基于 Polygon-RefineNet 的违禁品 X 线图像自动标注方法

马博文 贾 同 刘益辄 滑心语

(东北大学信息科学与工程学院 沈阳 110819)

摘 要 近年来,随着深度学习的快速发展,其在智慧安检领域的应用已经成为了当下的研究热点.众所周知,深度学习方法是以海量训练数据为基础的,然而手工标注真值(ground truth)是一项十分繁琐的工作.为此,本文提出一种基于 Polygon-RefineNet 的违禁品 X 线图像自动标注方法,该方法在用户设定的包含感兴趣区域的初始边框(bounding box)内自动预测出目标的多边形轮廓,旨在生成可用真值的情况下最大限度地减少标注时间.由于违禁品 X 线图像存在大量的重叠现象,导致图像背景十分杂乱、违禁品轮廓模糊不清,因此本文首先引入多路径优化机制,通过有效利用深度网络下采样过程中提取的底层空间信息和高层语义信息来优化多边形轮廓的边缘细节,从而提高标注精度;其次,本文设计一种混合损失函数用于优化多边形轮廓的整体形状和位置,并同时消除真值本身存在的主观性误差使模型具有强大的泛化能力.最后,为了验证所提出方法的有效性,本文建立了一个违禁品 X 线数据集,该数据集包含 2623 张经过手工标注的 X 线图像,共 10 类 7257 个违禁品带有像素级真值.实验表明,本文提出的方法在标注违禁品时达到了 93.1%的准确率,且速度约是手工标注的 3.7 倍.本文进一步证明了该方法在 Cityscapes 数据集、MS COCO 数据集等其它域外数据集上的有效性.

关键词 深度学习;自动标注;X 线数据集;多路径优化;混合损失函数

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2021.00395

Automatic Annotation Approach for Prohibited Item in X-Ray Image Based on Polygon-RefineNet

MA Bo-Wen JIA Tong LIU Yi-Zhe HUA Xin-Yu

(*College of Information Science and Engineering, Northeastern University, Shenyang* 110819)

Abstract In recent years, with the rapid development of deep learning, its application in the field of smart security inspection has become a research hotspot. However, unfortunately, there is no X-ray segmentation dataset so far for research in relevant fields such as prohibited item instance segmentation and Threat Image Projection (TIP). At the meantime, it is universally acknowledged that deep learning approaches are data hungry and their performance is closely related to the amount of training data, though, but manually annotating ground truth instance masks is an extremely time-consuming task, especially in X-ray images. For these reasons, this paper proposes an efficient approach based on Polygon-RefineNet (PRN) for automatic annotation of prohibited items in X-ray images, aiming at minimizing the annotation time and yield high quality annotations. In particular, we construct a "fully convolutional" network to adapt prohibited items in different scales and styles, which takes as input an arbitrary-size initial bounding box

收稿日期:2019-08-29;在线发布日期:2020-05-13.本课题得到国家自然科学基金(U1613214)、国家重点研发计划(2018YFB14041)资助.**马博文**,硕士研究生,主要研究方向为机器学习、计算机视觉. E-mail: 2010285@stu.neu.edu.cn. **贾 同**(通信作者),博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为机器学习、模式识别、计算机视觉. E-mail: jiatong@ise.neu.edu.cn. **刘益辄**,硕士研究生,主要研究方向为机器学习、计算机视觉. **滑心语**,硕士研究生,主要研究方向为机器学习、计算机视觉.

given by a user containing region of interest and automatically produces a series of correspondingly-sized continuous vertices of the polygon outlining the prohibited item by efficient learning and inference. And our model allows the user to correct vertex at the end to produce as accurate annotation results as desired. Because of a large amount of overlapping phenomenon in X-ray images, the background of the images is complex and the boundary of the prohibited item is blurred. To solve this problem, we first introduce a multi-path refinement mechanism to refine the edge details of the polygon by making full use of the low-level spatial information and high-level semantic information that are extracted during the down-sampling process. Then, we design a mixed loss function by introducing the evaluation metric into the loss function for modifying the overall shape and position of the polygon and meanwhile eliminating the subjectivity error on account of the ground truth itself, which makes the network have strong generalization ability. Finally, to evaluate the proposed approach, we present a high-quality X-ray segmentation dataset named Prohibited Item X-ray (PIXray). The dataset consists of 2623 X-ray images by manually labeling, in which 10 classes of 7257 prohibited items have pixel-level ground truth. For a fair comparison, we validate our approach on our PIXray dataset and the public Cityscapes instance segmentation dataset, respectively. Experimental results demonstrate that our approach accelerates the annotation process by a factor of 3.7 in all classes of PIXray, while achieving 93.1% agreement after manual fine-tuning in IoU (Intersection over Union) with original ground truth, matching the typical agreement between human annotators. Besides, we outperform the baselines in 8 out of 10 categories, particularly well in the wrench, pliers, bat and razor. Our approach also obtains an IoU of 67.93% on Cityscapes dataset without using any fine-tuning. We further prove the effectiveness of our approach on Aerial Rooftop dataset, MS COCO dataset and PASCAL VOC dataset. The results show that the users can use our approach to achieve a relatively high reduction in time for annotating a new segmentation dataset.

Keywords deep learning; automatic annotation; X-ray dataset; multi-path refinement; mixed loss function

1 引 言

公共安全检查一直以来都在保障国民安全方面起着至关重要的作用. 在各种检查手段中,低成本、非接触、可成像的 X 射线安检机的应用最为普遍. 随着深度学习[1],尤其是卷积神经网络的快速发展,目前可基于一些深度网络模型自动地从 X 线图像中识别出违禁品[2-4],例如手枪和刀等. 随着对该领域研究的深入,许多问题被重新提出并尝试用深度学习方法加以解决,例如违禁品的分割[5]和图像注入. 但深度学习方法是以大量数据为基础的,并且模型的表现与可获得的训练数据量以及标签质量密切相关[6],这就需要手工标注大规模的数据集,最好是带有像素级(pixel-level)标注的分割数据集,由于其包含了标注物体的全部信息. 然而大多数分割数据集都是通过标注人员手工勾勒出目标的多边形轮廓来制作的[7-10],平均每个目标耗时 20～30 s[11],十分费时费力. 因此,不少研究旨在减少对像素级标注的依赖[12-15],例如 Miao 等人[15]利用弱监督学习方法定位违禁品的位置,但该方法的表现还达不到在实际中应用的标准. 其它研究,比如 Chen 等人[16]利用更容易获得的真值,如边界框(bounding box)真值,并利用类似 GrabCut[17]类型的方法在每个边界框内生成带有噪音的标签,但经研究表明,这种标签由于其内部固有的不精确性,只能用作辅助数据,不能作为官方的标准数据集[18-19]. 为此,针对具有复杂背景的 X 线图像,本文提出一种基于 Polygon-RefineNet (简称 PRN)的违禁品 X 线图像自动标注方法,旨在加快标注违禁品的速度,生成精准、可用的真值(ground truth),功能示意图如图 1 所示.

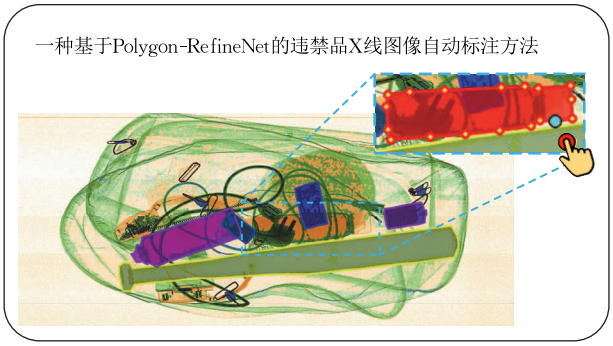


图 1　本文方法总体功能示意图(首先用户在待标注图像中设定一个包含感兴趣区域的任意尺寸的初始矩形边框,接着模型会自动预测出一系列勾勒违禁品轮廓的多边形顶点,最后允许用户手动微调预测的多边形以获得准确、可用的真值)

为了验证所提出方法的有效性,本文建立了一个违禁品 X 线数据集,名叫 Prohibited Item X-ray 数据集,简称 PIXray 数据集. 该数据集包含 2623 张经过手工标注的 X 线图像,共 10 类 7257 个违禁品带有像素级标签. X 线扫描图像不同于普通彩色图像[20],虽然行李箱(背包或者手提箱等)中的物品一般都随机、杂乱的堆叠在一起,存在大量的遮挡现象,但当行李箱通过 X 射线安检机的时候,由于 X 射线的透射性能,使得安检人员可以同时在扫描图像中看到前面的物品和后面被遮挡住的物品,这种现象被称为重叠现象[15]. X 线图像的这种特性对于计算机视觉领域来说既是优势也是挑战:优势是可以通过重叠现象标注出整个违禁品的轮廓,包括被遮挡的部分,得到违禁品的全部信息;挑战是 X 线图像背景本身就十分复杂,加之违禁品通常在尺度、形状和样式上具有相当大的变化,而且每个违禁品都有可能与任意数量和类型的安全物品或其它违禁品混合、重叠,这些原因都会导致违禁品的轮廓模糊不清.

