基于深度帧差卷积神经网络的运动目标检测方法研究

欧先锋[1],晏鹏程[1],王汉谱[1],涂 兵[1],何 伟[1],张国云[1],徐 智[2]
(1. 湖南理工学院信息科学与工程学院机器视觉与人工智能研究中心,湖南岳阳 414006;
2. 桂林电子科技大学广西图像图形智能处理重点实验室,广西桂林 541004)

摘 要: 复杂场景中的运动目标检测是计算机视觉领域的重要问题,其检测准确度仍然是一大挑战. 本文提出并设计了一种用于复杂场景中运动目标检测的深度帧差卷积神经网络(Deep Difference Convolutional Neural Network,DFDCNN). DFDCNN 由 DifferenceNet 和 AppearanceNet 组成,不需要后处理就可以预测分割前景像素. DifferenceNet 具有孪生 Encoder-Decoder 结构,用于学习两个连续帧之间的变化,从输入(t 帧和 $t+1$ 帧)中获取时序信息;AppearanceNet 用于从输入(t 帧)中提取空间信息,并与时序信息融合;同时,通过多尺度特征图融合和逐步上采样来保留多尺度空间信息,以提高网络对小目标的敏感性. 在公开标准数据集 CDnet2014 和 I2R 上的实验结果表明:DFDCNN 不仅在动态背景、光照变化和阴影存在的复杂场景中具有更好的检测性能,而且在小目标存在的场景中也具有较好的检测效果.

关键词: 运动目标检测; 复杂场景; 深度帧差卷积神经网络; 时序信息; 空间信息; 多尺度特征图融合
中图分类号: TP183 **文献标识码:** A **文章编号:** 0372-2112 (2020)12-2384-10
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.12.014

Research of Moving Object Detection Based on Deep Frame Difference Convolution Neural Network

OU Xian-feng[1], YAN Peng-cheng[1], WANG Han-pu[1], TU Bing[1], HE Wei[1], ZHANG Guo-yun[1], XU Zhi[2]
(1. *School of Information and Communication Engineering, Machine Vision & Artificial Intelligence Research Center, Hunan Institute of Science and Technology, Yueyang, Hunan* 414006, *China*;
2. *Guangxi Key Laboratory of Images and Graphics Intelligent Processing, Guilin University of Electronics Technology, Guilin, Guangxi* 541004, *China*)

Abstract: Moving object detection in complex scenes is an important problem in computer vision domain, and the detection accuracy is still a great challenge. In this paper, we propose and design a deep frame difference convolution neural network (DFDCNN) for moving object detection in complex scenes. DFDCNN consists of DifferenceNet and AppearanceNet, which can predict and segment the foreground pixels simultaneously without post-processing. DifferenceNet has Siamese Encoder-Decoder structure, which is used to learn changes between two consecutive frames and to obtain temporal information from inputs, while AppearanceNet is used to extract spatial information from the input frame, and fuse the temporal information and spatial information by fusion of feature maps. Finally, multi-scale spatial information is retained through multi-scale feature map fusion and stepwise up-sampling to improve the sensitivity to small objects. Experiments on two public standard datasets: CDnet2014 and I2R demonstrate that the proposed DFDCNN outperforms the classic algorithms significantly from both qualitative and quantitative aspects. The experimental results illustrate that the proposed DFDCNN shows much better detection performance in complex scenes where dynamic background, illumination variation and shadow exist, and there is improvement for scenes, in which small objects exist.

Key words: moving object detection; complex scenes; deep frame difference convolutional neural network; temporal information; spatial information; multi-scale feature map fusion

---

收稿日期:2020-04-23;修回日期:2020-09-27;责任编辑:覃怀银
基金项目:湖南省自然科学基金项目(No. 2020JJ4340, No. 2020JJ4343);国家自然科学基金(No. 61662014);湖南省教育厅优秀青年项目(No. 19B245);湖南省研究生教育创新工程和专业能力提升工程项目(No. CX20201114);湖南省三维重建与智能应用技术工程研究中心(No. 2019-430602-73-03-006049);湖南省应急通信工程技术研究中心(No. 2018TP2022);广西科技基地和人才专项(No. AD19110022)

1 引言

运动目标检测[1,2]是计算机视觉和数字图像处理领域的基本任务之一,在视频监控[3]、目标跟踪[4]、行人检测[5]等领域有着广泛的应用. 传统的运动目标检测方法是通过计算相邻帧之间、或背景模型与当前帧之间的差异来实现的. 但在背景运动、光照变化、阴影等存在的复杂场景中,传统方法的准确性容易受到影响.

目前关于运动目标检测算法[6~8]的研究,其中有一部分研究是基于帧差法展开的. 帧差法是一种简单快速的方法,它通过比较相邻帧中对应像素之间的差值来检测运动,能有效地检测运动物体,但是在复杂场景中往往会产生误报,实际中往往与其他方法结合使用. Zhou 等人[9]将基于帧差的双向匹配方法与图像滤波方法相结合,提出了一种改进的自适应高斯混合模型,解决了运动目标检测中的目标分割问题. Xu 等人[10]提出了一种改进的高斯混合模型,与三帧帧差法相结合,解决了前景提取中的虚影问题.

近年来,随着深度学习的发展,卷积神经网络(Convolutional Neural Network,CNN)在计算机视觉领域[11,12]得到了广泛的应用,也被用于运动目标检测中. Ou 等人[13]提出了一种基于 ResNet-18 的运动目标检测方法,网络采用编码器-解码器结构,对图像进行分割后再进行运动检测,该方法抑制了背景对前景提取的影响. Lim 等人[14]提出了一种基于编码器-解码器(encoder-decoder)结构 CNN 的背景差分法,并利用轮廓信息优化前景提取. 这些方法主要包括两个阶段:粗匹配(前处理)和精匹配(后处理). 也有一些方法可以一步直接提取出精细的前景,如 A. Dosovitskiy 等人[15]提出将光流估计问题转化为监督学习任务来解决,其设计的 CNNs 网络可以预测光流估计来确定当前物体的运动状态.

