基于改进 SSD 的交通大场景多目标检测

华夏[1*]，王新晴[1]，王东[1,2]，马昭烨[1]，邵发明[1]

[1]中国人民解放军陆军工程大学野战工程学院，江苏 南京 210007；

[2]南部战区陆军第二工程科研设计所，云南 昆明 650222

摘要 现有目标检测算法在复杂大场景下多目标检测的精度和实时性难以平衡，为此，受深度神经网络卷积核形态启发，模仿了人眼视觉机理，改进了基于深度学习的目标检测框架，即单向多框检测器(SSD)，提出了多目标检测框架——自适应感知 SSD，将其专用于复杂大交通场景多目标检测。设计了由多形态、彩色 Gabor 构成的特征卷积核库，训练筛选最优特征提取卷积核组替换原有网络的低级卷积核组，从而提高检测精度；将单图像检测框架与卷积长短期记忆网络结合，通过瓶颈-长短期记忆层提炼传播帧间的特征映射，实现网络帧级信息的时序关联，降低计算成本，从而实现对视频中受强干扰影响目标的追踪识别；同时加入自适应阈值策略，降低漏警率和虚警率。实验结果表明，相比于其他基于深度学习的目标检测框架，各类目标识别的平均准确率提高了 9%～16%，平均准确率均值提高了 14%～21%，多目标检测率提高了 21%～36%，检测帧率达到 32 frame·s^{-1}，实现了算法精度与实时性的平衡，取得较好的检测识别效果。

关键词 机器视觉；生物视觉；深度学习；卷积神经网络；Gabor 卷积核；递归神经网络

中图分类号 O436 **文献标识码** A **doi**：10.3788/AOS201838.1215003

Multi-Objective Detection of Traffic Scenes Based on Improved SSD

Hua Xia[1*], Wang Xinqing[1], Wang Dong[1,2], Ma Zhaoye[1], Shao Faming[1]

[1] *College of Field Engineering, PLA Army Engineering University, Nanjing, Jiangsu* 210007, *China*;

[2] *Second Institute of Engineering Research and Design, Southern Theatre Command, Kunming, Yunnan* 650222, *China*

Abstract Aiming at the problem that the accuracy and real-time of multi-target detection in complex and large scenes are difficult to balance in the existing target detection algorithms, we imitate the human visual mechanism inspired by the convolution kernel shape of the deep neural network. The target detection framework——the single shot multi-box detection (SSD) based on deep learning is improved, and a multi-target detection framework adaptive perceive SSD is proposed, which is specially used for the multi-target detection in complex and large traffic scenes. A feature convolution kernel library composed of multi-form Gabor and color Gabor is designed. The optimal feature extraction convolution kernel group is trained and screened to replace the low-level convolution kernel group of the original network, and effectively improves the detection accuracy. A single image detection framework is combined with a convolution long-short-term memory network, and the temporal association of network frame-level information is realized by extracting the characteristic mapping between propagation frames with a bottleneck-long-term and short-term memory layer. And the calculation cost is reduced, and the tracking and identification of targets affected by the strong interference in the video are realized. An adaptive threshold strategy is added to reduce the rate of missing and false alarms. The experimental results show that compared with other target detection frameworks based on deep learning, the average accuracy of various target recognition is increased by 9%～16%, the average accuracy is increased by 14%～21%, the multi-target detection rate is increased by 21%～36%, and the detection frame rate reaches 32 frame·s^{-1}, which achieves a balance between the accuracy and real-time performance of the algorithm and achieves better detection and recognition results.

Key words machine vision; biological vision; deep learning; convolution neural network; Gabor convolution kernel; recurrent neural network

OCIS codes 150.0155; 150.1135; 100.4996

收稿日期：2018-05-22；**修回日期**：2018-06-26；**录用日期**：2018-07-25

基金项目：国家重点研发计划(2016YFC0802904)、国家自然科学基金(61671470)、江苏省自然科学基金(BK20161470)、中国博士后科学基金第 62 批面上资助项目(2017M623423)

* **E-mail**：1614118084@qq.com

1　引　言

交通场景中的行人、车辆目标检测与识别是目标检测技术的重要分支，也是自动驾驶、机器人以及智能视频监控等研究领域的核心技术，具有重要的研究意义[1]。深度学习为基于深层人工神经网络的学习方法[2]，在人工神经网络结构中，深度卷积网络具有强大的特征提取能力，广泛应用于图像分类，并在图像识别、图像分割、目标检测、场景分类等视觉任务中，取得了非常好的效果[3-5]。

单向多框检测器(SSD)是 Liu 等[6]提出的目标检测算法，也是主要检测框架之一，相比更快的区域卷积神经网络(Faster RCNN)[7]具有明显的速度优势，相比 YOLO(You Only Look Once)[8-9]又具有明显的平均准确率均值(mAP)优势。尽管 SSD 在特定数据集上已经取得了较高的准确率，具有较好的实时性，但是模型的训练过程非常耗时，严重依赖训练样本的质和量。并且通过图像的颜色、边缘等信息来检测目标，对于弱小目标和大面积遮挡目标等缺乏图像信息的目标检测效果不佳。因此，该算法检测效率仍然有待提高，以满足工程实际应用对于实时性的要求。

综上，本文提出一种新的多目标检测框架自适应感知 SSD，主要对传统 SSD 算法进行了以下改进：1)设计构造了由多形态、彩色 Gabor 构成的特征卷积核库，通过训练筛选得到最优特征提取卷积核组以替换原有特征提取网络用于区域基础特征提取的低级卷积核组，得到新的特征提取网络 Gabor-VGGnet，大幅提高了检测精度；2)采用模糊阈值法调整自适应阈值策略降低漏警率和虚警率，在避免适应数据集的同时提高模型的决策能力；3)将单图像多目标检测框架与卷积长短期记忆网络(LSTM)相结合，形成交织循环卷积结构，实现了网络帧级信息的时序关联，极大降低了网络计算成本；4)利用 LSTM 的时序关联特性，结合动态卡尔曼滤波算法，实现对视频中受光照变化、大面积遮挡等强干扰影响目标的追踪识别。

2　本文模型

本文检测算法整体框架如图 1 所示，由 LSTM 和动态卡尔曼滤波(图 1 中绿色框)、AP-Gabor SSD(图 1 中红色框)组成。首先将单帧视频图像输入 AP-Gabor SSD，再结合 LSTM 网络传递的预测各层特征映射进行目标检测识别，获得初检测结果 R_1；2)通过 LSTM 网络传递获得当前帧的预测检测结果 R_2，采用动态卡尔曼滤波将初检测结果 R_1 和预测检测结果 R_2 结合，获得最终的检测识别结果 R_3；3)将当前帧检测过程中产生的各层特征映射以及检测结果 R_3 输入 LSTM 网络，便于对下一帧的检测结果进行指导。