为了在 X 线图像中标注出准确、可用的真值. 本文处理背景复杂问题的主要方法是:(1)高效地利用深度网络下采样过程中提取的各个层次的特征信息,特别是底层空间信息,其可以帮助在复杂场景下生成清晰、准确的边界;(2)提出了一种混合损失函数用于优化多边形的整体形状和位置.

本文的主要贡献有两个:(1)针对上述的 X 线图像特点,本文提出一种基于 Polygon-RefineNet 的违禁品 X 线图像自动标注方法,用于对 X 线图像进行像素级标注. 该模型的特点是可以在复杂背景下生成精准的边界,并且可以适应尺度、样式变化很大的违禁品,示意图如图 1 所示;(2)本文建立了一个高质量的违禁品 X 线数据集,PIXray 数据集,用于违禁品的高效标注研究,并为未来智慧安检领域的其它任务提供一个标准数据集. 在 PIXray 数据集上的实验结果表明,本文提出的方法在标注违禁品时达到了 93.1%的准确率,并且速度约是纯手工标注的 3.7 倍. 此外,本文进一步证明了该方法在 Cityscapes 数据集、MS COCO 数据集等其它域外数据集上的有效性.

2　相关工作

2.1　像素级标注

手工对图像进行像素级标注是一件十分费时费力的工作. 一直以来,许多交互式标注方法被提出来来加快该过程. 例如,Boykov 等人[21]通过用户手动标记某些像素为“目标”或者“背景”所提供的硬约束,再结合包含边缘信息和区域信息的软约束来完成分割. Shankar 等人[22]通过在“目标”和“背景”上使用多个涂鸦,并利用运动提示来标注视频中的目标. 文献[23]同样使用涂鸦来分割“目标”和“背景”,不同的是,Lin 等人之后使用卷积神经网络来进行图像语义分割. GrabCut[17]算法是 Graph Cuts[21]算法的改进,包括将基于灰度分布的模型替换为高斯混合模型以支持彩色图片,将一次性得到结果的算法改成强大的迭代过程,该算法利用了图像的 RGB 色彩信息和边界信息,只要少量的交互操作即可得到比较好的分割效果. 基于这个思路,Rajchl 等人[24]将 GrabCut 算法和卷积神经网络结合在一起来分割医疗图像. Chen 等人[16]利用 3D 边界框和点云来协同标注,等等. 大多数这些方法将目标分割视为像素级分类问题,这就导致很难对标注结果进行修改,并且在由阴影、图像饱和度或对象的低分辨率造成的模糊区域中表现不佳. 相对应地,文献[25]和文献[26]旨在寻找勾勒出目标的最优边界. 文献[27]中,Duan 等人通过在小的多边形中生成超像素,并将这些超像素合并为对象区域来标注航拍图像. 文献[11]和文献[28]采用了另一种新的处理像素级标注的思路,即通过预测勾勒出目标多边形轮廓的一系列顶点来实现分割,这样做更符合手工标注的行为. 且用少量的多边形顶点标注出的目标便于更改,仅需修改预测错误的顶点即可. 与文献[11]和文献[28]相同,本文也通过预测目标的多边形轮廓来实现分割,不同的是,本文通过在模型中加入多路径优化机制和混合损失函数来优化预测的多边形顶点

以在复杂场景下获得精准的标注结果.

2.2 X 线数据集

X 线图像是通过 X 射线照射物体并根据它们的光谱吸收率用伪彩色渲染它们得到的,例如,金属材质一般为蓝色,有机物一般为橙色,混合物一般为绿色. X 线图像最大的特点就是存在大量的重叠现象,而这种特点也为计算机视觉领域带来了全新的挑战. 随着目标检测领域的飞速发展,许多研究人员针对这些困难做了很多工作并取得了一些成果[29-33]. 但不幸的是,用于研究目的的公开 X 线数据集却少之又少,虽然近期发布了两个标准数据集:GDXray 数据集[34]和 SIXray[15]数据集,但其自身都存在着一定的局限性. GDXray 数据集中主要包含 3 类(手枪、手里剑和剃须刀片)约 1000 张带有违禁品的 X 线扫描图,但图像背景十分简单且是灰度图,不能模拟真实场景,因此应用受到了限制. 与 GDXray 数据集不同,SIXray 数据集提供了大量从地铁站收集的真实世界的 X 线扫描图像,其中 8929 张图像带有违禁物品,违禁品种类包含 6 种,分别是手枪、刀、扳手、钳子、剪刀和锤子,但作者仅为其中的大约 1400 张图像进行了边界框标注,其余的标注为图像级(image-level)真值. 背景简单、数据匮乏、标签弱等问题都为智慧安检在实际中的应用带来了困难.

3 Polygon-RefineNet 模型

由于 X 线扫描图像本身存在的由重叠现象造成的背景杂乱、违禁品轮廓模糊不清等问题,导致现有的图像标注模型并不适用于 X 线图像. 为此,本文提出 Polygon-RefineNet 模型,简称 PRN 模型,旨在生成精准、可用的真值(ground truth)的同时,加快标注违禁品的速度. 该模型仿照手工标注数据集的过程,通过预测目标的多边形轮廓(封闭)来实现标注. 特别地,本文将多边形轮廓参数化为一系列连续的 2D 顶点,其中两个连续的顶点组成一条边. 值得注意的是,由于封闭的多边形是由顶点构成的循环序列,因此通过选择不同的起点和不同的方向就可以得到多个等效的参数化形式,本文固定生成多边形的方向为顺时针方向,但起点位置是任意的,由模型自动选择.

本文假设用户在每次标注时都会事先设定一个包含感兴趣区域的任意大小的初始矩形边框,PRN 模型会自动裁剪出边框所包含的图像内容作为输入. 模型的总体结构采用编码器+解码器+优化器,如图 2 所示. 其中编码器用于提取图像特征,包括底层视觉特征和高层语义特征,解码器用于每一个时刻解码(预测)一个顶点位置,优化器用于优化解码器输出的顶点位置以在复杂背景下得到精准、细致的高精度结果. 输出的顶点位置表示为与输入图像尺寸相同的网格(大小为 $H\times W$,其中 H 表示输入图像的高度,W 表示输入图像的宽度)中的位置. PRN 模型允许用户在最后手动微调预测错误的顶点以获得准确、可用的真值. PRN 模型的具体结构如图 3 所示. 接下来将详细地介绍 PRN 模型.

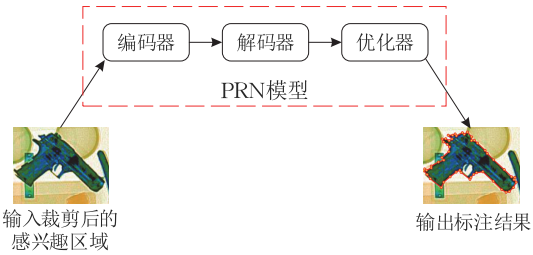


图 2 PRN 模型总体结构图

3.1 模型结构

在前人的工作中[11,28],作者将每一时刻模型预测出的多边形顶点位置处理为多分类问题,引入全连接层,导致模型的输入只能是固定尺寸的,进而造成许多长宽比差别很大的违禁品在缩放图像时边缘细节就已经丢失. 为了更好地适应不同违禁品在尺度上的差异性,减小由于缩放图像尺寸而造成的精度损失,本文建立了一个全卷积网络. 接下来将分别详细地介绍模型的各个部分,编码器、解码器以及由多路径优化机制构成的优化器.