这些方法在一定程度上显示了其有效性,但在复杂场景或小目标场景中,它们的性能还可以进一步提高. 本文提出了一种深度帧差卷积神经网络(Deep Frame Difference Convolutional Neural Network, DFDCNN),该网络由两部分组成,第一部分是 DifferenceNet,具有孪生编码器-解码器结构,输入为 t 和 $t+1$ 帧图像,并使用相应的人工标签进行监督;第二部分是 AppearanceNet,具有编码器-解码器结构,输入为 t 帧图像,并使用 $t+1$ 帧对应的 Groundtruth 进行监督. DifferenceNet 编码器生成的特征图将在 AppearanceNet 中融合,并通过多尺度特征图融合和逐步上采样来保留多尺度空间信息,提高对小目标的敏感性.

2 孪生网络

孪生网络(siamese network)从数据中学习相似度度量用来度量对象之间的相似度,也可以反映对象之间的差异[16]. 孪生网络由两个权值和参数共享的主干网络(backbone)组成,任何基础的 CNN(如 VGGNet[17]、ResNet[18]、DenseNet[19]等)都可以作为主干网络. 帧差法是通过比较相邻帧之间的差异来检测运动目标的,因此,可以使用孪生网络来学习相邻帧之间的变化.

设为一幅图像为 $F(X)$,可表示为:

$$F(X)=A_1X^n+A_2X^{n-1}+\cdots A_{n-1}X+A_n \tag{1}$$

其中图像 $F(X)$ 可以看作由高频部分(如 A_1X^n、A_2X^{n-1} 等)和低频部分(如 $A_{n-1}X$、A_n 等)组成的数字信号. 高频部分表示图像中像素快速变化的区域(边缘),低频部分表示平滑区域(纹理). A 表示各项的常系数. 则帧差法可以表示为:

$$\begin{aligned}F(X_t)-F(X_{t+1})&=(A_1X_t^n+A_2X_t^{n-1}+\cdots+A_n)\\&\quad-(B_1X_{t+1}^n+B_2X_{t+1}^{n-1}+\cdots+B_n)\end{aligned} \tag{2}$$

其中 $F(X_t)$ 和 $F(X_{t+1})$ 为连续的两帧图像,A 和 B 分别表示两帧图像中各项的常系数. 在复杂场景中,由于常系数通常不相同,因此帧差法容易产生大量的噪声. 然而,孪生网络可以将低阶的输入图像映射到高维的特征图,通过比较高阶特征来检测运动目标,会大大削弱复杂场景中噪声的影响.

3 深度帧差卷积神经网络

本文结合帧差法和孪生网络用于检测复杂场景中的运动目标,提出了一种深度帧差卷积神经网络(DFDCNN). 网络主要由两部分组成,图 1 为 DFDCNN 的总体结构图.

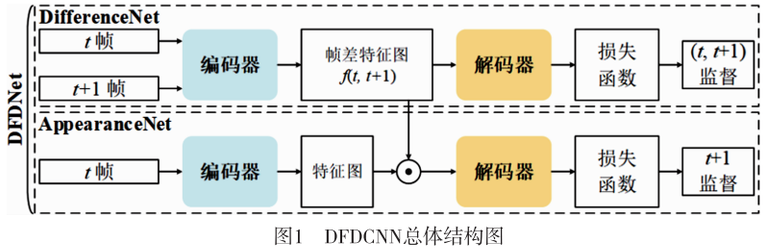

图1　DFDCNN总体结构图

3.1 DifferenceNet

运动目标检测可以看作像素的二分类任务(前景和背景),DifferenceNet 需要学习两帧图像之间的变化以获取时序信息,采用孪生编码器-解码器结构;并且考虑到 ResNet 的优点[13],采用 ResNet 作为主干网络提取特征. 首先训练 DifferenceNet,DifferenceNet 将生成反映了两帧之间的变化的帧差特征图,然后将帧差特征图加入到 AppearanceNet 中进行训练.

3.1.1 DifferenceNet 网络结构

图 2 为 DifferenceNet 网络结构图. 左虚线框内部分表示编码器,右虚线框内部分表示解码器. 编码器通过卷积从输入中提取特征得到特征图,其中包含两个主干网络 Backbone 1 和 Backbone 2(图 2 中蓝色部分),两个主干网络结构相同(都包含 4 个 Residual Block[18]). 将 t 和 $t+1$ 帧分别输入到 Backbone 1 和 Backbone 2,得到相应的特征图 $f(t)$ 和 $f(t+1)$. $f(t,t+1)$ 是由 $f(t)-f(t+1)$ 得到的帧差特征图. 解码器通过 4 次分步上采样将特征图的尺寸还原到与输入相同的大小. 表 1 给出了 DifferenceNet 的网络详细参数,由于 Backbone 1 和 Backbone 2 权值共享,所以没有重复给出相同的参数.

