆检测中的最佳使用，因为通过移动窗口和补丁划分后的补丁块的尺寸 $H/P\times W/P$ 比原始图像分辨率 $H\times W$ 小得多，这会导致低层次细节的丢失。因此，为了补偿这种信息损失，本实验组采用一种融合边缘增强和超分辨率重建的视觉 Transformer 的架构提取交通图像特征，然后通过 Cascade Mask R-CNN[14]的回归头实现精确车辆检测，如图 6 所示。

在训练阶段，使用在 ImageNet 上预训练的视觉 Transformer 编码器来进行特征提取，维度参数 C 设置为 96，Transformer 块的数量 n 设置为 18。Cascade Mask R-CNN 回归模块包括一个分类器 h_x 和一个回归器 f_x，分类器将一个图像块 x 分配给 $M+1$ 类中的一个，多出的一类代表背景类，分类器的损失设置为

$$R_{\mathrm{cls}}[h]=\sum_{i=1}^{N}L_{\mathrm{cls}}\left[h\left(x_i\right),y_i\right],\tag{4}$$

式中：$h(x)$ 是类别后验分布的 $M+1$ 维估计；x_i 是网络输入；y_i 是类别编号；L_{cls} 为交叉熵损失；N 为批处理

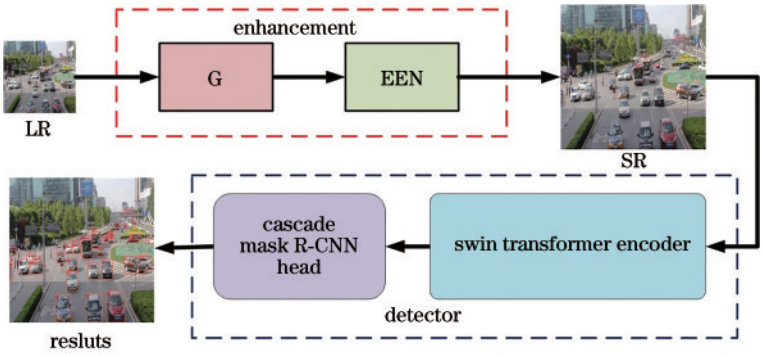

图 6　车辆检测流程

Fig. 6　Vehicle detection process

大小。在标签推测时，因为边界框通常包括一个对象和一些背景，很难确定检测的对错，所以通过 IoU 指标来解决，如果高于阈值 u，则图像块 x 负责该目标的预测，假设 x 的类别标签是关于 u 的函数，根据 u 进行推测：

$$y=\begin{cases} g_y, & \mathrm{IoU}(x,g)\geqslant u \\ 0, & \mathrm{IoU}(x,g)<u \end{cases}, \tag{5}$$

式中：g_y 是真实框位置标签；g 为真实类别。回归器的任务是使用回归器 $f(x,b)$ 将一个候选框 b 回归到真实目标框 g 的位置，一个框包含 (b_x,b_y,b_w,b_h) 4 个坐标，回归器的损失设置为

$$R_{\mathrm{loc}}[f]=\sum_{i=1}^{N}L_{\mathrm{loc}}\left[f\left(x_i,b_i\right),g_i\right], \tag{6}$$

式中：L_{loc} 为 L_2 损失。另外在每个训练阶段 t 对 IOU 阈值进行优化，优化的级联损失定义为

$$L\left(x^t,g\right)=L_{\mathrm{cls}}\left[h_t\left(x^t\right),y^t\right]+\lambda\left\{y^t\geqslant 1\right\}\times L_{\mathrm{loc}}\left[f_t\left(x^t,b^t\right),g\right], \tag{7}$$

式中：$b^t=f_{t-1}\left(x^{t-1},b^{t-1}\right)$；$g$ 是 x^t 的真实框对象；权衡系数 $\lambda=1$；$\{\cdot\}$ 是指标函数；y^t 是标签。级联损失保证了经过有效训练的检测器对位置的检测效果是不断提高的，而且在推断时，通过应用相同的级联过程，假设的质量也将被顺序提高，以此来提高检测效果。