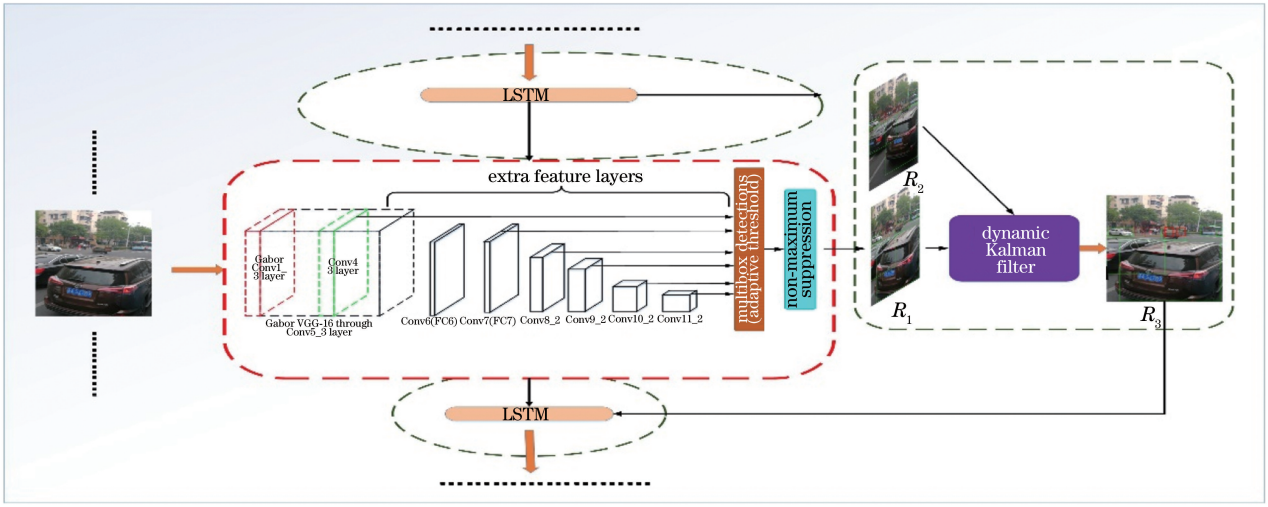

图 1　改进后检测算法整体框架

Fig. 1　Improved detection algorithm overall framework

3　改进的特征提取网络 Gabor-VGGnet

3.1　仿光感细胞的 Gabor 卷积核设计

训练深度卷积神经网络(CNN)的某一个卷积层实际上是在训练一系列的滤波器，让这些滤波器对特定的目标有高的敏感激活度，以达到深度 CNN 的识别、检测等目的。在训练开始之时，卷积层的滤波器是完全随机的，它们不会对任何特征激活即不

能检测任何特征[10-11]。通过深度 CNN 可视化工具箱（yosinski/deep-visualization-toolbox）[12]，对 CNN 模型进行可视化得到的各级特征卷积核示例如图 2 所示。

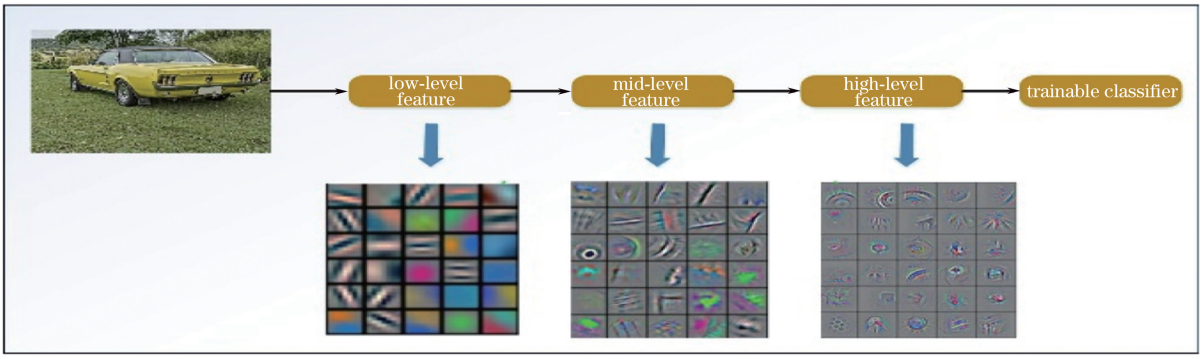

图 2　CNN 模型提取各级特征卷积核示例

Fig. 2　Example of CNN model extracting feature convolution kernels at various levels

人眼视网膜上主要光感受器为视杆细胞和视锥细胞。视锥细胞主司昼光觉，有色觉，光敏感性差，但视敏度高。视杆细胞对暗光敏感，光敏感度较高，视物无色觉[13-14]。Gabor 小波与人类视觉系统中简单细胞的视觉刺激响应非常相似[15]。通过对比分析发现，深度 CNN 通过训练获取的提取基础特征的卷积核在形态上与 Gabor 的卷积核存在极大的相似度。用二维 Gabor 卷积核模拟视杆细胞功能，用彩色 Gabor 卷积核模拟视锥细胞功能。在空域，一个二维的 Gabor 滤波器是一个正弦平面波和高斯核函数的乘积，二维 Gabor 函数的数学表达式为

$$g(x,y,\lambda_1,\theta_1,\psi_1,\sigma_2,\gamma_1)=\exp\left(-\frac{x'^2+\gamma_1^2y'^2}{2\sigma_2^2}\right)\exp\left[\mathrm{i}\left(2\pi\frac{x'}{\lambda_1}+\psi_1\right)\right],\tag{1}$$

其中其实部与虚部分别为

$$g_{\mathrm{real}}(x,y,\lambda_1,\theta_1,\psi_1,\sigma_2,\gamma_1)=\exp\left(-\frac{x'^2+\gamma_1^2y'^2}{2\sigma_2^2}\right)\cos\left(2\pi\frac{x'}{\lambda_1}+\psi_1\right),\tag{2}$$

$$g_{\mathrm{imag}}(x,y,\lambda_1,\theta_1,\psi_1,\sigma_2,\gamma_1)=\exp\left(-\frac{x'^2+\gamma_1^2y'^2}{2\sigma_2^2}\right)\sin\left(2\pi\frac{x'}{\lambda_1}+\psi_1\right),\tag{3}$$

式中：x、y 代表像素点坐标，$x'=x\cos\theta+y\sin\theta$，$y'=-x\sin\theta+y\cos\theta$；$\lambda_1$ 表示正弦函数波长；θ_1 表示核函数的方向；ψ_1 表示相位偏移；σ_2 表示高斯函数的标准差；γ_1 表示空间的宽高比。实部可以对图像进行平滑滤波，虚部可以用来边缘检测。通过实验发现，Gabor 滤波器卷积核的形态对 Gabor 滤波器边缘增强的对象和效果具有决定性的影响。不同结构类型 Gabor 滤波器对与其尺度、方向、中心位置、相位、结构类型相一致的图像内容形成最优响应。

为了让 Gabor 滤波器能够提取更加复杂、丰富的边缘和纹理特征信息，引入了参数 k_1、k_2、k_3、k_4、k_5 对 Gabor 卷积核实部进行调整，即

$$g_{\mathrm{real}}(x,y,\lambda_1,\theta_1,\psi_1,\sigma_2,\gamma_1,k_1,\cdots,k_5)=\exp\left(-\frac{x'^2+\gamma_1^2y'^2}{2\sigma_2^2}\right)\cos\left[2\pi\frac{(k_1\cdot x'^{k_2}+y'^{k_3}+k_4)^{k_5}}{\lambda_1}+\psi_1\right]。\tag{4}$$