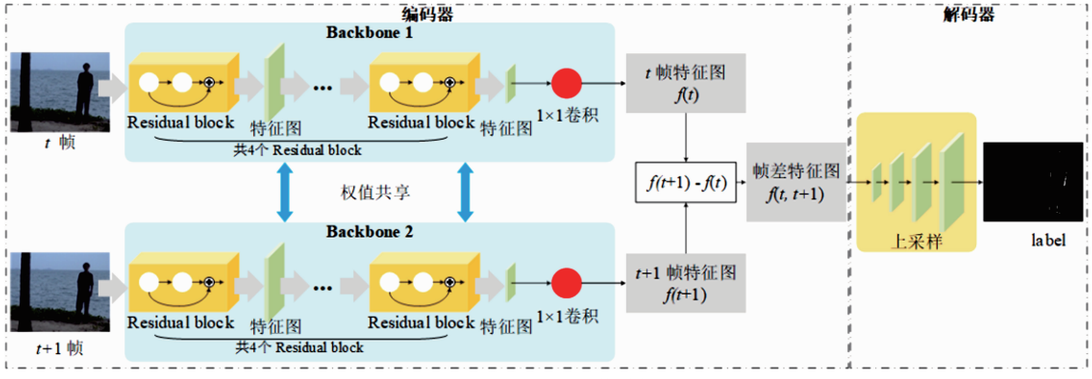

图2 DifferenceNet网络结构图

表 1 DifferenceNet 网络详细参数

	各层名称	输入	输出	输出大小	核大小和通道
编码器	Diff_conv1	t frame, $t+1$ frame	Diff_conv1(t), Diff_conv1($t+1$)	$\frac{M}{2}\times\frac{N}{2}$	$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}$
	Diff_conv2	Diff _conv1(t), Diff _conv1($t+1$)	Diff_conv2(t), Diff_conv2($t+1$)	$\frac{M}{4}\times\frac{N}{4}$	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}$
	Diff_conv3	Diff _conv2(t), Diff _conv2($t+1$)	Diff_conv3(t), Diff_conv3($t+1$)	$\frac{M}{8}\times\frac{N}{8}$	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}$
	Diff_conv4	Diff _conv3(t), Diff _conv3($t+1$)	Diff_conv4(t), Diff_conv4($t+1$)	$\frac{M}{16}\times\frac{N}{16}$	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}$
	Diff_conv5	Diff _conv4(t), Diff _conv4($t+1$)	$f(t)$, $f(t+1)$	$\frac{M}{16}\times\frac{N}{16}$	$[1\times1,2]$
	$f(t,t+1)=f(t+1)-f(t)$				
解码器	Diff_deconv1	$f(t,t+1)$	Diff_deconv1 ($f(t,t+1)$)	$\frac{M}{8}\times\frac{N}{8}$	$[3\times3,2]$
	Diff_deconv2	Diff_deconv1 ($f(t,t+1)$)	Diff_deconv2 ($f(t,t+1)$)	$\frac{M}{4}\times\frac{N}{4}$	$[3\times3,2]$
	Diff_deconv3	Diff_deconv2 ($f(t,t+1)$)	Diff_deconv3 ($f(t,t+1)$)	$\frac{M}{2}\times\frac{N}{2}$	$[3\times3,2]$
	Diff_deconv4	Diff_deconv3 ($f(t,t+1)$)	changes feature map ($t,t+1$)	$M\times N$	$[3\times3,2]$

DifferenceNet 的主干网络通过以下过程提取特征:

$$y_{h,w}^{(o)} = \sum_{o}\sum_{m}\sum_{n}\theta_{m,n}^{(i,o)}x_{((h-1)s-2p+m),((w-1)s-2p+n)}^{(i)} + b^{(o)} \quad (3)$$

其中,x 和 y 分别表示输入和输出的特征图;θ 表示卷积核的权重;b 表示输出通道的偏置;h 和 w 分别表示输出特征图的坐标索引;m 和 n 代表卷积核的坐标索引;o 和 i 分别代表输出和输入的通道;p 是输入的 padding 值;s 是卷积操作的步长 stride 值.

如表 1 所示,DifferenceNet 的主干网络采用了 Residual Block 结构[18]. 以 Diff_conv4 层为例:卷积核大小为 3 × 3,padding 为 0,stride 为 2,通道数为 512. 因此,DifferenceNet 的主干网络 Diff_conv4 层可以表示为:

$$y_{\frac{M}{16},\frac{N}{16}}^{(512)} = \sum_{512}\sum_{3}\sum_{3}\theta_{3,3}^{(256,512)}x_{\frac{M}{8},\frac{N}{8}}^{(256)} + b^{(512)} \quad (4)$$

运动目标检测仅需要分割前景和背景像素,因此 Diff_conv5 层的通道数应为 2,故可以利用 1 × 1 卷积将 $y_{\frac{M}{16},\frac{M}{16}}^{(512)}$ 降维[20]. 对于输入 t,Diff_conv5 层的输出 $f(t)$ 可以表示为:

$$f(t) = y_{\frac{M}{16},\frac{M}{16}}^{(2)} = \sum_{2}\theta^{(512,2)}x_{\frac{M}{16},\frac{M}{16}}^{(512)}(t) + b^{(2)} \quad (5)$$

同理,由于 Backbone 1 和 Backbone 2 共享权值,对于输入 $t+1$,相应的输出 $f(t+1)$ 可以表示为:

$$f(t+1) = y_{\frac{M}{16},\frac{M}{16}}^{(2)} = \sum_{2}\theta^{(512,2)}x_{\frac{M}{16},\frac{M}{16}}^{(512)}(t+1) + b^{(2)} \quad (6)$$

帧差特征图 $f(t,t+1)$ 可以表示为:

$$\begin{aligned} f(t,t+1) &= \left[\sum_{2}\theta^{(512,2)}x_{\frac{M}{16},\frac{N}{16}}^{(512)}(t) + b^{(2)}\right] \\ &\quad - \left[\sum_{2}\theta^{(512,2)}x_{\frac{M}{16},\frac{N}{16}}^{(512)}(t+1) + b^{(2)}\right] \\ &= \sum_{2}\theta^{(512,2)}\left[x_{\frac{M}{16},\frac{N}{16}}^{(512)}(t) - x_{\frac{M}{16},\frac{N}{16}}^{(512)}(t+1)\right] \end{aligned} \quad (7)$$

编码器通过上述过程得到帧差特征图,然后解码器通过上采样将帧差特征图还原到与输入相同的尺寸大小,并密集地对每个像素进行预测. 为了避免一次性上采样造成的信息丢失,解码器通过四次转置卷积将特征图进行上采样.

3.1.2　DifferenceNet 网络输入和监督标签

DifferenceNet 要学习连续图像的变化,因此监督标签(label)必须反映这些变化. 本文利用数据集相应的 Groundtruth 制作 Label,图 3 为制作 Label 的过程.