3.1.1 编码器

大多数模型通过执行重复的下采样操作来获得高层语义信息,进而实现分割任务. 但这样做不仅牺牲了输出的分辨率[35-36],而且忽略了大量的空间视觉信息. 为了缓解这种情况,本文在模型中充分利用下采样过程中提取的各个层次的特征信息,包括底层空间信息,例如边信息和角信息,和高层语义信息. 整个编码器的主干卷积神经网络的选择是任意的,本文采用 VGG-16 结构[37],首先舍弃了全连接层和最后的最大池化层,pool5. 接着分别取 VGG-16 结构中第二个、第三个和第四个最大池化层(pool2、pool3 和 pool4)的输出作为不同层次(不同分辨率)的特征信息,并表示为 $\{P_2,P_3,P_4\}$. 同时为了节省内存空间,本文同样不采用第一个最大池化层的输出,编码器结构如图 3 所示.

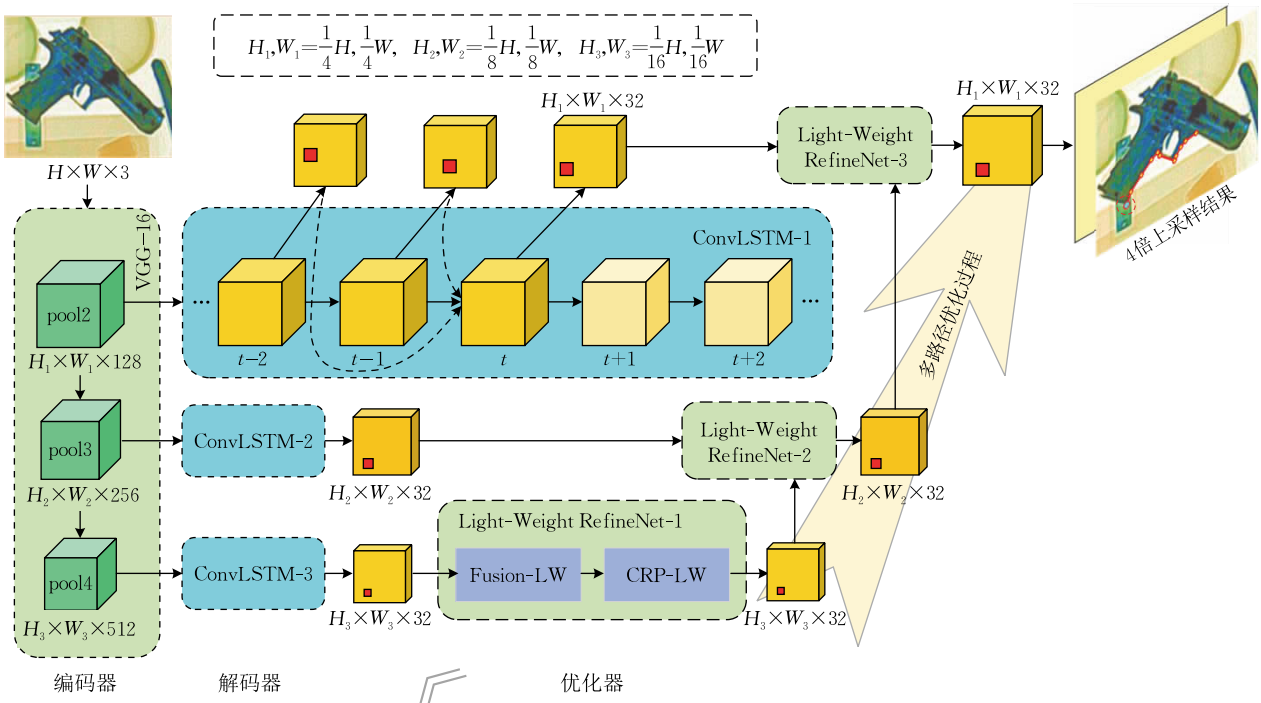


图 3　PRN 模型结构示意图(在每一个时间步,3 个解码器(ConvLSTM-1、2、3)分别接收:(1)编码器(VGG-16)提取的不同层次(不同分辨率)的特征信息;(2)前两个时间步对应解码器输出的顶点位置信息;(3)该分辨率下的起点位置信息.在解码器之后,本文使用由多路径优化机制构成的优化器来逐步融合、细化当前时刻解码器输出的不同分辨率下的顶点位置信息以得到一个准确、高分辨率下的顶点位置,最后加入层和空间损失来实现端到端的训练)

3.1.2　解码器

循环神经网络(Recurrent Neural Network,简称 RNN)是一种十分强大的可以处理时序数据的神经网络,其通过采用线性和非线性函数来传递历史信息.但本文不仅希望 RNN 可以对勾勒出目标轮廓的多边形顶点进行连续的预测,而且希望 RNN 能够识别物体的形状.因此,本文采用卷积长短时记忆网络(Convolutional LSTM Network,简称 ConvLSTM[38])作为解码器,由于其可以同时处理时间信息和空间信息,并且相比于全连接 RNN,ConvLSTM 由于其采用卷积操作从而减少了需要学习的参数.一个简单形式的单层 ConvLSTM 的公式定义如下:

$$\begin{pmatrix} i_t \\ f_t \\ g_t \\ o_t \end{pmatrix} = W_x * X_t + W_h * H_{t-1} + b$$

$$C_t = \sigma(f_t) \circ C_{t-1} + \sigma(i_t) \circ \tanh(g_t) \qquad (1)$$

$$H_t = \sigma(o_t) \circ \tanh(C_t)$$

其中,X_t 表示输入,H_{t-1} 和 H_t 分别表示前一时刻的输出和当前时刻的输出,C_{t-1} 和 C_t 分别表示前一时刻的细胞状态和当前时刻的细胞状态,i_t,f_t,o_t 分别表示输入门、遗忘门和输出门,W_x 和 W_h 分别表示输入层到当前状态的权重矩阵和隐藏层到当前状态的权重矩阵,σ 表示 sigmoid 函数,"$\circ$"表示元素积,"$*$"表示卷积操作.

在同一时刻,本文采用 3 个卷积核大小为 3×3,卷积核个数为 32 的单层的 ConvLSTM 来分别解码不同分辨率下的 2D 顶点位置 y_{1t},y_{2t} 和 y_{3t},每一时刻每个 ConvLSTM 接受 4 种信息作为输入,分别是对应的不同层次的图像特征信息,对应 ConvLSTM 在前两个时刻预测的顶点位置 $y_{i(t-2)}$ 和 $y_{i(t-1)}$,以及为了确定何时结束预测而需要的起点位置信息 y_{i1},每一时刻每个 ConvLSTM 的输出表示为 $H_i\times W_i\times 32$ 维的特征信息,其中前 $H_i\times W_i$ 维表示不同分辨率下对应的顶点可能存在的 2D 位置,最后一维表示通道数.其中 $i=1,2,3$ 分别表示 3 个单层的 ConvLSTM.3 个 ConvLSTM 输出的分辨率分别是输入图像大小的 1/4×1/4,1/8×1/8 和 1/16×1/16,解码器结构如图 3 所示.

在 PRN 模型中,给定两个连续的多边形顶点,下一个顶点是被唯一确定的[11,28],但初始顶点(起点)除外.为此,本文提出一种高效的初始顶点选取方法,利用两个虚拟顶点来自动选择初始顶点的位

置.模仿人类的视觉特点,取输入图像中相邻的两个特征点作为两个虚拟顶点,比如角点或边界.为此,本文首先将 P_2 送入到一个卷积核大小为 1×1,卷积核个数为 1 的卷积层中用于提取输入数据中的特征点,并将输出表示为 P_2',接着取 P_2' 中最大值的位置作为该分辨率下的第一个虚拟顶点,表示为 $y_{1(-2)}$,之后在 $y_{1(-2)}$ 顶点附近的 8 像素内取次大值的位置作为第二个虚拟顶点,表示为 $y_{1(-1)}$.最后分别取 $y_{1(-2)}$,$y_{1(-1)}$ 坐标的 1/2,1/4 作为 P_3,P_4 分辨率下的虚拟顶点,并表示为 $y_{2(-2)}$、$y_{2(-1)}$,$y_{3(-2)}$、$y_{3(-1)}$,虚拟顶点选取过程如图 4 所示.