3　实验与结果分析

3.1　实验环境与实验数据集

硬件环境：CPU 为 Intel（R）Xeon（R）Silver 4210R CPU@2.40 GHz，GPU 为 4 个 Nvidia GeForce RTX 2080Ti，操作系统为 Ubuntu18.04。软件环境：CUDA10.1、Python3.7、OpenCV3.4、PyTorch1.6、MMDetection2.12.0 和 mmcv-full1.3.4。

本实验训练车辆检测模型所使用的数据集是 UA-DETRAC 公开数据集，包括 82085 张道路车辆图片，主要拍摄于北京和天津的道路过街天桥，多为摄像头俯视视角，示例如图 7 所示，训练集、验证集、测试集比为 7∶1∶2。

图 7　UA-DETRAC 数据集示例

Fig. 7　Examples of UA-DETRAC dataset

3.2　车辆检测模型训练过程

在检测器的训练中，采用Transformer在ImageNet上预训练的模型作为特征提取模型，优化器使用的是基于权重衰减的AdamW优化算法，初始学习率设置为6×10^{-5}，并使用线性学习率衰减，权重衰减设置为每1500次迭代衰减0.01，并对每张图片进行随机水平翻转，在[0.5，2.0]比例范围内进行随机缩放和随机光度失真。所提模型在8块RTX 2080Ti显卡上进行了33万次迭代，共耗时32 h，模型收敛，损失函数曲线如[图8(b)]所示，其中损失下降到0.24左右，在IOU为0.50的情况下，车辆类别的检测精度(AP)达到99.0%，较[图8(a)]Swin Transformer提高2.6个百分点。

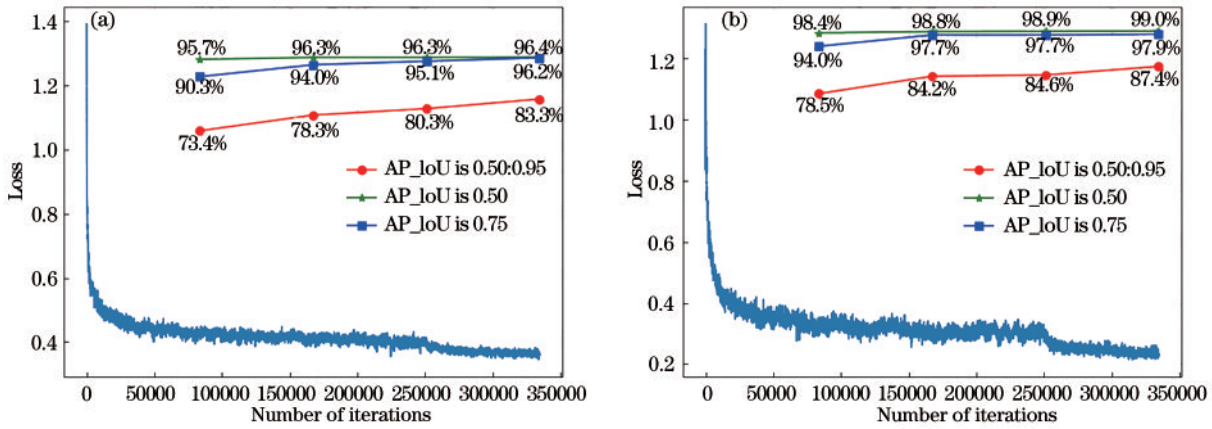

图8　Swin Transformer和所提模型训练过程。(a) Swin Transformer；(b)所提模型

Fig. 8　Training process of Swin Transformer and proposed model. (a) Swin Transformer; (b) proposed model