图 3 是由(4)式构造的部分二维 Gabor 滤波器卷积核，参数 k_2、k_3、k_5 决定了卷积核的结构类型，参数 k_1、k_4 决定 Gabor 滤波器卷积核的方向与相位，从而实现更加复杂、丰富的边缘和纹理特征信息的提取。

受深度 CNN 训练获得的彩色卷积核启发，以神经网络训练得到的彩色卷积核为参考，通过重构的方式构造了三维彩色 Gabor 滤波器，用于对彩色图像颜色特征的激活[17]。RGB 空间使用红、绿、蓝三原色的亮度来定量表示颜色，是以 R（红）、

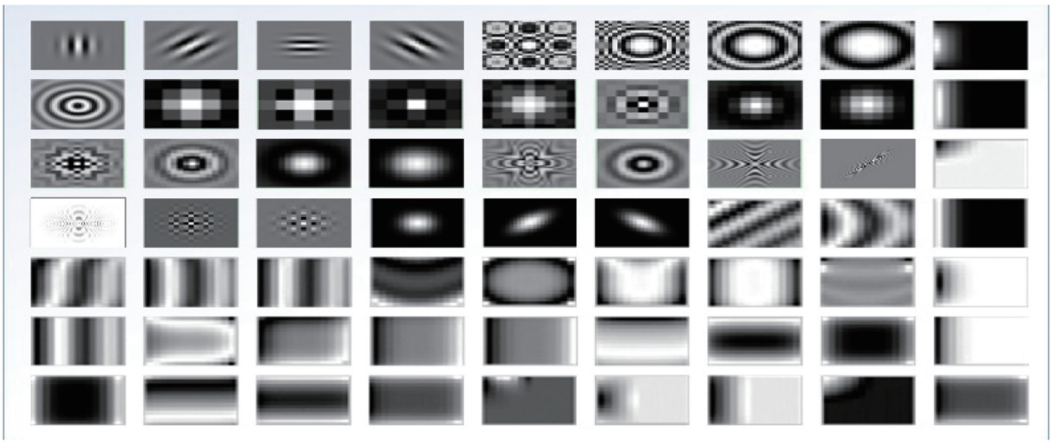

图 3　二维 Gabor 滤波器卷积核

Fig. 3　Two-dimensional Gabor filter convolution kernel

G(绿)、B(蓝)三色光互相叠加来实现混色的方式。模仿人眼的视觉机理，将一个二维的 Gabor 滤波器视为对三维颜色空间的一个颜色分量进行颜色特征检测的滤波器，依据需要提取的目标颜色特性，构造三个颜色分量的相互关系，分别得到其他两个颜色分量的 Gabor 滤波器，将这三个二维滤波器通过合成即可获得用于提取指定目标颜色特征的彩色 Gabor 滤波器，即

$$G_c = G_R + G_G + G_B, \tag{5}$$

式中 G_R 代表彩色 Gabor 在 R 颜色通道上的二维 Gabor 滤波器，G_R 卷积核的形态。由(5)式所确定。G_G、G_B 分别为彩色 Gabor 在 G、B 颜色通道上的二维 Gabor 滤波器。

在经典的感受野中，包含有红、绿、蓝、黄 4 个分量，拥有 4 种感受野[17]。为了模仿人眼视觉细胞对颜色的感知，以神经网络训练得到的彩色卷积核为参考，通过模仿重构的方式，总结出各颜色通道之间的数学关系为

$$G_G = \begin{cases} 255 - G_R, \text{R\&G} \quad \text{or} \quad \text{Y\&B} \\ G_R, \text{R\&G\&B\&Y} \quad \text{or} \quad \text{R\&B} \end{cases}, \tag{6}$$

$$G_B = \begin{cases} 255 - G_R, \text{R\&G\&B\&Y} \\ G_R, \text{R\&G} \\ 255 - G_G, \text{Y\&B} \\ G_G, \text{G\&B} \end{cases}。 \tag{7}$$

约束条件为彩色 Gabor 敏感的目标颜色，例如 R&G 表示目标主色为红色和绿色，Y 代表黄色。部分彩色 Gabor 效果如图 4 所示。

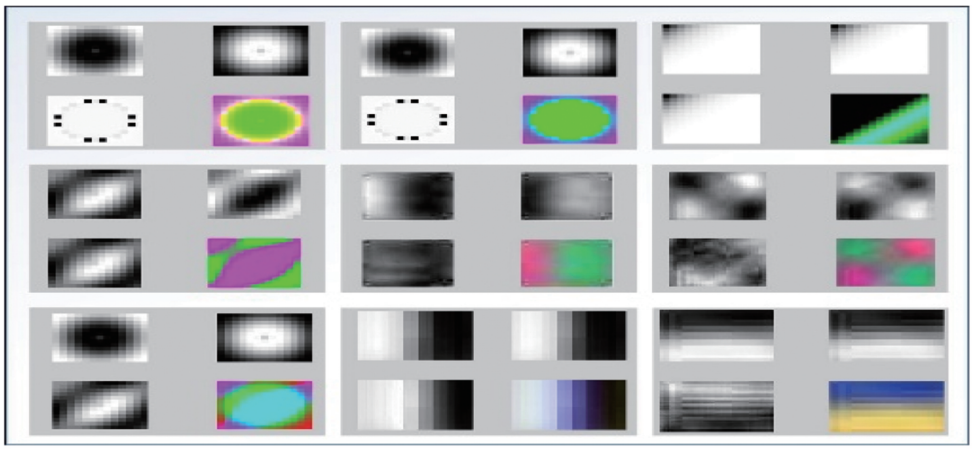

图 4　三维 Gabor 滤波器卷积核

Fig. 4　Three-dimensional Gabor filter convolution kernel

3.2　智能优化最优 Gabor 卷积核组筛选

在实验中用 SSD512 模型结合训练数据集中的部分车辆目标数据集训练并测试了几种不同卷积核个数变化对目标识别率的影响，分析实验结果，为了保证尽可能高的检测精度，选取卷积层深度依次为 128、256、384、384、384。人的视网膜中，视锥细胞数

量约为 600～800 万，视杆细胞总数达 1 亿以上，两者的比例近似为 10∶1，由此本文设计第一层 Gabor 卷积核组中二维 Gabor 卷积核数量为 110，彩色 Gabor 卷积核数量为 18。用 KITTI 数据集中的车辆目标数据集训练并测试了几种不同卷积核尺寸变化对识别率的影响，通过实验总结，发现二维 Gabor 滤波器卷积核取 3×3 大小，彩色 Gabor 滤波器卷积核取 5×5，并往网络的 Inception 结构中加入 1×1 的卷积核进行降维时，组合的滤波器组能够获得更佳的目标特征敏感。为了能够有效提高算法整体的检测精度，构造合理的 Gabor 特征提取卷积核组提取具有区分度的多特征具有重要意义。最优 Gabor 卷积核组的筛选流程如图 5 所示。