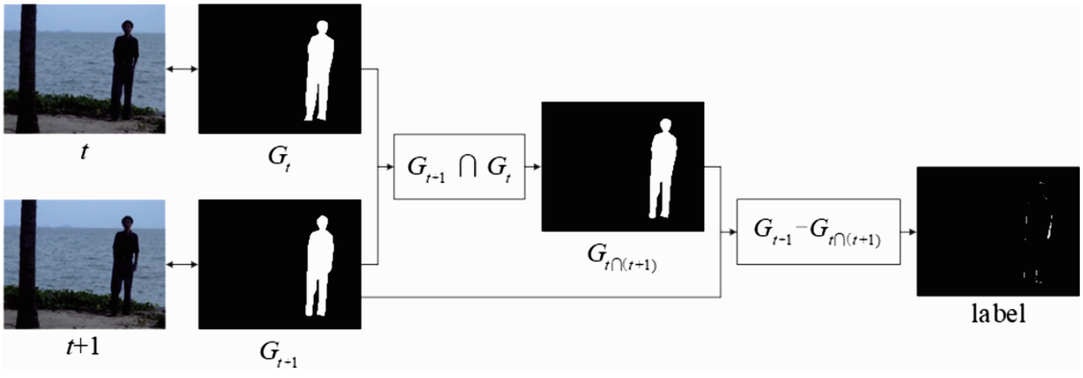

图3　Lable制作过程图

设 t 帧和 $t+1$ 帧是视频序列中连续的两帧图像,其对应的 Groundtruth 为 G_t 和 G_{t+1}. $G_{t\cap(t+1)}$ 为 G_t 和 G_{t+1} 前景区域重叠的部分. 如果物体移动,则它会在 $t+1$ 帧中产生一些新的前景像素,这反映了 t 帧和 $t+1$ 帧前景的变化. 其计算公式如下:

$$\text{Label} = G_{t+1} - G_{t\cap(t+1)} \quad (8)$$

3.2　AppearanceNet

图 4 为 AppearanceNet 的网络结构图.

AppearanceNet 具有编码器-解码器结构,并通过多尺度特征图融合来改进主干网络. AppearanceNet 的编码器由 4 个 Residual Block 组成,解码器通过 4 次转置卷积进行上采样. 表 2 给出了 AppearanceNet 的网络详细参数.

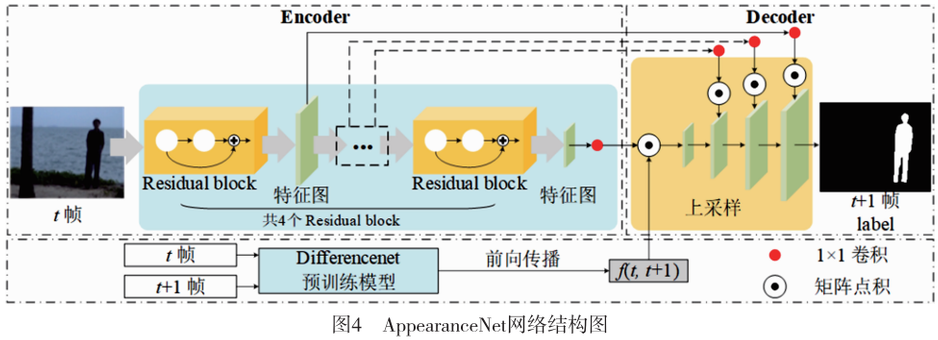

图4　AppearanceNet网络结构图

AppearanceNet 的输入为 t 帧图像,通过编码器提取特征得到特征图 App_conv5(t),其中包含空间信息,帧差特征图 $f(t,t+1)$ 包含时序信息,两者被用于时空信息融合. 此外,为了提高对小目标的敏感性,保留多尺度空间信息,采用了多尺度特征图融合和逐步上采样结构.

3.3 损失函数

由于网络输出为二进制值(前景为 1,背景为 0),因此使用了 Softmax Cross Entropy Loss 损失函数,表示为:

$$L = -\frac{1}{MN}\sum_{M}\sum_{N} G_{MN}\log\left(\frac{e^{Y_{MN}}}{\sum e^{Y_{MN}}}\right) \tag{9}$$

其中 G_{MN} 为 Groundtruth;Y_{MN} 为输出;M 和 N 为输出大小,与输入图像大小相同,且会随输入大小而变化;(x,y) 为特征图中的像素位置索引.

3.4 网络训练

为了加快网络训练速度,在编码器和解码器中对每个卷积层(包括转置卷积层)执行批量归一化(Batch-Norm)[21],激活函数为 Rectified Linear Units (ReLU)[22]. 本文使用 PyTorch 深度学习库对网络进行训练和测试. 为了使网络适应目标的多样性,针对每一个视频序列随机抽取 20% 连续成对的视频帧用于训练.

第一步:DifferenceNet 训练. 输入为 t 帧和 $t+1$ 帧图像,使用由两帧图像 Groundtruth 制作的标签进行监督. 通过 SGD 对网络进行训练,迭代训练 5000 次. 初始学习率为10^{-3},学习率衰减系数为 0.1,每 500 次迭代,学习率衰减一次,最终衰减到10^{-6},Batchsize 设置为 4.

第二步:AppearanceNet 训练. 输入为 t 帧图像,使用 $t+1$ 帧的 Groundturth 作为监督标签. 还需要将 t 帧和 $t+1$ 帧输入到 DifferenceNet 的预训练模型中,得到帧差特侦图 $f(t,t+1)$ 并在 AppearanceNet 中融合. 通过 SGD 训练网络,迭代训练 5000 次. 初始学习率为10^{-3},学习率衰减系数为 0.1,每 500 次迭代,学习率衰减 1 次,最终衰减到10^{-6},Batchsize 设置为 8.