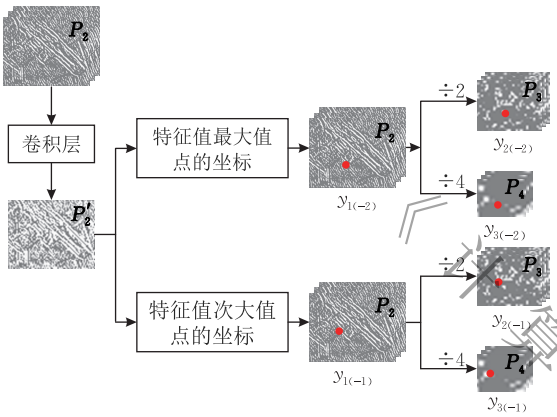


图 4 虚拟顶点选取过程示意图

3.1.3 优化器

现有的一些图像标注模型[11,28]虽然采用普通加和方式融合了底层视觉信息和高层语义信息,但其在融合信息的过程中或融合后送入到解码器的过程中又重复使用了下采样操作,导致底层视觉信息的作用大大降低.为解决这个问题,实现模型在背景复杂的 X 线图像中既快速又精准地标注违禁品的目标,本文提出一种基于多路径优化机制的网络模型设计方法,如图 3 所示.由多路径优化机制构成的优化器是由三个 Light-Weight RefineNet 优化模块组成的三级优化过程,Light-Weight RefineNet 优化模块,简称 RefineNet-LW 模块[39],是利用多层抽象特征来逐步细化预测的顶点位置以得到精准的高分辨率结果,其中高层语义特征用于把握对象或类别信息,底层视觉特征用于生成精准细致的边界.每个 RefineNet-LW 模块由两部分组件组成,分别是融合块(简称 Fusion-LW)单元和链式残差池化块(Chained Residual Pooling-LW,简称 CRP-LW)单元.Fusion-LW 单元和 CRP-LW 单元的详细结构分别如图 5、图 6 所示.

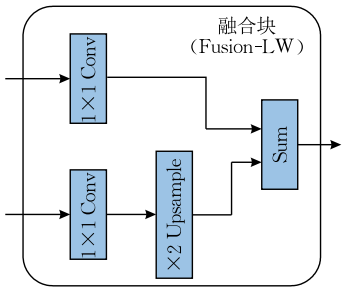


图 5 Fusion-LW 单元结构示意图

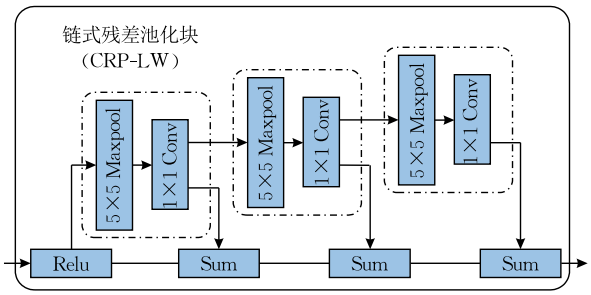


图 6 CRP-LW 单元结构示意图

如图 3 所示,整个优化过程开始于第三个解码器(ConvLSTM-3)的输出 y_{3t},首先经过 CRP-LW 单元用于提取 y_{3t} 中最显著的特征(顶点最可能存在的位置),值得注意的是,如果 RefineNet-LW 模块只有一个输入,比如图 3 中的第一个优化模块,RefineNet-LW-1 模块,则该输入就直接通过 Fusion-LW 单元,不发生任何变化.CRP-LW 单元是一条由多个池化块组成的链式结构,每个池化块由一个池化窗口为 5×5 的最大池化层和一个卷积核大小为 1×1,卷积核个数为 32 的卷积层组成,每个池化块都利用最大池化层来提取输入池化块数据中的显著特征,之后进行逐元素相加来保证每次提取的显著特征都被保留,重复 3 次直到将大区域中的显著特征提取出来.之后,经过 CRP-LW 单元提取的大区域显著特征和第二个解码器(ConvLSTM-2)的输出 y_{2t} 一起被送入到 Fusion-LW 单元用于融合不同层次的特征信息.在 Fusion-LW 单元内部,每条路径都首先经过卷积核大小为 1×1,卷积核个数为 32 的卷积层用于匹配输入数据,生成相同通道数的特征图,之后小分辨率的输入经过上采样层上采样到该路径的最大分辨率后与大分辨率的输入通过逐元素相加方式融合.然后,再将融合后的特征信息送入到 CRP-LW 块中用于提取融合信息中的大区域显著特征,重复此过程直到达到所需的分辨率,优化器优化过程热力图如图 7 所示.最后将优化结果经过 4 倍上采样后输入到一个卷积核大小为 3×3、卷积核个数为 1 的卷积层中来产生得分图(score map).每一时

刻得分图的输出维度为 $H\times W$ 维，表示当前时刻顶点最可能存在的 2D 位置. 整体得分图的输出表示为 $n\times H\times W$ 维，其中 n 表示预测的顶点数. 整个模型可以实现端到端的训练，更重要的是，优化器的设计可以实现细粒度底层视觉信息和粗粒度高层语义信息的合理融合，使预测出的顶点位置得以细化.

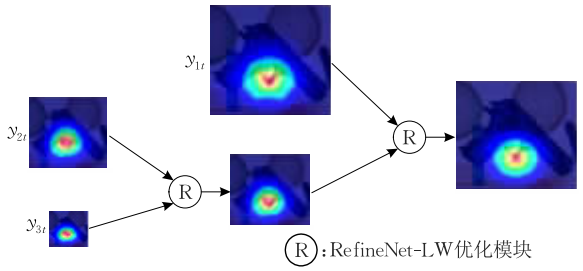


图 7　优化器优化过程热力图

3.2　混合损失函数

本文设计一种混合损失函数用来评估模型的预测值与真实值的不一致程度. 该损失函数分为两个部分：分类损失函数和多边形回归损失函数. 其中分类损失函数用于让模型“学习”手工标注的标注方式和标注特点；多边形回归损失函数用于修正预测多边形整体的形状和位置，并同时消除真值本身存在的主观性误差，防止过拟合现象的发生，使模型具有强大的泛化能力. 整个模型的损失函数定义如下：

$$L=L_{\text{cls}}+\lambda L_{\text{reg}} \tag{2}$$

其中，L_{cls} 表示分类损失函数，L_{reg} 表示多边形回归损失函数，λ 表示平衡参数，用于平衡分类损失函数和回归损失函数的数量级，使模型在训练时均匀考虑两种损失函数. 接下来将详细介绍混合损失函数的每一部分.

(1) 分类损失函数. 由于本文将每一时刻预测的顶点位置处理为与输入图像尺寸相同的网格中每个格子的二元分类问题，所以对每个格子采用二元分类的交叉熵损失函数，定义如下：

$$L_{\text{cls}}=-\sum_i(\hat{Y}_{t_i}\log(Y_{t_i})+(1-\hat{Y}_{t_i})\log(1-Y_{t_i})) \tag{3}$$

其中，$\hat{Y}_{t_i}$ 表示当前时刻一个格子的顶点真值(0 或 1)，Y_{t_i} 表示当前时刻对应格子的预测值(概率值)，且 $Y_{t_i}\in(0,1)$.

(2) 多边形回归损失函数. 对于任意一个多边形，其每一条连续的边都可以离散成无数个顶点，因此，就可以通过无数种组合来勾勒出一个多边形. 但由于手工标注训练数据集时存在的主观性(训练集的顶点坐标真值(ground truth)是固定的)而导致不能客观反映现实中选择多边形顶点的多样性. 即如果单纯使用分类损失函数作为整个模型的损失函数，就会造成只有当顶点坐标预测值与顶点坐标真值完全重合时损失值才会下降的情况，这就违背了现实中多边形顶点具有多样性的条件. 示意图如图 8 所示，当顶点坐标预测值为第一种情况时，只使用分类损失函数作为模型的总损失函数是没有问题的，但当顶点坐标预测值为第二种情况时(现实中还有很多类似情况，图中只举一例)，虽然顶点坐标不完全重合，但预测结果准确率却接近 100%，如果此时模型只使用分类损失函数作为总损失函数，就会造成严重的过拟合现象. 为缓解这种情况，本文引入 GIoU 损失函数(Generalized Intersection over Union)[40] 作为多边形回归损失函数，将评价指标引入到损失函数当中，用于消除真值本身存在的主观性误差，并同时优化预测多边形整体的形状和位置.