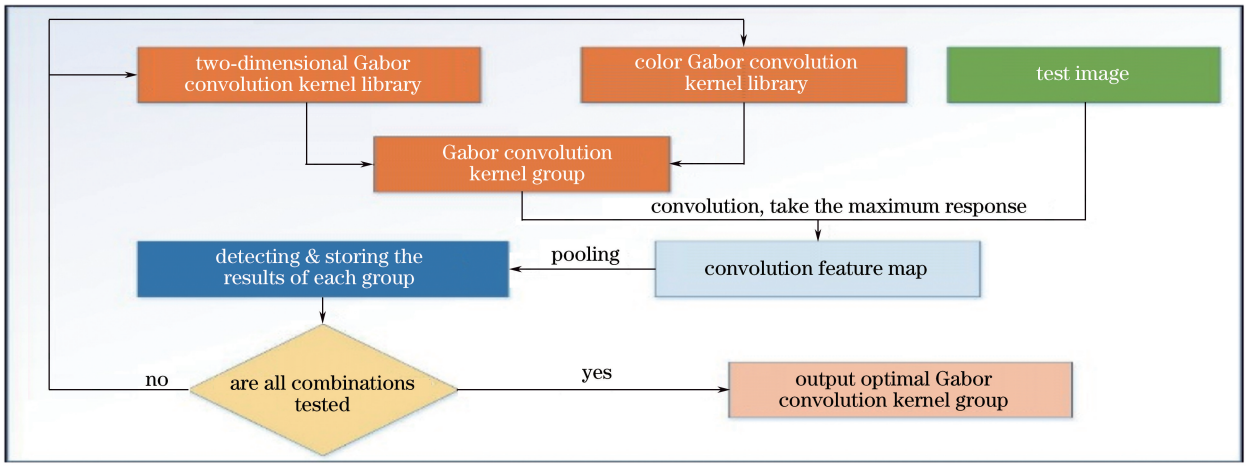

图 5　最优 Gabor 卷积核组的训练流程

Fig. 5　Training process for optimal Gabor convolution kernel group

首先由(1)式和(4)式，通过变换参数的方式构造一个包含多种形态的二维 Gabor 库，由(6)式和(7)式可以构造一个同等规模的彩色 Gabor 库，再构造分别只单独含有“人”、“骑行者”、“车辆”的小规模测试图像集(三个目标各 20 张，共 60 张)。从两个 Gabor 库中各随机不重复抽取卷积核，组成卷积核组，每个卷积核组对测试集的图像逐张进行卷积，通过非极大值抑制获得对应的特征映射，将特征映射经过池化转换为特征向量，输入通过小样本数据训练好的传统 SSD 检测框架中全连接层即 Softmax 分类器，可获得测试图像目标的检测置信度，将测试集全部置信度的均值作为该卷积核组特征提取有效性的评价分数，取最高评价分数对应的卷积核组作为最佳卷积核组。

Gabor 库的规模，依据实际需求的卷积核组中卷积核的个数来合理确定。为了避免因为组合过多造成的数据爆炸，以及数据库规模太小造成的特征提取不全面，在进行卷积核抽选时，将二维 Gabor 卷积核每 10 个随机组成一组，将彩色 Gabor 卷积核每 18 个随机组成一组，以组为单位进行组合，构造的 Gabor 库规模为 180 个卷积核。

4　置信度自适应阈值判定

在 SSD 用 Softmax 为候选区域进行分类的最后阶段，候选区域会得到属于各个类别的置信度(即属于各个类别的概率)，当属于某类的置信度高于设定阈值时则将此候选区域判为该类目标，若同一候选区域有多个类别置信度高于阈值则取最高者。针对 SSD 检测固定置信度阈值不够灵活的缺陷，采用模糊自适应阈值法调整自适应阈值策略降低漏警率和虚警率。

模糊程度是由模糊率函数来确定的，当模糊率最低的时候，这时候分割效果最好。其中模糊率与隶属函数相关，模糊数学的基本思想是隶属度的思想[18]。检测一张图像默认得到 N 个候选区域送入 SSD，最后每个候选区域都得到 M 个用来表示属于 M 个类别的置信度，故共可以得到 N 个 $M\times1$ 的数组。取出每个数组中的最大值并由大到小排序，舍去其中小于 0.1 的值(若 N 个值全部小于 0.1，则判为没有目标)，得到 $N\times1$ 的数组 $\boldsymbol{C}$。$\mu(x)$ 是隶属度函数，$\mu(C_k)$ 为数组 $\boldsymbol{C}$ 中置信度取 C_k 的区域的隶属度。数组 $\boldsymbol{C}$ 的模糊率 $\gamma(\boldsymbol{C})$ 是对数组 $\boldsymbol{C}$ 的模糊性度量，令 $h(C_k)$ 为数组 $\boldsymbol{C}$ 中置信度取 C_k 的元素个数，则数组 $\boldsymbol{C}$ 的模糊率 $\gamma(\boldsymbol{C})$ 定义为

$$\gamma(\boldsymbol{C})=\frac{2}{n}\sum_{k=0}^{n-1}T(C_k)h(C_k),\tag{8}$$

式中 $T(C_k)=\min\{\mu(C_k),1-\mu(C_k)\}$。数组 $\boldsymbol{C}$ 的

模糊率 $\gamma(\boldsymbol{C})$ 取决于隶属度函数 $\mu(x)$，若取隶属度函数为 S 函数，即

$$\mu(x)=\begin{cases}0, & 0\leqslant x\leqslant q-\Delta q\\ 2\left[\dfrac{(x-q+\Delta q)}{2\Delta q}\right]^2, & q-\Delta q\leqslant x\leqslant q\\ 1-2\left[\dfrac{(x-q+\Delta q)}{2\Delta q}\right]^2, & q<x\leqslant q+\Delta q\\ 1, & q+\Delta q<x\leqslant C_n\end{cases}。\tag{9}$$

此时 $\mu(x)$ 由窗宽 $c=2\Delta q$ 和参数 q 决定，一旦选定了窗宽，则 $\gamma(\boldsymbol{C})$ 就只与参数 q 有关。模糊阈值法的求解过程是预先设定窗宽，系数常设定为0.3。改变 q 使得隶属度函数 $\mu(x)$ 在置信度区间 $[C_0, C_{n-1}]$ 上滑动，通过计算模糊率 $\gamma_q(\boldsymbol{C})$ 获得模糊率曲线，该曲线的谷点即为 $\gamma_q(\boldsymbol{C})$ 取得极小值的 q，也就是自适应阈值。

5　基于时间感知特征映射的视频目标检测

人类的思想具有连贯性，但传统的神经网络无法做到，然而递归神经网络（RNN）较好地解决了这个问题。LSTM 是一种特殊的 RNN，可以解决长期依赖的问题。

研究了在保证运行速度和低运算资源消耗的前提下，通过增加时间感知来构建视频检测模型的策略，在最终检测结果和特征空间中添加时间感知机制，通过递归网络体系结构将每个帧的特征映射调整到先前帧的相应特征映射上来利用特征级的连续性。提出了一种将卷积 LSTM 结合到单图像检测框架中的方法，将其作为跨时间传播帧级信息的手段，网络结构如图 6 所示。