表 2 AppearanceNet 网络详细参数

	各层名称	输入	输出	输出大小	核大小和通道数
编码器	App_conv1	tframe	App_conv1(t)	$\frac{M}{2}\times\frac{N}{2}$	$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}$
	App_conv1_1	App_conv1(t)	App_conv1_1(t)		[1×1,2]
	App_conv2	App_conv1(t)	App_conv2(t)	$\frac{M}{4}\times\frac{N}{4}$	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}$
	App_conv2_1	App_conv2(t)	App_conv2_1(t)		[1×1,2]
	App_conv3	App_conv2(t)	App_conv3(t)	$\frac{M}{8}\times\frac{N}{8}$	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}$
	App_conv3_1	App_conv3(t)	App_conv3_1(t)		[1×1,2]
	App_conv4	App_conv3(t)	App_conv4(t)	$\frac{M}{16}\times\frac{N}{16}$	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}$
	App_conv5	App_conv4(t)	App_conv5(t)	$\frac{M}{16}\times\frac{N}{16}$	[1×1,2]
解码器	App_deconv1	App_conv5(t), $f(t,t+1)$	App_deconv1($t+1$)	$\frac{M}{8}\times\frac{N}{8}$	[3×3,2]
	App_deconv2	App_deconv1($t+1$), App_conv3_1(t)	App_deconv2($t+1$)	$\frac{M}{4}\times\frac{N}{4}$	[3×3,2]
	App_deconv3	App_deconv2($t+1$), App_conv2_1(t)	App_deconv3($t+1$)	$\frac{M}{2}\times\frac{N}{2}$	[3×3,2]
	App_deconv4	App_deconv3($t+1$), App_conv1_1(t)	feature map ($t+1$)	$M\times N$	[3×3,2]

4 实验结果与分析

在本节中,通过计算 F-measure 评价指标来验证所提出的 DFDCNN 的有效性,其计算公式为:

$$F\text{-measure} = \frac{2\cdot \text{Precision}\cdot \text{Recall}}{\text{Precision}+\text{Recall}} \tag{10}$$

其中,Recall 表示被正确检测到的像素中,属于运动目标像素的百分比;Precision 表示运动目标像素中,被正确检测到的百分比. Recall 和 Precision 分别被定义为:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{12}$$

其中 TP 为被检测到的运动目标前景像素对应于 Groundtrurh 中前景像素的像素数量. FN 是没有被检测到的前景像素对应于 Groundtruth 中前景像素的像素数量. FP 是被检查到的像素中对应于 Groundtruth 中背景像素的像素数量.

为了验证 DFDCNN 的有效性,在 I2R[23] 和 CDnet2014[24] 两个数据集上进行实验,包括动态背景、小运动目标、多运动目标、光照变化和阴影等多种复杂场景,从定性和定量两个方面进行说明和对比.

4.1　I2R 数据集上的实验

I2R 数据集包含 9 个不同场景的视频序列,分别是 AirportHall,Bootstrap,ShoppingMall 为静态背景;Lobby 为光照变化;Curtain、Escalator 为室内动态背景;Fountain、Campus、WaterSurface 为室外动态背景. I2R 数据集为每个视频提供 20 帧图像的 Groundtruth. 图 5 为 DFDCNN 与现有其他运动目标检算法的定性结果比较,包括每个视频序列的部分可视化结果. 可以看到,文献[13]、MSFgNet[7] 和 DFDCNN 等基于深度学习的算法能更好的处理背景噪声,相比于文献[13]中算法,DFDCNN 能更好的获取结构信息,因此它检测到的前景目标轮廓细节更加丰富. 表 3 为 DFDCNN 与现有其他运动目标检算法的定量结果比较,计算了每个视频序列的 F-measure 与数据集的平均 F-measure 进行比较,为方便表示,各视频序列做了缩写处理. 在对比方法中,文献[13]算法和 MSFgNet[7] 为基于深度学习的运动目标检测方法. DFDCNN 在 I2R 数据集上的平均 F-measure 达到了 84.95%,且相比于其他算法仅在某些序列上表现较好(如 LSD[26] 在 WaterSurface 上的 F-measure 为 90.50% 而在 Bootstrap 上仅有 58.42%,MSFgNet[7] 在 Curtain 上的 F-measure 为 99.04% 而在 Lobby 上仅有 62.82%),本文提出的 DFDCNN 在各个视频序列上的表现都较为稳定,在 I2R 数据集上优于其它运动目标检测方法.

表 3　I2R 数据集上的 F-measure(%) 性能表(第一:粗体,第二:下划线)

对比算法	Airp	Curt	Boot	Foun	Shop	Camp	Lobby	Esca	Wate	均值
文献[25]	76.44	91.98	70.19	58.55	65.48	82.73	23.72	47.17	88.67	67.21
文献[26]	72.22	85.57	58.42	83.71	73.62	76.13	73.13	72.14	90.50	76.16
文献[27]	77.21	92.54	61.17	82.53	72.43	65.88	**83.47**	66.47	93.14	77.20
文献[7]	69.44	**99.04**	75.92	**95.15**	69.09	86.37	62.82	**84.48**	**99.36**	82.41
文献[13]	72.42	92.12	70.70	86.64	78.42	**94.40**	76.92	64.62	95.82	81.34
本文 DFDCNN	**81.32**	93.24	**80.50**	85.66	**78.66**	93.49	79.41	76.67	95.64	**84.95**