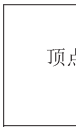
		第一种情况(顶点坐标完全重合)	第二种情况(顶点坐标不重合)
顶点坐标真值(GT)			
顶点坐标预测值			
损失值(Loss)	分类损失函数	很小	非常大
	分类损失函数+回归损失函数	很小	较小
准确度(IoU)		100%	≈100%

图 8　主观性误差对训练 PRN 模型时产生的影响

GIoU 损失函数主要解决了当 IoU 作为损失函数时，预测框与真值框不重合，损失值为 0 无法优化的问题. GIoU 损失函数的公式定义如下：

$$L_{\text{reg}}=L_{\text{GIoU}}=1-\left(\frac{|A\cap B|}{|A\cup B|}-\frac{|C\backslash(A\cup B)|}{|C|}\right) \tag{4}$$

其中，A、B 分别表示预测多边形和真值多边形，C 表示包围 A、B 的最小凸多边形(为了方便，本文采用矩形).

4　PIXray 数据集

为了验证 Polygon-RefineNet 模型的有效性，本文建立了一个违禁品 X 线数据集，PIXray 数据集. 该数据集共有 2623 张 X 线图像，包含 10 种类型共 7257 个违禁品. 违禁品种类包括手枪(Gun)、

刀(Knife)、扳手(Wrench)、钳子(Pliers)、剪刀(Scissors)、打火机(Lighter)、电池(Battery)、棒球棍(Bat),剃须刀片(Razor)和压力容器(Pressure Vessel,简称 PV).PIXray 数据集的详细统计信息如表 1 所示.PIXray 数据集由两部分构成,第一部分,本文从 SIXray 数据集[15]中随机选取了其中较清楚的 553 张带有违禁品的图像作为 PIXray 数据集的一部分,这一部分包含了手枪、刀、扳手、钳子和剪刀这 5 类违禁品,由于 SIXray 数据集中锤子的样本太少,所以本文没有采用.第二部分,其余的 2070 张 X 线图像是由实验室人员模拟真实场景采集的,包括刀、打火机、电池、棒球棍、剃须刀片和压力容器这 6 类十分常见但却容易被忽视的违禁品.每张图像都由型号为 UNX5030E 的 X 射线安检机扫描得到,所有图像均保存成 JPG 格式.之后使用开源软件 LabelMe[41]来手工标注全部的 X 线图像,效果图如图 9 所示,其中每个违禁品的真值都由勾勒出该违禁品轮廓的多边形的顶点坐标组成,每张图像都有一个用名字对应的 JSON 文件用于保存真值信息.

表 1 PIXray 数据集的种类分布表

	种类	数量/个
PIXray 数据集 (共 2623 张 X 线图像)	手枪	457
	刀	974
	扳手	175
	钳子	136
	剪刀	48
	打火机	840
	电池	1423
	棒球棍	692
	剃须刀片	1135
	压力容器	1377

图 9 PIXray 数据集手工标注效果图

5 实验结果与分析

本文的实验环境为 GTX 1080Ti GPU,深度学习框架为 Pytorch,编程语言为 Python.本节将分为两个部分来展示实验结果,域内结果和域外结果.域内结果,即 PRN 模型在 PIXray 数据集上的各项指标测试结果.另外,为了更直观地展示 PRN 模型的优点并同时证明 PIXray 数据集的合理性,本文分别在 PIXray 数据集上测试了其它现有的主流标注模型,例如 DeepMask 模型[42]、SharpMask 模型[43]、Polygon-RNN 模型[11]以及 Polygon-RNN++模型[28];域外结果,为了进一步分析 PRN 模型的各项性能,本文同样展示模型在其它域外数据集上的表现,包括 Cityscapes 数据集[10]、Aerial Rooftop 数据集[26]、MS COCO 数据集[9]和 PASCAL VOC 数据集[7].值得注意的是,以下实验结果中提到的自动标注结果均指模型自动预测的结果,即不经过任何手工微调.

接下来,本文将主要从以下三个方面来评价 PRN 模型:(1)评价模型自动预测多边形轮廓的质量,即计算每一个预测多边形和对应的真值多边形的拟合程度,采用标准 IoU(Intersection over Union)值来度量(以百分数形式表示),并求其平均值;(2)评价模型标注违禁品轮廓的速度,即计算平均标注一个违禁品所需的时间;(3)评价用户最后手工微调预测多边形的人工操作量,以修改次数衡量,即对于一个模型自动标注的多边形,统计需要几次的鼠标点击修改才能使多边形达到预期的准确率.

5.1 域内实验

5.1.1 PIXray 数据集

如上所述,PIXray 数据集共包含 2623 张带有违禁品的 X 线图像,本文随机选取了其中的 2100 张图像作为训练集,173 张图像作为验证集,剩余的 350 张图像作为测试集.表 2 展示了将 PIXray 数据集拆分后每种违禁品对应的个数.

表 2 PIXray 数据集拆分结果表

拆分	手枪/个	刀/个	扳手/个	钳子/个	剪刀/个	打火机/个	电池/个	棒球棍/个	剃须刀片/个	压力容器/个	违禁品总数/个	图像总数/张
训练集	367	766	114	83	22	635	1114	542	999	1176	5818	2100
验证集	35	92	17	17	8	59	33	47	18	65	391	173
测试集	55	116	44	36	18	146	276	103	118	136	1048	350

5.1.2　质量评估

表 3 展示了不同模型在 PIXray 数据集上的测试结果，包括 DeepMask 模型[42]、SharpMask 模型[43]、Polygon-RNN 模型[11]、Polygon-RNN++模型[28]以及本文提出的 Polygon-RefineNet 模型. 为了保证公平，本文均展示没有经过任何手工微调的结果. 从表 3 可知，现有的标注模型在违禁品 X 线图像上的表现并不理想，DeepMask 模型由于只采用深度卷积神经网络中的高层语义信息进行分割，导致之后利用双线性上采样操作将分割结果放大到与原图像相同尺度后，物体边缘十分粗糙. SharpMask 模型、Polygon-RNN 模型和 Polygon-RNN++模型虽然融合了底层空间信息与高层语义信息来实现边缘细节的优化，但这些模型并没有考虑底层视觉信息的利用率. 而且在损失函数的设计上仅考虑了顶点的预测位置，并没有关注生成多边形的整体位置和形状. 相比而言，本文提出的 PRN 模型在 PIXray 数据集上的表现十分优秀，10 种违禁品类别中有 8 种类别的准确率超过了其它模型，平均准确率也超过了最先进的模型(Polygon-RNN++模型)约 5.6%，特别是在扳手、钳子、棒球棍以及剃须刀片这四类存在严重重叠现象或尺度差异较大的违禁品上，PRN 模型标注的准确率更是大幅度领先其它模型. 图 10 展示了使用 PRN 模型自动标注违禁品 X 线图像的结果. 同时为了更直观地突出 PRN 模型的优点，本文在图 11 中进一步展示了使用 Polygon-RNN++模型自动标注图 10 中(d)组图像的结果，从对比结果可知，相比于 Polygon-RNN++模型，PRN

表 3　在 PIXray 数据集上的表现结果(IoU 值/%)(不经过任何手工微调)

模型	手枪	刀	扳手	钳子	剪刀	打火机	电池	棒球棍	剃须刀片	压力容器	平均值
DeepMask	56.92	63.57	49.32	38.60	57.38	74.19	77.53	59.58	61.14	64.18	60.24
SharpMask	59.93	68.63	52.84	44.37	62.96	78.15	81.72	66.25	64.45	67.38	64.67
Polygon-RNN	61.37	71.01	54.79	47.67	68.65	78.82	82.48	65.04	66.52	69.73	66.61
Polygon-RNN++	71.45	77.28	62.95	57.25	**77.91**	88.23	**88.45**	72.20	79.07	80.11	75.49
Polygon-RefineNet	**78.34**	**79.44**	**76.43**	**66.64**	77.02	**89.11**	87.86	**83.55**	**86.21**	**86.09**	**81.07**