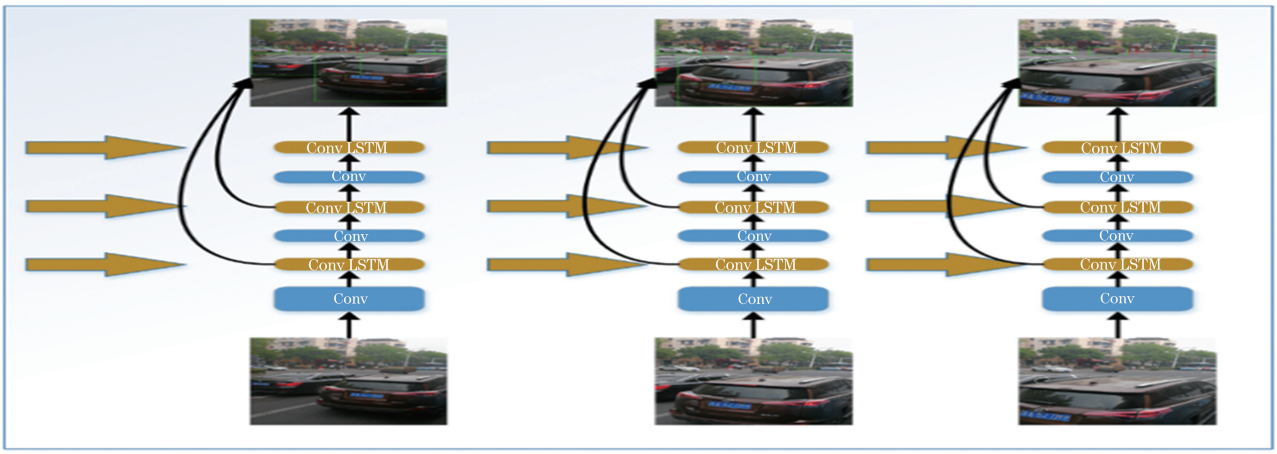

图 6　基于时间感知特征映射的移动视频目标检测框架

Fig. 6　Mobile video target detection framework based on time-aware feature mapping

网络当中的某层 Conv LSTM 接收了上一帧对应位置的 Conv LSTM 传递的特征映射和当前帧前一层卷积层传递的特征映射后对检测结果进行预测，并把特征映射继续传递给下一层卷积层和下一帧对应位置的 Conv LSTM，Conv LSTM 的输出将在以后的所有计算中替换之前的特征映射，继续执行检测任务。然而，LSTM 的简单集成会导致较大的运算量，妨碍网络实时运行。为了解决这个问题，引入了一个 Bottleneck-LSTM[19]，利用它具有深度可分离卷积和 Bottleneck 设计原则的特性，以降低计算成本。

视频数据可以视为多帧图像组成的序列，$V=\{I_0, I_1, \cdots, I_n\}$，目标是得到帧级的检测结果 $\{D_0, D_1, \cdots, D_n\}$，其中 D_k 表示对图像帧 I_k 的检测结果，包括各个目标检测框的位置，以及各个目标识别置信度。考虑构造一种在线学习机构，使得检测结果 D_k 可以由图像帧 I_{k-1} 进行预测和修正。将预测模型当做函数，即

$$F(I_t, \boldsymbol{s}_{t-1})=(D_t, \boldsymbol{s}_t),\tag{10}$$

式中 $\boldsymbol{s}_k=\{s_k^0, s_k^1, s_k^2, \cdots, s_k^{m-1}\}$，表示描述视频第 k 帧图像的特征映射向量，构造一个具有 m 层 LSTM 卷积层的神经网络来近似地实现这个函数功能。这个神经网络把特征映射向量 $\boldsymbol{s}_{t-1}$ 中的每个特征映射作为 LSTM 卷积层的输入，可以得到对应的特征映射向量 $\boldsymbol{s}_t$。因此，要获得整个视频的检测结果，只需通过网络顺序运行每帧图像。

将单帧图像目标检测器定义为函数 $G(I_t)=D_t$，该函数用于构造具有 m 个 LSTM 层的复合网络。再将 LSTM 卷积层看作是函数 G 划分为 $m+1$ 个合适的子网络 $\{g_0, g_1, \cdots, g_m\}$，则

$$G(I_t)=(g_m\circ\cdots,g_1\circ g_0)I_t,\tag{11}$$

式中∘表示哈达玛乘积。同样将任意一层 LSTM 卷积层定义为函数，即

$$L_k(M,s_{t-1}^k)=(M_+,s_t^k),\tag{12}$$

式中 M、M_+ 都是同维度的特征映射。则按照时序进行计算，即

$$\begin{cases}(M_+^0,s_t^0)=L_0[g_0(I_t),s_{t-1}^0]\\(M_+^1,s_t^1)=L_1[g_1(M_+^0),s_{t-1}^1]\\\quad\vdots\\(M_+^{m-1},s_t^{m-1})=L_{m-1}[g_{m-1}(M_+^{m-2}),s_{t-1}^{m-1}]\\D_t=g_m(M_+^{m-1})\end{cases}。\tag{13}$$

图 7 所示为整个模型在处理视频时的输入和输出。

由于需要在单个前向通道中计算多个门，所以 LSTM 对计算资源要求较高，这极大地影响了网络的整体效率。为了解决这个问题，首先，调整 LSTM 的维度，扩展文献[19]中定义的通道宽度乘子 α_δ，可以更好地控制网络结构。引入了三个新的参数 α_{base}、α_{SSD}、α_{LSTM}，分别控制网络不同部分的信道尺寸。具有 N 个输出通道的基本移动网络中的任何给定层被修改为具有 $N_{\alpha\,\text{base}}$ 个基本输出通道，而 α_{SSD} 应用于所有 SSD 特征映射，α_{LSTM} 应用于 LSTM 层。对于所提出的网络，设置 $\alpha_{\text{base}}=\alpha$，$\alpha_{\text{SSD}}=0.5\alpha$，$\alpha_{\text{LSTM}}=0.25\alpha$。每个 LSTM 的输出均为输入大小的 1/4，大大减少了所需的计算量。