3.3　实验结果对比分析

为了验证所提基于Transformer车辆检测算法的有效性，分别与Swin-Transformer及常用的两种基于CNN的车辆检测算法Cascade R-CNN和YOLOV3展开了对比实验，使用公开车辆数据集UA-DETRAC(图9第1、2组)、VOC数据集中包含车辆类别的2000张图片及自行标定的100张在苏州昆山采集的远距离小目标车辆图片(图9第3组)对算法性能进行了测试。测试结果如表1所示，在AP方面，公开数据集UA-DETRAC和VOC-vehicle上的数据表明，所提算法比基于CNN的YOLOV3检测算法精度分别高出11.4个百分点和7.2个百分点，在小目标数据集Small-100中表现更为出色，精度高出了30.4个百分点，这得益于Transformer利用注意力的方式来捕获全局的上下文信息，从而建立远距离目标依赖，提取出更强有力的图像特征；此外，本实验组通过添加图像增强模块，将Swin-Transformer的检测精度在公开数据集UA-DETRAC和VOC-vehicle中分别提高了2.6个百分点和2.9个百分点，在小目标数据集Small-100中提高了7.9个百分点。检测速度方面，所提算法单张图片检测耗时68.4 ms，与Swin-Transformer相当，但与CNN方法还有一定差距，这是由于图像的信息量远大于文本数据，需要设计更加适合图像的Transformer结构来减少计算开销，这也是本课题接下来的研究方向。在实际车辆检测应用中，所提算法采用隔帧检测的方式依然可以达到实时的效果，检测速

表1　不同检测算法性能对比结果

Table 1　Comparison results of different detection algorithm performance

Testset	Model	Backbone	AP /%	Time /ms
UA-DETRAC	Cascade R-CNN	ResNet-101	79.8	48.8
UA-DETRAC	YOLOV3	Darknet-53	87.6	42.3
UA-DETRAC	Swin-Transformer	Swin-T	96.4	65.3
UA-DETRAC	Proposed model	Enhance-Swin-T	99.0	68.4
VOC-vehicle	Cascade R-CNN	ResNet-101	86.7	48.8
VOC-vehicle	YOLOV3	Darknet-53	90.9	42.3
VOC-vehicle	Swin-Transformer	Swin-T	95.2	65.3
VOC-vehicle	Proposed model	Enhance-Swin-T	98.1	68.4
Small-100	Cascade R-CNN	ResNet-101	49.3	48.8
Small-100	YOLOV3	Darknet-53	57.9	42.3
Small-100	Swin-Transformer	Swin-T	80.4	65.3
Small-100	Proposed model	Enhance-Swin-T	88.3	68.4

度可达 29.2 frame/s。

图像增强效果和检测效果如图 9 所示。[图 9(a)]表示在光照条件不佳的夜间摄像头拍到的图像，[图 9(b)]代表超分辨率重建后的图像，从[图 9(a)、(b)]左下与右上局部放大图可以看出，重建后的图像视觉效果明显增强。[图 9(c)]代表原图检测结果，左下侧圈井盖误检成了车辆，右上特征较少的车辆未检测出，[图 9(d)]代表超分辨率重建后使用 Transformer 方法检测结果的对比图，左下误检的井盖得到了纠正，令人惊讶的是，右上边缘的车辆也能准确地检测出来。