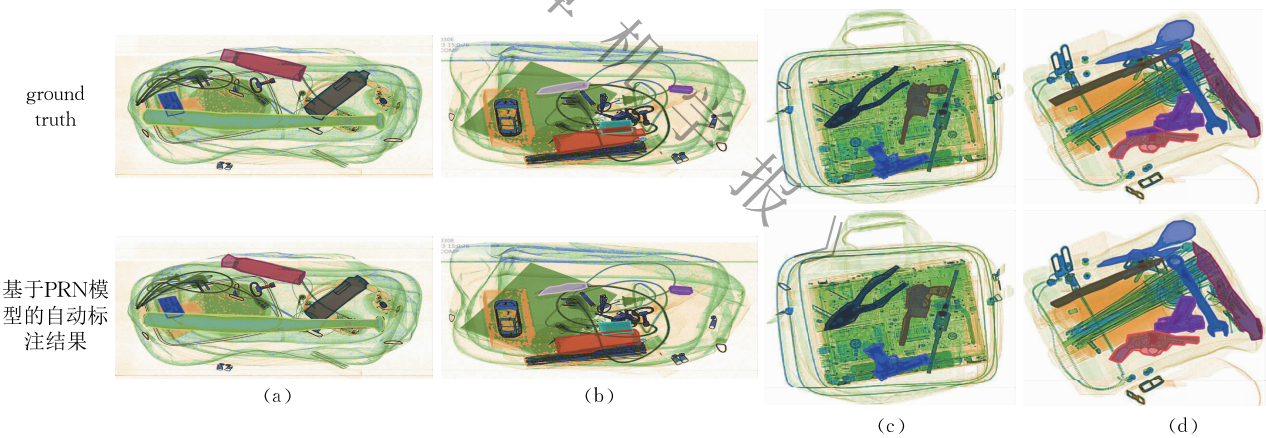


图 10　在 PIXray 数据集上的标注结果图(每一组图像((a)(b)(c)(d))中上方的图像为真值图(ground truth)，下方的图像为自动标注结果图(没有经过任何手工微调))

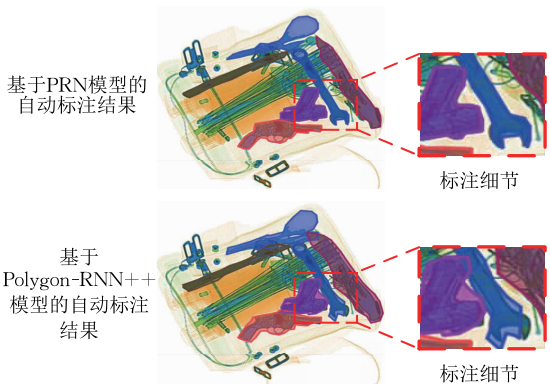


图 11　分别使用 PRN 模型和 Polygon-RNN++模型标注同一张 X 线图像的结果对比图

模型能够在图像背景十分杂乱的情况下更精准地把握违禁品的边缘轮廓细节，并且可以更好地标注长宽比差别很大的违禁品，例如图 11 中的刀. 但 PRN 模型本身也存在一定的不足，比如当两个重叠物品颜色十分相近时，PRN 模型的表现并不理想，如图 11 中的两个手枪.

5.1.3　初始边框鲁棒性分析

由于用户在设定包含感兴趣区域的初始边框时存在极大的主观因素，因此本文进一步分析给模型输入不同尺寸的初始边框时模型的准确率变化情况，即以初始边框宽度和高度的百分比随机扩展初

始边框，本文分别取 0 扩展(无扩展)、0～5%扩展、5%～10%扩展以及 10%～15%扩展这四个阶段的初始边框进行实验，结果如表 4 所示. 从实验结果可知，本文提出的 PRN 模型对于 0～5%扩展的初始边框具有良好的鲁棒性，即使对于 5%～10%和 10%～15%扩展的初始边框，模型也能够较为准确地预测出目标的多边形轮廓.

表 4 PRN 模型对初始边框的鲁棒性统计

初始边框扩展量/%	准确率(IoU 值/%)
0	81.07
0～5	80.43
5～10	78.24
10～15	74.62

5.1.4 速度评估

PRN 模型的另一优势就是允许用户在最后手工微调预测的多边形以获得准确、可用的真值标签. 为此，本文在 PIXray 数据集中随机选取了 30 张 X 线图像，并分别统计通过纯手工标注、模型自动标注(给定包含感兴趣区域的初始边框)以及模型半自动标注(模型自动标注＋手工微调，给定包含感兴趣区域的初始边框)的准确率和时间，并求其平均值. 结果展示在表 5. 从表 5 可知，经过简单的手工微调后，PRN 模型标注的准确率可以达到 93.1%，且标注速度约是纯手工标注的 3.7 倍.

表 5 平均标注一个违禁品所需的时间和对应的 IoU 值

数据集	模型	标注方式	时间/s	IoU 值/%
PIXray 数据集		纯手工标注	28.8	100.0
	Polygon-RNN++模型	自动标注	0.4	75.5
		半自动标注(自动标注＋手工微调)	9.3	92.4
	PRN 模型	自动标注	0.3	81.1
		半自动标注(自动标注＋手工微调)	**7.8**	**93.1**

5.1.5 手工微调分析

本文进一步分析手工微调每类违禁品使其达到 90%以上准确率所需的平均点击次数，实验图像依旧采用上述的 30 张 X 线图像，实验结果展示在表 6 中. 从表 6 可知，使用 PRN 模型自动标注的违禁品在最后手工微调时仅需平均 6.6 次鼠标点击修改就可以达到 93.1%的准确率. 其中，打火机、电池和剃须刀片这 3 类形状简单的违禁品仅需 1～3 次点击就可以达到很高的准确率；而钳子、剪刀这 2 类形状为复杂“X”型的违禁品，则需要相对较多的点击次数来修正多边形的轮廓. 每类违禁品在微调过程中点击次数与所达到准确率的变化情况如图 12 所示.

表 6 手工微调每类违禁品平均所需的点击次数和对应达到的 IoU 值

违禁品种类	平均点击次数/次	所达到的 IoU 值/%
手枪	8.4	92.4
刀	5.1	93.8
扳手	9.6	91.4
钳子	13.2	92.0
剪刀	11.4	90.8
打火机	1.6	96.4
电池	1.8	95.8
棒球棍	5.3	92.5
剃须刀片	3.4	93.5
压力容器	5.7	92.2
平均值	6.6	93.1

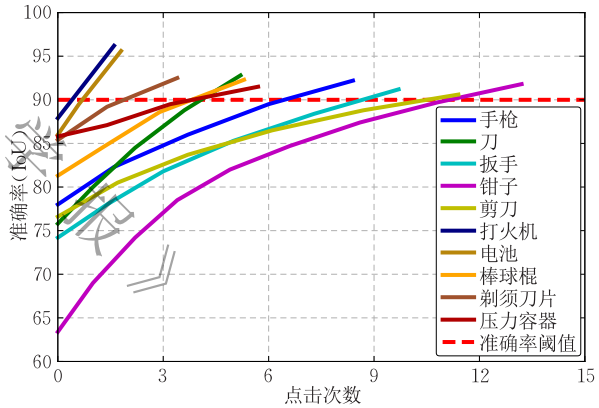


图 12 修改点数与对应准确率变化曲线图

5.2 域外实验

为了全面分析 PRN 模型的性能，本文进一步在其它域外数据集上做了一系列实验，包括 Cityscapes 数据集[10]、Aerial Rooftop 数据集[26]、MS COCO 数据集[9]以及 PASCAL VOC 数据集[7]. 首先在 Cityscapes 数据集上重新训练了 PRN 模型，并详细地展示了实验结果. 之后验证了在 Cityscapes 数据集上训练过的模型对其它未知数据集的泛化能力，包括 Aerial Rooftop 数据集、MS COCO 数据集和 PASCAL VOC 数据集.