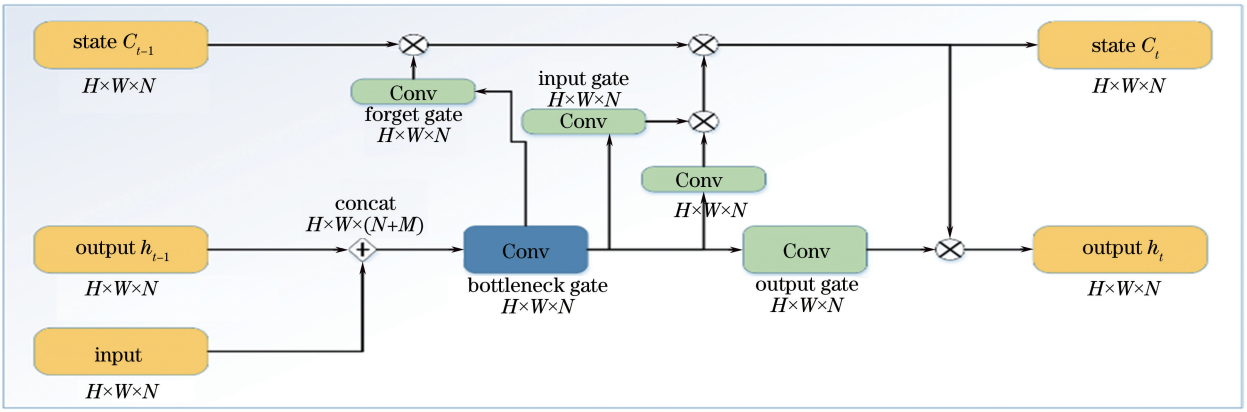

图 7　模型在处理视频输入和输出示意图

Fig. 7　Model processing video input and output schematics

同时采用 Bottleneck-LSTM 提高传统 LSTM 的运算效率，即

$$b_t=\phi({}^{M+N}W_b^N*[x_t,h_{t-1}]),\tag{14}$$

式中：x_t、h_{t-1} 为输入的特征映射；$\phi(x)=\text{ReLU}(x)$，修正线性单位(ReLU)表示稀疏激活；${}^jW^k*X$ 表示具有权重 W、输入 X、j 输入通道和 k 输出通道的深度可分离卷积。使用瓶颈特征映射减少门内的计算量，在所有实际场景中均优于标准 LSTM。

训练有素的 CNN 无法应对大面积遮挡等强干扰造成目标图像信息严重缺失。对此本文从之前的检测结果中获取有用的先验信息来合理预测少量候选区域，增加目标被检测的几率。因此选择卡尔曼滤波[20]作为前一帧和当前帧之间传递目标信息的工具，结合目标检测任务设计卡尔曼滤波模型。$D_k=\{X_k^0,X_k^1,\cdots,X_k^n\}$ 表示使用未加入滤波的检测器对图像帧 I_k 的检测结果，其中 $X_k^t=[x_k^t,y_k^t,a_k^t,b_k^t,c_k^t,d_k^t]$，$x$、$y$、$a$、$b$ 分别为第 k 帧某一目标 t 外接矩形框的左上角坐标和宽、高，c 为目标置信度，d 为目标所属类别。通过 LSTM 可以获得视频第 $k+1$ 帧的检测结果 D_{K+1} 的预测值 $\hat{D}'_{k+1}$。由于预测过程中存在噪声等因素干扰而产生误差，如果不对预测结果加以修正，那么在视频检测的过程中误差将因为迭代过程而被无限地放大。为了避免出现这种情况，将视频第 $k+1$ 帧的初检测结果 z_{k+1} 作为测量值对 LSTM 的预测值 $\hat{D}'_{k+1}$ 进行修正，即采用“预测＋测量反馈”的方式获得视频第 $k+1$ 帧的检测结果 D_{K+1} 的估计值 $\hat{D}'_{k+1}$。则系统的估计值滤波方程为

$$\hat{X}_{k+1}^t=A_k\hat{X}_k^{t\prime}+K_{k+1}(Z_{k+1}^t-H_{k+1}A_k\hat{X}_k^{t\prime})。\tag{15}$$

系统的测量方程为 $Z_{k+1}^t=\boldsymbol{H}X_{k+1}^t+\boldsymbol{v}_{k+1}$，卡尔曼增益方程为 $K_{k+1}=\boldsymbol{P}_{k+1/k}\boldsymbol{H}^{\mathrm{T}}(\boldsymbol{H}\boldsymbol{P}_{k+1/k}\boldsymbol{H}^{\mathrm{T}}+v_{k+1})^{-1}$。预测误差协方差矩阵方程为 $\boldsymbol{P}_{k+1/k}=\boldsymbol{A}\boldsymbol{P}_k\boldsymbol{A}^{\mathrm{T}}+w_k$，修正误差协方差矩阵方程为 $\boldsymbol{P}_{k+1}=$

$(I-K_{k+1}\boldsymbol{H})\boldsymbol{P}_{k+1/k}$，其中 $\boldsymbol{A}$ 为状态转移矩阵，$\boldsymbol{H}$ 为观测矩阵，w_k 为状态噪声，v_k 为观测噪声，均为高斯白噪声。

6　实验分析与讨论

6.1　实验的条件与数据集

本文实验使用 DELL Precision R7910 (AWR7910)图形工作站，处理器为 Intel Xeon E5-2603 v2(1.8 GHz/10M)，采用 NVIDIA Quadro K620 GPU 加速运算。SSD 基于深度学习框架 Caffe 运行。本文在 YFCC100M 收集的交通场景数据集(WD)和 KITTI 数据集上进行了实验。选用 KITTI 数据集中第一个图片集 Download left color images of object data set 和标注文件 Download training labels of object data set，实验数据集设置三个类别分别为 Car、Cyclist、Pedestrian。YFCC100M 数据集包含大约 1 亿张图片以及摘要、标题和标签。为了更好地展示本文方法的效果，通过搜索关键词“行人”、“道路”和“车辆”从 YFCC100M 数据集收集了 1000 幅分辨率较高的测试图像。对于该数据集，使用至少 16 pixel 宽度和小于 50%遮挡对所有目标进行注释。图像在较长的一侧被重新缩放到 2000 pixel，以适合 GPU 内存。

6.2　实验的参数设置

对 SSD 系列中的 SSD512 进行改进。为了优化调参过程以及快速选取自适应池化纠正误差项的最佳值，制作了小样本数据集(200 张图像)，大幅节约了时间成本，提高了调参选值效率。在不使用自适应阈值时，阈值设置为 0.7；将所有实验中经过非极大抑制留下的候选区域数量设置为 100(默认设置为 300)。其他设置保持默认不变，后续所有实验都在以上设置基础上进行。对于 LSTM，通道宽度乘子 $\alpha_\delta=1$，模型学习率为 0.003，其他参数与文献[19]一致。

6.3　评价指标

假设图像目标 $Z\{z_1,z_2,\cdots,z_n\}$，其中 $z_i=[x_z^i,y_z^i,a_z^i,b_z^i,c_z^i,d_z^i]$，算法对该图像输出为 $W\{w_1,w_2,\cdots,w_m\}$，目标 $w_j=[x_w^j,y_w^j,a_w^j,b_w^j,c_w^j,d_w^j]$。评价过程包含以下步骤：

1)建立目标和假设结果间的最优一一对应关系。采用欧氏距离来计算真实目标和假设目标的空间位置对应关系。欧氏距离的阈值 T 设置为假设和目标最少重叠时两者中心的距离。完成对应关系的目标数目为 N_T，漏检目标个数 $L_P=n-N_T$。

2)依据真实目标和假设目标对应的目标所属类别 d，将检测结果分为准确检测和误检两种情况。统计准确检测到的目标数目为 T_R，统计误检的目标数目为 T_W。比较真实目标个数 n 和检测的目标个数 m，如果 $n<m$，则存在虚警的情况，虚假目标个数 $F_P=m-N_T$。