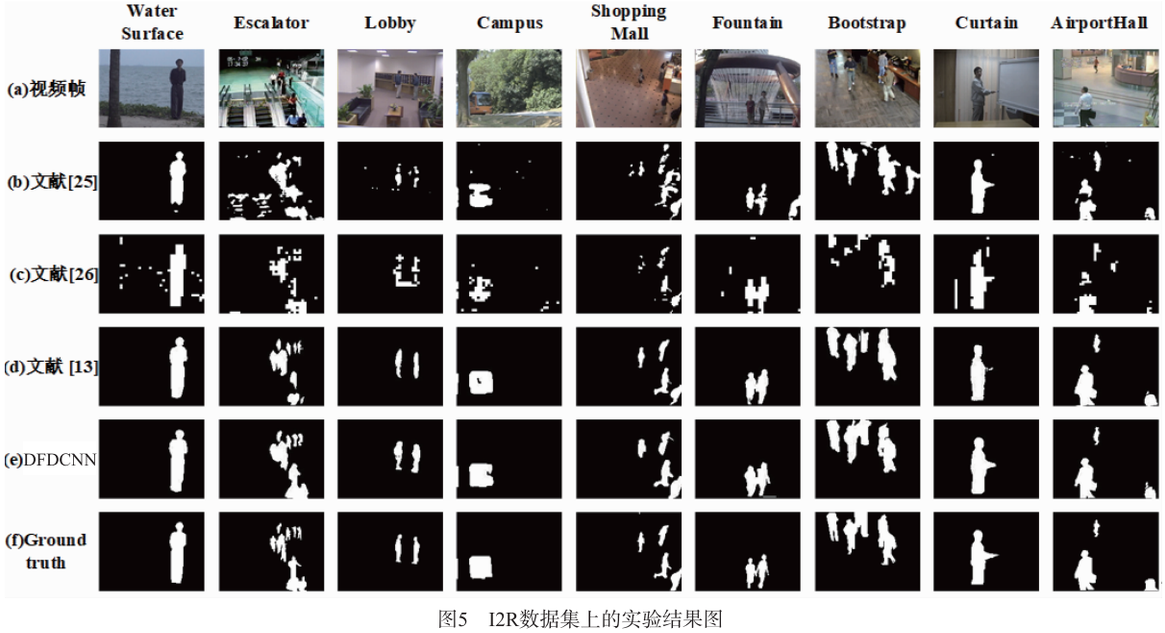

图5　I2R数据集上的实验结果图

4.2 CDnet2014 数据集上的实验

CDnet2014 数据集包含 53 个不同的视频序列,分为 11 类:Bad Weather(BW),Low Framerate(LF),Night Videos(NV),Pan Tilt Zoom(PTZ),Turbulence(Tu),Baseline(Ba),Dynamic Background(DB),Camera Jitter(CJ),Intermittent Object Motion(IOM),Shadow(Sh),Thermal(Th). CDnet2014 数据集为每个视频提供了所有视频帧图像的 Groundtruth. 图 6 为 DFDCNN 与现有其它运动目标检算法的定性结果比较,包括了每个类别的部分视频序列的可视化结果. 表 4 为 DFDCNN 与现有其他运动目标检算法的定量结果比较,计算了每个类别的平均 *F*-measure 和数据集的平均 *F*-measure 进行比较. 在对比算法中,DeepBS[32]、BSUV-Net 和 BSUV-Net + SemanticBGS[33]、MSFgNet[7] 是基于深度学习的方法.

表 4 CDnet2014 数据集上的 *F*-measure(%)性能表(第一:粗体,第二:下划线)

对比算法	BW	LF	NV	PTZ	Tu	Ba	DB	CJ	IOM	Sh	Th	均值
文献[28]	81.36	62.75	49.46	23.22	79.78	93.22	68.57	82.31	69.39	89.69	74.48	70.38
文献[29]	82.89	79.11	51.32	47.03	85.07	95.67	**89.02**	83.32	72.96	90.84	83.03	78.21
文献[30]	82.33	73.74	58.07	45.45	77.35	92.14	86.45	74.11	70.92	87.89	85.81	75.83
文献[31]	86.16	64.45	57.01	33.67	83.04	94.87	83.76	82.28	72.64	89.84	81.52	75.39
文献[32]	83.01	60.02	58.35	31.33	84.55	95.80	87.61	**89.90**	60.98	93.04	75.83	74.48
文献[33]	87.13	67.97	69.87	62.82	70.51	96.93	79.67	77.43	74.99	92.33	85.81	78.68
文献[33]	**87.30**	67.88	68.15	65.62	76.31	96.40	81.76	77.88	76.01	**96.64**	84.55	79.86
文献[7]	85.04	**84.22**	80.99	78.70	**86.42**	92.11	85.14	83.13	77.97	93.45	80.16	84.30
本文 DFDCNN	86.46	80.24	**82.10**	**81.48**	79.86	**97.92**	82.76	85.64	**80.25**	95.34	**85.99**	**85.28**