5.2.1 域外数据集

Cityscapes 数据集是目前最全面的图像分割标准数据集之一. 该数据集的图像取自 27 个德国城市

和其邻国城市的道路场景，共标注了 8 类街拍物体的像素级真值，分别是行人(Person)、骑手(Rider)、汽车(Car)、卡车(Truck)、公交车(Bus)、列车(Train)、摩托车(Motorcycle)和自行车(Bicycle)，包含 2975 张训练图像，500 张验证图像和 1525 张测试图像. 为了保证实验结果的公平性，本文使用与文献[11]相同的数据集拆分方式.

Aerial Rooftop 数据集中的图像是带有建筑物屋顶的乡村场景的航拍图像. MS COCO 数据集和 PASCAL VOC 数据集都是十分常见的大规模数据集，其图像内容主要是复杂的日常场景.

5.2.2　域外定量结果

表 7 中展示了 PRN 模型在 Cityscapes 数据集上的测试结果，本文同样采用上述的 DeepMask 模型、SharpMask 模型、Polygon-RNN 模型以及 Polygon-RNN++模型作为本实验的对比模型. 从表 7 可知，本文提出的 PRN 模型在标注 Cityscapes 数据集中的街景图像时也可以较好地满足标注需求，其在自行车、公交车、列车以及摩托车这 4 类街拍目标上的表现超过了最先进模型，准确率分别达到了 63.22%、82.89%、64.34%和 65.26%，值得注意的是，Polygon-RNN++模型采用更强大的 ResNet 结构(本文采用 VGG-16 结构).

5.2.3　域外定性结果

图 13 展示了使用 PRN 模型自动标注的城市街景图. 图 14 展示了经过 Cityscapes 数据集训练的

表 7　在 Cityscapes 数据集上的表现结果(IoU 值/%)(不经过任何手工微调)　　(单位：%)

模型	自行车	公交车	行人	列车	卡车	摩托车	汽车	骑手	平均值
DeepMask	47.19	69.82	47.93	62.20	63.15	47.47	61.64	52.20	56.45
SharpMask	52.08	73.02	53.63	64.06	65.49	51.92	65.17	56.32	60.21
Polygon-RNN	52.13	69.53	63.94	53.74	68.03	52.07	71.17	60.58	61.40
Polygon-RNN++	63.06	81.38	**72.41**	64.28	**78.90**	62.01	**79.08**	**69.95**	**71.38**
Polygon-RefineNet	**63.22**	**82.89**	61.84	**64.34**	73.28	**65.26**	73.37	59.20	67.93

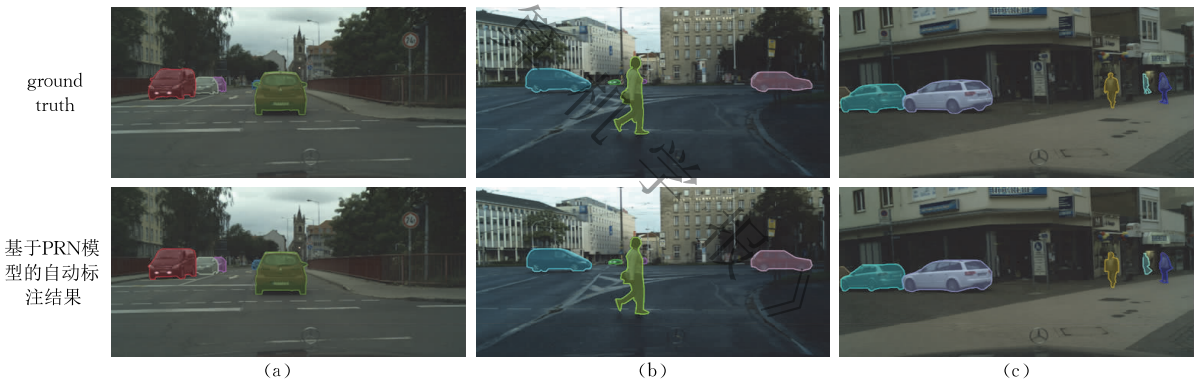


图 13　在 Cityscapes 数据集上的标注结果图(每一组图像((a)(b)(c))中上方的图像为真值图(ground truth)，下方的图像为自动标注结果图(没有经过任何手工微调))

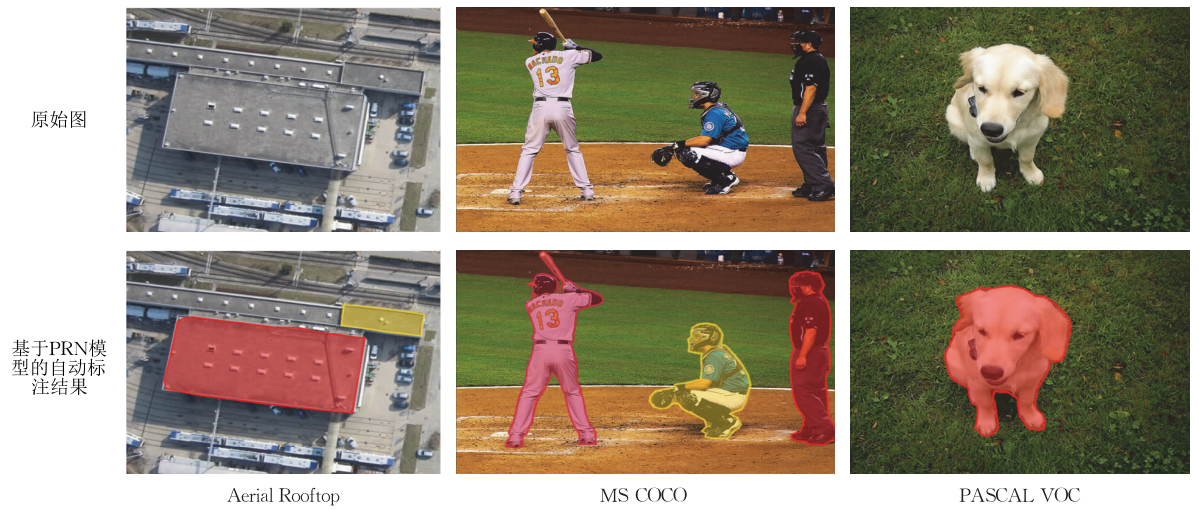


图 14　在未知数据集上的标注结果图(从左到右，图像分别来自 Aerial Rooftop 数据集、MS COCO 数据集和 PASCAL VOC 数据集. 其中上方的图像为原始图，下方的图像为自动标注结果图(没有经过任何手工微调))

PRN 模型分别在 Aerial Rooftop 数据集、MS COCO 数据集和 PASCAL VOC 数据集上自动标注的结果.从图 13、图 14 可知,本文提出的 PRN 模型同样适用于普通彩色图像,并且对未知数据表现出较强的泛化能力.

6 结 论

本文首先提出了一种基于 Polygon-RefineNet 的违禁品 X 线图像自动标注方法,旨在尽可能节省人力和时间的情况下完成更多的违禁品像素级标注.Polygon-RefineNet 模型通过预测目标的多边形轮廓来完成标注,并允许用户在最后手工微调多边形轮廓以得到满意的标注结果,该模型最大的优势就是有效地利用不同层次的特征信息,特别是底层空间信息,其可以帮助模型在背景复杂的违禁品 X 线图像中生成清晰、准确的边界.为了验证提出模型的有效性,本文接着提供了一个带有像素级标注的 X 线图像数据集,PIXray 数据集.实验结果显示,本文的方法在完成精准标注的同时,速度约是纯手工标注的 3.7 倍.在其它域外数据集上的实验证明了本文所提出的方法具有强大的表达能力和泛化能力.