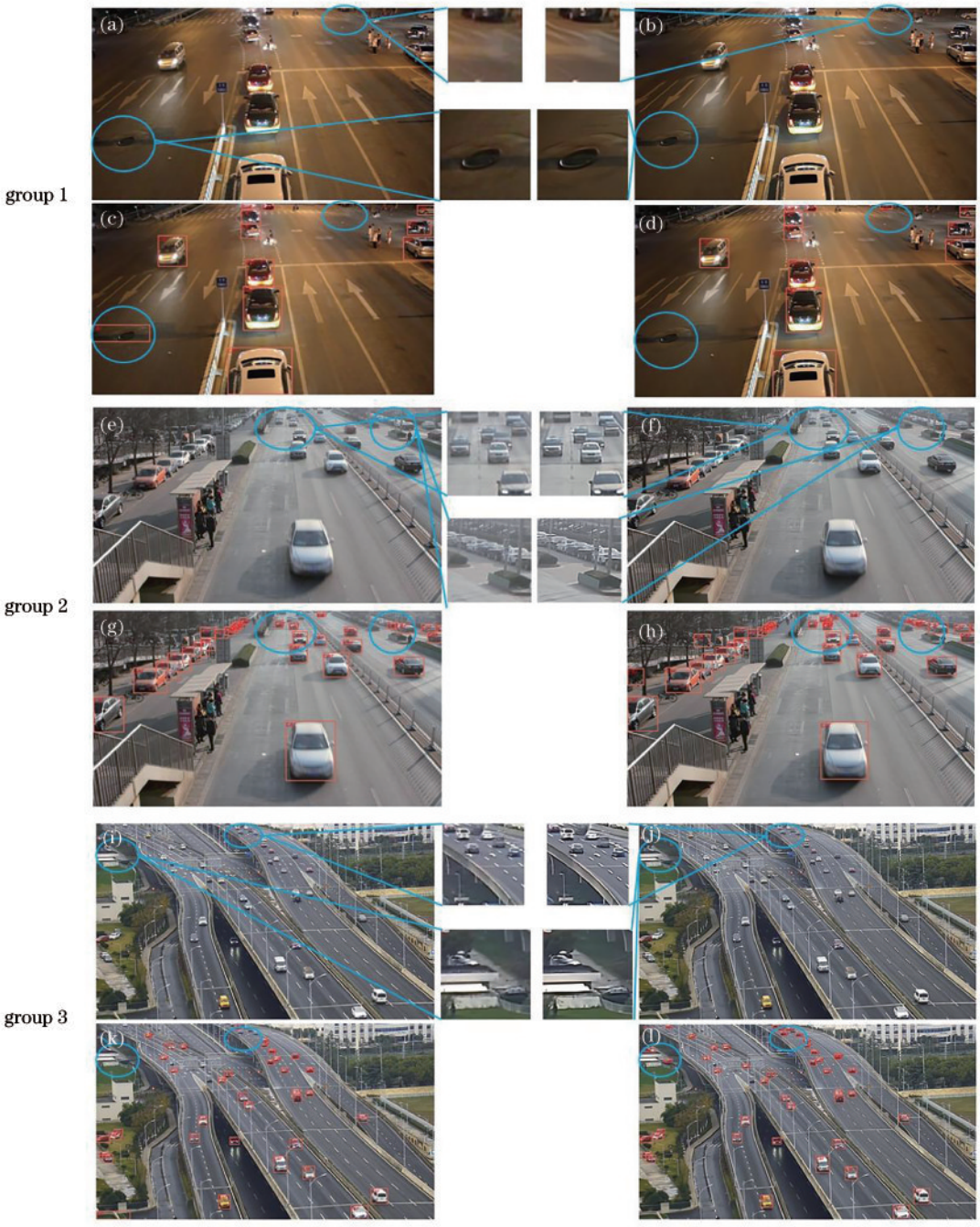

图 9　原始图像与图像增强后清晰度对比及检测效果对比。(a)(b)(c)(d)光照变化；(e)(f)(g)(h)模糊、密集小目标；(i)(j)(k)(l)远距离小目标

Fig. 9　Contrast of sharpness and detection effect between original image and image enhancement. (a) (b) (c) (d) Illumination change; (e) (f) (g) (h) blurred, dense small targets; (i) (j) (k) (l) long range small targets

[图 9(e)]表示拍到的包含远距离小尺度车辆的图像，[图 9(f)]代表超分辨率重建后的图像，从[图 9(e)、(f)]可知，左侧圆圈远距离模糊的小尺度车辆、右侧圆圈密集的停车区小目标车辆在重建后特征更加明显，对比原图与重建后图的检测结果[图 9(g)、(h)]可知，模糊车辆、路边密集停车区域检测召回率有了极大提升；[图 9(i)]代表摄像头距离公路较远时采集的图像，对于远距离的小目标，超分辨率重建之后的图像[图 9(j)]