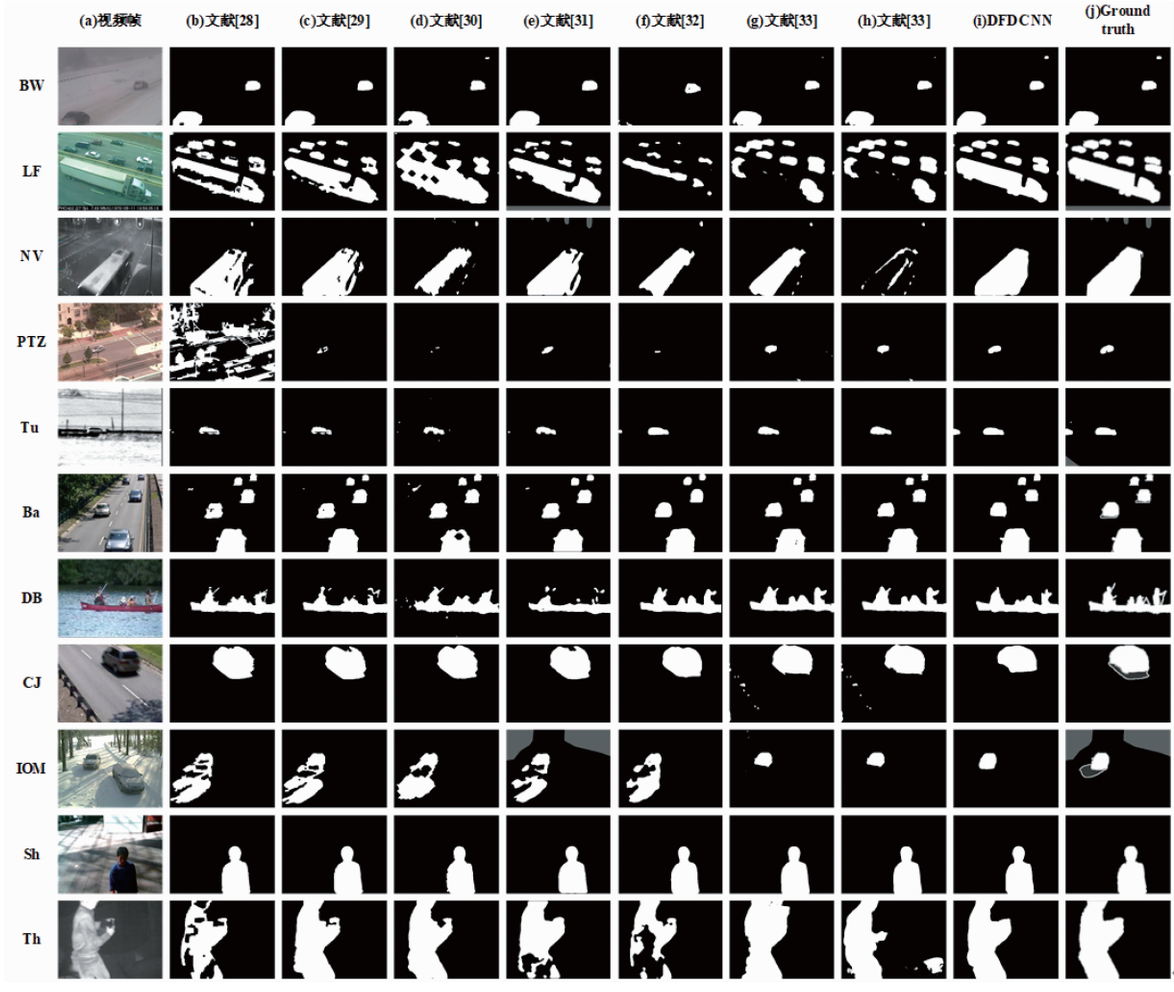

图6 CDnet2014数据集上的实验结果图

由图6所示,相比于 M^4CD[28]、IUTIS-5[29]、SWCD[30]、WisenetMD[31]等算法,对于动态背景视频,DFDCNN 能很好的抑制噪声;相比于 DeepBS[32]、BSUV-Net[33]等基于深度学习的算法,前景更加完整,目标的轮廓细节更丰富;对于存在小运动目标的视频序列 DFDCNN 也能够检测到完整的前景区域. 由表4中可知,本文提出的 DFDCNN 在 CDnet2014 数据集上的平均 F-measure 为 85.28%(共53个视频序列),相比于目前的算法提升了0.98% ~14.9%,在 CDnet2014 数据集上优于其它运动目标检测方法.

5 结论

本文提出并设计了一种用于复杂场景中运动目标检测的深度帧差卷积神经网络(DFDCNN),该网络由具有编码器-解码器结构的 DifferenceNet 和 AppearanceNet 两部分组成. DFDCNN 利用 DifferenceNet 学习 t 帧与 $t+1$ 帧之间的变化以获取时序信息;AppearanceNet 通过融合由 DifferenceNet 生成的帧差特征图,将时序和空间信息结合起来,通过融合多尺度空间信息,使网络对小目标更加敏感. DFDCNN 不需要后处理可以直接预测分割出精细的运动前景,在 I2R 和 CDnet2014 数据集上的实验结果表明,本文提出的 DFDCNN 具有良好的运动目标检测性能. 由于 DifferenceNet 与 AppearanceNet 分为两部分训练,会大大增加训练时间,在后续的研究中将两部分网络进行融合,研究更为简洁的网络模型. 另外,当物体运动缓慢时,相邻两帧图像的变化十分微小,监督标签难以反映这些差异,从而导致 DifferenceNet 不能学习到准确的帧间变化关系而产生误检. 因此,在后续工作中要研究更加合理的监督标签制作方法.

参考文献

[1] Yazdi M, Bouwmans T. New trends on moving object detection in video images captured by a moving camera: A survey [J]. Computer Science Review, 2018, 28: 157 – 177.

[2] Joshi K A, Thakore D G. A survey on moving object detection and tracking in video surveillance system [J]. International Journal of Soft Computing and Engineering, 2012, 2(3): 44 – 48.

[3] Mabrouk A B, Zagrouba E. Abnormal behavior recognition for intelligent video surveillance systems: A review [J]. Expert Systems with Applications, 2018, 91: 480 – 491.

[4] Wang Q, Zhang L, et al. Fast online object tracking and segmentation: A unifying approach [A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition [C]. Long Beach, USA: ACM, 2019. 1328 – 1338.

[5] Zeng Z, Li Z, et al. Two-stream multirate recurrent neural network for video-based pedestrian reidentification [J]. IEEE Transactions on Industrial Informatics, 2017, 14(7): 3179 – 3186.

[6] Li L, Hu Q, et al. Moving object detection in video via hierarchical modeling and alternating optimization [J]. IEEE Transactions on Image Processing, 2018, 28(4): 2021 – 2036.

[7] Patil P W, Murala S. Msfgnet: A novel compact end-to-end deep network for moving object detection [J]. IEEE Transactions on Intelligent Transportation Systems, 2018, 20(11): 4066 – 4077.

[8] Lim L A, Keles H Y. Foreground segmentation using convolutional neural networks for multiscale feature encoding [J]. Pattern Recognition Letters, 2018, 112: 256 – 262.

[9] Wei Z, Li P, et al. A foreground-background segmentation algorithm for video sequences [A]. Proceedings of the International Symposium on Distributed Computing and Applications for Business, Engineering and Science [C]. Washington DC, USA: ACM, 2015. 340 – 343.