3)由步骤 2)的统计结果，可以通过计算算法的虚警率、漏警率、检测率、误检率来衡量算法的检测效果，分别表示为

$$\begin{cases}P_f=\dfrac{F_P}{n}\\ P_m=\dfrac{L_P}{n}\\ P_d=\dfrac{T_R}{n}\\ P_e=\dfrac{T_W}{n}\end{cases}。\tag{16}$$

平均准确率(AP)是评价深度学习检测模型准确性最直观的标准，AP 从召回率和准确率两个角度衡量检测算法的准确性，可以用来分析单个类别的检测效果。平均准确率均值(mAP)是各个类别 AP 的平均值，mAP 越高表示模型在全部类别中检测的综合性能越高[1]。

6.4　各改进策略有效性验证

首先将各个策略与 SSD512 进行单独结合，并进行相应的对比实验，表明各个策略的作用。然后将所有策略与 SSD512 结合，对最终的改进算法进行整体测评。采用训练集训练原始 SSD512，将此模型记为 M0。在 M0 基础上加入新的特征提取网络 Gabor-VGGnet 策略，生成模型 M1。在 M0 基础上加入自适应阈值策略，生成模型 M2。在 M0 基础上加入基于时间感知特征映射和动态卡尔曼滤波的目标检测改进策略，生成模型 M3。最后将 M0 与所有策略结合在一起，生成模型 M4。使用两数据库测试集对 M0、M1、M2、M3、M4 进行测试和对比。表 1 对比了模型 M0、M1、M2、M3、M4 在 KITTI 和 WD 数据集上普通测试集的检测效果。

对比表 1 中 M0 和 M4 检测结果可知，在 KITTI 数据集中，各类目标检测的 AP 提高了 19%～25%，mAP 提高了约 21.76%，虚警率降低了15.02%，检测率提高了 40.15%，漏警率降低了 12.21%，误检率降低了 12.92%；在 WD 数据集中，各类目标检测的AP提高了21%～23%，mAP提高

表 1　各模型识别和检测效果比较

Table 1　Comparison of model identification and detection effects

Model	Dataset	AP /%			mAP /%	P_f/%	P_m/%	P_d/%	P_e/%
		Person	Car	Cyclist					
M0	KITTI	73.36	71.53	65.32	70.07	20.21	19.34	41.32	19.13
	WD	71.59	69.63	62.75	67.99	19.25	21.38	38.83	20.54
M1	KITTI	87.53	82.16	78.28	82.66	16.48	17.91	57.38	8.23
	WD	85.64	80.59	74.34	80.19	18.95	19.28	51.42	10.35
M2	KITTI	77.18	72.35	68.69	72.74	12.31	13.29	57.84	16.56
	WD	73.52	70.45	64.83	69.61	15.17	14.49	52.45	17.89
M3	KITTI	88.42	81.73	74.38	81.51	9.53	11.69	64.25	14.53
	WD	74.92	72.34	65.63	70.96	16.24	15.19	51.16	17.41
M4	KITTI	92.42	92.23	90.85	91.83	5.19	7.13	81.47	6.21
	WD	88.46	87.38	83.24	86.36	8.26	11.27	71.05	9.42

了约 18.37%，虚警率降低了 11.99%，检测率提高了 32.22%，漏警率降低了 8.07%，误检率降低了 11.12%。各项指标提升明显，表明本文策略总体对于弥补 SSD512 缺陷具有有效性。由表 1 可知，在 KITTI 数据集和 WD 数据集中，M1 相较于 M0，对目标的识别准确性得到了较大提高，多目标检测的误检率降低明显。M2 相较于 M0，对多目标的检测率得到了较大提高，多目标检测的虚警率和漏警率降低明显。M3 相较于 M0，多目标的检测率得到了较大提高，多目标检测的虚警率和漏警率降低明显，对各目标的识别精度和平均识别精度同样获得了较大的提高。而且，由于 WD 数据集是静态图像数据集，时空上下文策略无法生效，改进效果不如在视频数据集 KITTI 上的效果明显。

6.5　与其他算法对比实验

另外本文选取了 Faster R-CNN、深度监督对象检测器 300(DSOD300)检测框架[21]、YOLO 系列检测框架中的 YOLOv2 544[22]和 SSD 改进模型去卷积单镜头探测器(DSSD513)[23]作为深度学习对比算法，与 M4 对比在 Web Dataset 和 KITTI 数据集上的检测效果。检测识别效果如表 2 所示，其中 FPS 代表算法运行的速度、帧率。

表 2　不同算法检测和识别效果比较

Table 2　Comparison of detection and recognition with different algorithms

Method	Dataset	AP /%			mAP /%	P_d/%	FPS /(frame·s^{-1})
		Person	Car	Cyclist			
Faster R-CNN	KITTI	83.26	74.13	75.42	77.61	45.22	13.15
	WD	81.49	71.33	68.65	73.82	36.63	11.64
DSOD300	KITTI	77.43	72.26	68.38	72.69	58.68	58.23
	WD	70.73	69.39	67.04	69.05	52.32	50.35
DSSD513	KITTI	75.46	69.53	68.34	71.11	59.42	46.34
	WD	72.19	68.83	66.45	69.16	49.79	39.38
YOLOv2 544	KITTI	79.43	71.25	67.32	72.66	60.82	56.74
	WD	73.29	69.63	68.85	70.59	54.86	49.28
M4	KITTI	92.42	92.23	90.85	91.83	81.47	31.86
	WD	88.46	87.38	83.24	86.36	71.05	19.83

对比表 2 中 M4 和其他深度学习对比算法检测结果可知，在 KITTI 数据集中，各类目标识别的 AP 提高了 9%～16%，mAP 提高了约 14%～21%，检测率提高了 21%～36%；在 WD 数据集中，各类目标识别的 AP 提高了 7%～11%，mAP 提高了约 13%～16%，检测率提高了 11%～35%。M5 模型检测效果如图 8 所示。

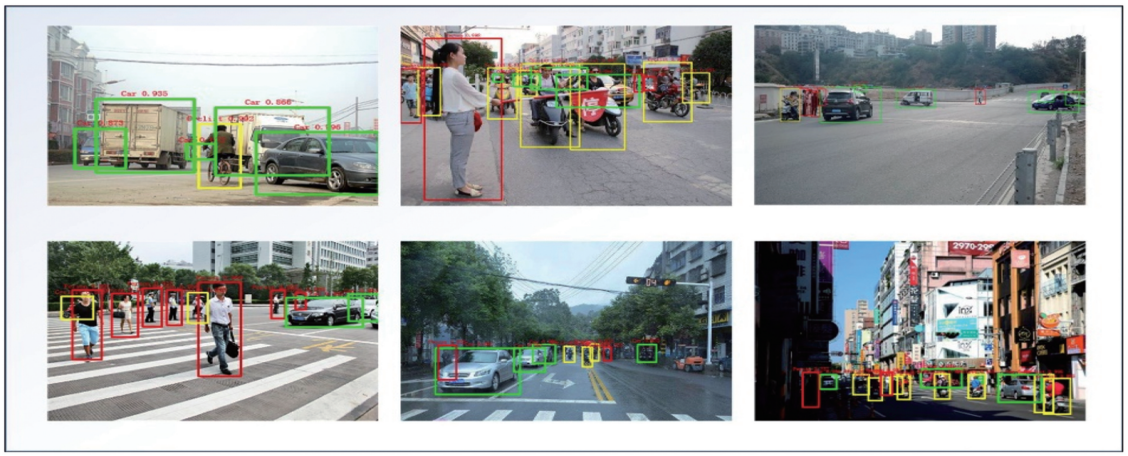

图 8　M4 模型检测结果示例

Fig. 8　Example of M4 model detection results

综上可知，M4 模型不仅在检测精度和识别精度上高于其他算法，而且检测速率达到了 32 frame·s^{-1}，验证了本文算法能够实现精度和实时性平衡，实现了既快又好，综合性能明显优于其他算法，具有较强的应用前景。