视觉效果明显增强。从[图 9(k)、(l)]中的局部检测图可以看出,在肉眼难以分辨的情况下,所提方法在远距离小目标上表现出了极好的性能,漏检率大幅降低。

4 结 论

针对目前复杂场景下车辆检测算法对远距离和遮挡情况小目标检测准确率低的情况,提出了一种基于GAN的融合边缘增强和超分辨率重建的交通图像增强算法及一种基于视觉Transformer的车辆检测算法。实验结果表明,所采用的对交通道路上原始低分辨率图像进行超分辨率重建和边缘增强后再运用具有移位窗口的层级视觉Transformer进行车辆检测的检测效果极佳,在公开车辆数据集UA-DETRAC中的检测精度达到99%,减少了车辆检测中小目标漏检、误检的情况,提高了车辆检测的检测距离和检测精度。本研究具有一定的理论研究意义和实际意义。同时,研究的相关内容还可以拓展至其他应用领域,例如行人检测、机场安检、军事边防预警等领域。

参 考 文 献

[1] 李汉冰, 徐春阳, 胡超超. 基于YOLOV3改进的实时车辆检测方法[J]. 激光与光电子学进展, 2020, 57(10): 101507.
Li H B, Xu C Y, Hu C C. Improved real-time vehicle detection method based on YOLOV3[J]. Laser & Optoelectronics Progress, 2020, 57(10): 101507.

[2] 段仲静, 李少波, 胡建军, 等. 深度学习目标检测方法及其主流框架综述[J]. 激光与光电子学进展, 2020, 57(12): 120005.
Duan Z J, Li S B, Hu J J, et al. Review of deep learning based object detection methods and their mainstream frameworks[J]. Laser & Optoelectronics Progress, 2020, 57(12): 120005.

[3] Cai Z W, Fan Q F, Feris R S, et al. A unified multi-scale deep convolutional neural network for fast object detection[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9908: 354-370.

[4] Hu X W, Xu X M, Xiao Y J, et al. SINet: a scale-insensitive convolutional neural network for fast vehicle detection[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 20(3): 1010-1019.

[5] Kampffmeyer M, Salberg A B, Jenssen R. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops, June 26-July 1, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 680-688.

[6] 刘峰, 郭猛, 王向军. 基于跨尺度融合的卷积神经网络小目标检测[J]. 激光与光电子学进展, 2021, 58(6): 0610012.
Liu F, Guo M, Wang X J. Small target detection based on cross-scale fusion convolution neural network[J]. Laser & Optoelectronics Progress, 2021, 58(6): 0610012.

[7] 刘力荣, 唐新明, 赵文吉, 等. 基于影像与激光数据的小交标检测与地理定位[J]. 中国激光, 2020, 47(9): 0910002.
Liu L R, Tang X M, Zhao W J, et al. Detection and geo-localization of small traffic signs based on images and laser data[J]. Chinese Journal of Lasers, 2020, 47(9): 0910002.

[8] Ren J, Chen X H, Liu J B, et al. Accurate single stage detector using recurrent rolling convolution[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 752-760.

[9] Takeki A, Trinh T T, Yoshihashi R, et al. Combining deep features for object detection at various scales: finding small birds in landscape images[J]. IPSJ Transactions on Computer Vision and Applications, 2016, 8(1): 5.

[10] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale[EB/OL]. (2020-10-22) [2021-05-04]. https://arxiv.org/abs/2010.11929.

[11] Rabbi J, Ray N, Schubert M, et al. Small-object detection in remote sensing images with end-to-end edge-enhanced GAN and object detector network[J]. Remote Sensing, 2020, 12(9): 1432.

[12] Wang X T, Yu K, Wu S X, et al. ESRGAN: enhanced super-resolution generative adversarial networks[M]// Leal-Taixé L, Roth S. Computer vision-ECCV 2018 workshops. Lecture notes in computer science. Cham: Springer, 2019, 11133: 63-79.

[13] Liu Z, Lin Y T, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]// 2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2021: 9992-10002.

[14] Cai Z W, Vasconcelos N. Cascade R-CNN: delving into high quality object detection[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: 6154-6162.