[10] Xu Z, Zhang D, et al. Moving object detection based on improved three frame difference and background subtraction [A]. Proceedings of the International Conference on Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration [C]. Wuhan, China: ACM, 2017. 79 – 82.

[11] Xiang C, Zhang L, et al. MS-CapsNet: A novel multi-scale capsule network [J]. IEEE Signal Processing Letters, 2018, 25(12): 1850 – 1854.

[12] Han C, Duan Y, et al. Dense convolutional networks for semantic segmentation [J]. IEEE Access, 2019, 7: 43369 – 43382.

[13] Ou X, Yan P, et al. Moving object detection method via ResNet-18 with encoder-decoder structure in complex scenes [J]. IEEE Access, 2019, 7: 108152 – 108160.

[14] Lim K, Jang W D, et al. Background subtraction using encoder-decoder structured convolutional neural network [A]. Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance [C]. Lecce, Italy: ACM, 2017. 1 – 6.

[15] Dosovitskiy A, Fischer P, et al. Flownet: Learning optical flow with convolutional networks [A]. Proceedings of the IEEE International Conference on Computer Vision [C]. Washington DC, USA: ACM, 2015. 2758 – 2766.

[16] Chopra S, Hadsell R, et al. Learning a similarity metric discriminatively, with application to face verification [A]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition [C]. San Diego, USA: ACM, 2005. 539 – 546.

[17] Simonyan K, Zisserman A. Very Deep Convolutional Net-

works for Large-scale Image Recognition [DB/OL]. arXiv preprint arXiv:1409.1556,2014.
[18] He K, Zhang X, et al. Deep residual learning for image recognition[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Las Vegas, USA: ACM, 2016. 770 – 778.
[19] Huang G, Liu Z, et al. Densely connected convolutional networks[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Honolulu, USA: ACM, 2017. 4700 – 4708.
[20] Long J, Shelhamer E, et al. Fully convolutional networks for semantic segmentation[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition [C]. Boston, USA: ACM, 2015. 3431 – 3440.
[21] Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift [DB/OL]. arXiv preprint arXiv:1502.03167, 2015.
[22] Nair V, Hinton G E. Rectified linear units improve restricted boltzmann machines[A]. Proceedings of the International Conference on Machine Learning[C]. Washington DC, USA: ACM, 2010. 807 – 814.
[23] Li L, Huang W, et al. Statistical modeling of complex backgrounds for foreground object detection[J]. IEEE Transactions on Image Processing, 2004, 13(11): 1459 – 1472.
[24] Wang Y, Jodoin P M, et al. CDnet 2014: An expanded change detection benchmark dataset[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Columbus, USA: ACM, 2014. 387 – 394.
[25] Maddalena L, Petrosino A. The 3dSOBS + algorithm for moving object detection[J]. Computer Vision and Image Understanding, 2014, 122: 65 – 73.
[26] Liu X, Zhao G, et al. Background subtraction based on low-rank and structured sparse decomposition[J]. IEEE Transactions on Image Processing, 2015, 24(8): 2502 – 2514.
[27] Yong H, Meng D, et al. Robust online matrix factorization for dynamic background subtraction[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(7): 1726 – 1740.
[28] Wang K, Gou C, et al. M4CD: A robust change detection method for intelligent visual surveillance[J]. IEEE Access, 2018, 6: 15505 – 15520.
[29] Bianco S, Ciocca G, et al. Combination of video change detection algorithms by genetic programming[J]. IEEE Transactions on Evolutionary Computation, 2017, 21(6): 914 – 928.
[30] Isik S, Özkan K, et al. SWCD: a sliding window and self-regulated learning-based background updating method for change detection in videos[J]. Journal of Electronic Imaging, 2018, 27(2): 023002.
[31] Lee S, Lee G, et al. Wisenetmd: Motion detection using dynamic background region analysis [J]. Symmetry, 2019, 11(5): 621.
[32] Babaee M, Dinh D T, et al. A deep convolutional neural network for video sequence background subtraction[J]. Pattern Recognition, 2018, 76: 635 – 649.
[33] Tezcan O, Ishwar P, et al. BSUV-Net: a fully-convolutional neural network for background subtraction of unseen videos[A]. IEEE Winter Conference on Applications of Computer Vision[C]. Snowmass, USA: ACM, 2020. 2774 – 2783.

作者简介

欧先锋 男,1983 年 7 月生于湖南郴州. 现为湖南理工学院副教授、硕士生导师. 主要研究方向为计算机视觉、高光谱遥感图像处理.
E-mail: ouxf@hnist.edu.cn

晏鹏程 男,1995 年 6 月生于湖南益阳. 现为湖南理工学院硕士生. 主要研究方向为深度学习框架与算法、计算机视觉.
E-mail: 530865028@qq.com

王汉谱 男,1997 年 4 月生于江苏盐城. 现为湖南理工学院硕士生. 主要研究方向为深度学习框架与算法、计算机视觉.
E-mail: 1215051195@qq.com

涂 兵 男,1983 年 1 月生于湖南岳阳. 现为湖南理工学院副教授、硕士生导师. 主要研究方向为计算机视觉、高光谱遥感图像处理.
E-mail: tubing@hnist.edu.cn

何　伟　男,1983 年 1 月生于湖南岳阳. 现为湖南理工学院副教授、硕士生导师. 主要研究方向为计算机视觉、机器学习.
E-mail:hewei@ hnist. edu. cn

张国云(通信作者)　男,1971 年 1 月生于湖南郴州. 现为湖南理工学院教授、硕士生导师. 主要研究方向为计算机视觉.
E-mail:gyzhang@ hnist. edu. cn

徐　智(通信作者)　男,1971 年 1 月生于四川眉山. 现为桂林电子科技大学副研究员、硕士生导师. 主要研究方向机器学习.
E-mail:xuzhi@ guet. edu. cn