参 考 文 献

[1] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436

[2] Mery D. X-ray testing by computer vision//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Portland, USA, 2013: 360-367

[3] Mery D, Svec E, Arias M, et al. Modern computer vision techniques for X-ray testing in baggage inspection. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2016, 47(4): 682-692

[4] Mery D, Arteta C. Automatic defect recognition in X-ray testing using computer vision//Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Santa Rosa, USA, 2017: 1026-1035

[5] Wang Huai-Ying, Yang Li-Rui, Zhang Yu-Jin. Contraband segmentation of Compton back-scattering images based on CNN. Acta Electronica Sinica, 2011, 39(3): 549-554 (in Chinese)
(王怀颖, 杨立瑞, 章毓晋. 基于 CNN 的康普顿背散射图像中违禁品分割方法. 电子学报, 2011, 39(3): 549-554)

[6] Sun C, Shrivastava A, Singh S, et al. Revisiting unreasonable effectiveness of data in deep learning era//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 843-852

[7] Everingham M, Van Gool L, Williams C K I, et al. The PASCAL visual object classes (VOC) challenge. International Journal of Computer Vision, 2010, 88(2): 303-338

[8] Mottaghi R, Chen X, Liu X, et al. The role of context for object detection and semantic segmentation in the wild//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 891-898

[9] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context//Proceedings of the European Conference on Computer Vision. Zurich, Switzerland, 2014: 740-755

[10] Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 3213-3223

[11] Castrejon L, Kundu K, Urtasun R, et al. Annotating object instances with a polygon-RNN//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 5230-5238

[12] Kuettel D, Guillaumin M, Ferrari V. Segmentation propagation in ImageNet//Proceedings of the European Conference on Computer Vision. Firenze, Italy, 2012: 459-473

[13] Xu J, Schwing A G, Urtasun R. Tell me what you see and I will show you where it is//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 3190-3197

[14] Dutt Jain S, Grauman K. Active image segmentation propagation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 2864-2873

[15] Miao C, Xie L, Wan F, et al. SIXray: A large-scale security inspection X-ray benchmark for prohibited item discovery in overlapping images//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 2119-2128

[16] Chen L C, Fidler S, Yuille A L, et al. Beat the MTurkers: Automatic image labeling from weak 3D supervision//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 3198-3205

[17] Rother C, Kolmogorov V, Blake A. GrabCut: Interactive foreground extraction using iterated graph cuts. ACM Transactions on Graphics, 2004, 23(3): 309-314

[18] Zhang Z, Schwing A G, Fidler S, et al. Monocular object instance segmentation and depth ordering with CNNs//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 2614-2622

[19] Uhrig J, Cordts M, Franke U, et al. Pixel-level encoding and depth layering for instance-level semantic labeling//Proceedings of the German Conference on Pattern Recognition. Hannover, Germany, 2016: 14-25

[20] Wang X, Peng Y, Lu L, et al. ChestX-Ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 2097-2106

[21] Boykov Y Y, Jolly M P. Interactive graph cuts for optimal boundary & region segmentation of objects in ND images//Proceedings of the Eighth IEEE International Conference on Computer Vision. Vancouver, Canada, 2001: 105-112

[22] Shankar Nagaraja N, Schmidt F R, Brox T. Video segmentation with just a few strokes//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 3235-3243

[23] Lin D, Dai J, Jia J, et al. ScribbleSup: Scribble-supervised convolutional networks for semantic segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 3159-3167

[24] Rajchl M, Lee M C H, Oktay O, et al. DeepCut: Object segmentation from bounding box annotations using convolutional neural networks. IEEE Transactions on Medical Imaging, 2016, 36(2): 674-683

[25] Zhang Z, Fidler S, Waggoner J, et al. Superedge grouping for object localization by combining appearance and shape information//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Rhode Island, USA, 2012: 3266-3273

[26] Sun X, Christoudias C M, Fua P. Free-shape polygonal object localization//Proceedings of the European Conference on Computer Vision. Zurich, Switzerland, 2014: 317-332

[27] Duan L, Lafarge F. Towards large-scale city reconstruction from satellites//Proceedings of the European Conference on Computer Vision. Amsterdam, Holland, 2016: 89-104

[28] Acuna D, Ling H, Kar A, et al. Efficient interactive annotation of segmentation datasets with polygon-RNN++//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 859-868

[29] Akçay S, Kundegorski M E, Devereux M, et al. Transfer learning using convolutional neural networks for object classification within X-ray baggage security imagery//Proceedings of the IEEE International Conference on Image Processing. Phoenix, USA, 2016: 1057-1061

[30] Franzel T, Schmidt U, Roth S. Object detection in multi-view X-ray images//Proceedings of the Joint DAGM (German Association for Pattern Recognition) and OAGM Symposium. Graz, Austria, 2012: 144-154

[31] Mery D, Svec E, Arias M. Object recognition in baggage inspection using adaptive sparse representations of X-ray images//Proceedings of the Image and Video Technology. Auckland, New Zealand, 2015: 709-720

[32] Roomi M, Rajashankarii R. Detection of concealed weapons in X-ray images using fuzzy K-NN. International Journal of Computer Science, Engineering and Information Technology, 2012, 2(2): 187-196

[33] Turcsany D, Mouton A, Breckon T P. Improving feature-based object recognition for X-ray baggage security screening using primed visualwords//Proceedings of the IEEE International Conference on Industrial Technology. Cape Town, South Africa, 2013: 1140-1145

[34] Mery D, Riffo V, Zscherpel U, et al. GDXray: The database of X-ray images for nondestructive testing. Journal of Nondestructive Evaluation, 2015, 34(4): 42

[35] Chen L C, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected. arXiv preprint arXiv:1412.7062, 2014

[36] Pohlen T, Hermans A, Mathias M, et al. Full-resolution residual networks for semantic segmentation in street scenes//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 4151-4160

[37] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014

[38] Xingjian S H I, Chen Z, Wang H, et al. Convolutional LSTM network: A machine learning approach for precipitation nowcasting//Proceedings of the Advances in Neural Information Processing Systems. Montréal, Canada, 2015: 802-810

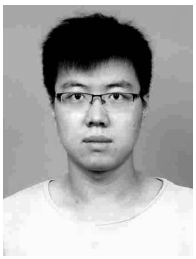
[39] Nekrasov V, Shen C, Reid I. Light-weight RefineNet for real-time semantic segmentation. arXiv preprint arXiv:1810.03272, 2018

[40] Rezatofighi H, Tsoi N, Gwak J Y, et al. Generalized intersection over union: A metric and a loss for bounding box regression//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 658-666

[41] Russell B C, Torralba A, Murphy K P, et al. LabelMe: A database and Web-based tool for image annotation. International Journal of Computer Vision, 2008, 77(1-3): 157-173

[42] Pinheiro P O, Collobert R, Dollár P. Learning to segment object candidates//Proceedings of the Advances in Neural Information Processing Systems. Montréal, Canada, 2015: 1990-1998

[43] Pinheiro P O, Lin T Y, Collobert R, et al. Learning to refine object segments//Proceedings of the European Conference on Computer Vision. Amsterdam, Holland, 2016: 75-91



MA Bo-Wen, M. S. candidate. His research interests include machine learning and computer vision.

JIA Tong, Ph. D., professor. His research interests include machine learning, pattern recognition and computer vision.

LIU Yi-Zhe, M. S. candidate. His research interests include machine learning and computer vision.

HUA Xin-Yu, M. S. candidate. Her research interests include machine learning and computer vision.

Background

In recent years, Artificial Intelligence (AI) has received extensive attention. As a technology of realizing artificial intelligence, deep learning, is triggering the upsurge of research and the development of new technologies in the field of smart security inspection. Deep learning as a method based on representation learning of data, in which data are the basis, however, manually labeling datasets is both time consuming and expensive, especially for pixel-level annotation. Since Boykov et al. proposed the Graph Cuts algorithm in 2006, interactive image segmentation has entered people's vision ensuingly. An accumulating number of approaches for annotating segmentation datasets using interactive image segmentation technology has been proposed, though, but they are applied to annotate natural images or other X-ray images such as medical images. In our knowledge, there is no annotation approach specialized for the security inspection field so far. As a result, this paper proposes an automatic approach based on Polygon-RefineNet for annotation of prohibited items in X-ray images, aiming at speeding up the annotating process and yield high quality annotations. There are numerous overlapping phenomena in prohibited item X-ray images, which leads to the complex background and blurred edges of a prohibited item. Other approaches do not perform satisfying in prohibited item X-ray images. It is well-known that, in a deep convolutional neural network, high-level semantic features helps to extract global and contextual information, which low-level visual features helps to generate sharp, detailed boundaries. Therefore, we introduce a multi-path refinement network to generate high-resolution and detailed results by effectively combining high-level semantics and low-level features. Besides, we also design a mixed loss function for modifying the overall shape and position of the prediction polygon. Through specific experiments, we can prove that our approach has compelling improvements over other existing approaches and can annotate prohibited item instances accurately and quickly.

This research is supported by the National Natural Science Foundation of China (No. U1613214) and the National Key Research and Development Program of China (No. 2018YFB14041).