7　结　　论

针对现有基于大数据和深度学习的目标检测算法在复杂大场景下多目标检测的精度和实时性难以平衡的问题，改进了基于深度学习的目标检测框架 SSD，提出一种新的多目标检测框架——自适应感知 SSD，将其专用于复杂大交通场景多目标检测。实验结果表明，改进后的自适应感知 SSD 在应对弱小目标、多目标、杂乱背景、光照变化、模糊、大面积遮挡等检测难度较大的情况时，均能获得较好的效果，为深度学习在特定目标检测的应用提供了实例和新的思路。但是算法的处理效率距离工程实际应用的需求仍然有差距，后期如何降低运算量、提高算法的实时性和针对低分辨率弱小目标的检测和识别将是主要的研究方向。

参　考　文　献

[1] Feng X Y, Mei W, Hu D S. Aerial target detection based on improved Faster R-CNN[J]. Acta Optica Sinica, 2018, 38(6): 0615004.
冯小雨, 梅卫, 胡大帅. 基于改进 Faster R-CNN 的空中目标检测[J]. 光学学报, 2018, 38(6): 0615004.

[2] Yu K, Jia L, Chen Y Q, *et al*. Deep learning: yesterday, today, and tomorrow[J]. Journal of Computer Research and Development, 2013, 50(9): 1799-1804.
余凯, 贾磊, 陈雨强, 等. 深度学习的昨天、今天和明天[J]. 计算机研究与发展, 2013, 50(9): 1799-1804.

[3] Liu F, Shen T S, Ma X X. Convolutional neural network based multi-band ship target recognition with feature fusion[J]. Acta Optica Sinica, 2017, 37(10): 1015002.
刘峰, 沈同圣, 马新星. 特征融合的卷积神经网络多波段舰船目标识别[J]. 光学学报, 2017, 37(10): 1015002.

[4] Xin P, Xu Y L, Tang H, *et al*. Fast airplane detection based on multi-layer feature fusion of fully convolutional networks[J]. Acta Optica Sinica, 2018, 38(3): 0315003.
辛鹏, 许悦雷, 唐红, 等. 全卷积网络多层特征融合的飞机快速检测[J]. 光学学报, 2018, 38(3): 0315003.

[5] Lu Y F, Jin Q H, Jing J, *et al*. Detection and segmentation algorithm for bioresorbable vascular scaffolds struts based on machine learning[J]. Acta Optica Sinica, 2018, 38(2): 0215005.
鲁逸峰, 金琴花, 荆晶, 等. 基于机器学习的可降解支架检测与分割算法[J]. 光学学报, 2018, 38(2): 0215005.

[6] Liu W, Anguelov D, Erhan D, *et al*. SSD: single shot multibox detector[C]//European Conference on

Computer Vision, 2016: 21-37.

[7] Ren S Q, He K M, Girshick R, *et al*. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[8] Redmon J, Divvala S, Girshick R, *et al*. You only look once: unified, real-time object detection[C]// IEEE Conference on Computer Vision and Pattern Recognition, 2016: 779-788.

[9] Lin T Y, Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection[C]// IEEE Conference on Computer Vision and Pattern Recognition, 2017: 936-944.

[10] Zhou F Y, Jin L P, Dong J. Review of convolutional neural network[J]. Chinese Journal of Computers, 2017, 40(6): 1229-1251.
周飞燕, 金林鹏, 董军. 卷积神经网络研究综述[J]. 计算机学报, 2017, 40(6): 1229-1251.

[11] Chang L, Deng X M, Zhou M Q, *et al*. Convolutional neural networks in image understanding[J]. Acta Automatica Sinica, 2016, 42(9): 1300-1312.
常亮, 邓小明, 周明全, 等. 图像理解中的卷积神经网络[J]. 自动化学报, 2016, 42(9): 1300-1312.

[12] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]//European Conference on Computer Vision, 2014: 818-833.

[13] Wilson H R. Spatiotemporal characterization of a transient mechanism in the human visual system[J]. Vision Research, 1980, 20(5): 443-452.

[14] Tang P J, Wang H L, Kwong S. G-MS_2 F: GoogLeNet based multi-stage feature fusion of deep CNN for scene recognition[J]. Neurocomputing, 2017, 225: 188-197.

[15] Jain A K, Farrokhnia F. Unsupervised texture segmentation using Gabor filters[C]// IEEE International Conference on Systems, Man, and Cybernetics Conference Proceedings, 1990: 14-19.

[16] Keil A, Stolarova M, Moratti S, *et al*. Adaptation in human visual cortex as a mechanism for rapid discrimination of aversive stimuli[J]. Neuroimage, 2007, 36(2): 472-479.

[17] Liu Z H, Yin J, Jin Z. An adaptive feature and weight selection method based on Gabor image for face recognition[J]. Acta Photonica Sinica, 2011, 40(4): 636-641.
刘中华, 殷俊, 金忠. 一种自适应的Gabor图像特征抽取和权重选择的人脸识别方法[J]. 光子学报, 2011, 40(4): 636-641.

[18] Chen G, Zuo H F. The image adaptive thresholding by index of fuzziness[J]. Acta Automatica Sinica, 2003, 29(5): 791-796.
陈果, 左洪福. 图像的自适应模糊阈值分割法[J]. 自动化学报, 2003, 29(5): 791-796.

[19] Liu M, Zhu M. Mobile video object detection with temporally-aware feature maps[J]. arXiv preprint arXiv:1711.06368, 2017.

[20] Zhan J P, Huang X Y, Shen Z H, *et al*. Target tracking based on mean-shift and Kalman filter[J]. Journal of Chongqing University of Technology (Natural Science), 2010, 24(3): 76-80.
詹建平, 黄席樾, 沈志熙, 等. 基于均值漂移和卡尔曼滤波的目标跟踪方法[J]. 重庆理工大学学报(自然科学), 2010, 24(3): 76-80.

[21] Shen Z Q, Liu Z, Li J G, *et al*. DSOD: learning deeply supervised object detectors from scratch[C]// IEEE International Conference on Computer Vision, 2017: 1937-1945.

[22] Zhang J M, Huang M T, Jin X K, *et al*. A real-time Chinese traffic sign detection algorithm based on modified YOLOv2[J]. Algorithms, 2017, 10(4): 1-13.

[23] Fu C Y, Liu W, Ranga A, *et al*. DSSD: deconvolutional single shot detector[J]. arXiv preprint arXiv:1701.06659, 2017.