计算机科学
COMPUTER SCIENCE

基于深度学习的红外视频显著性目标检测

朱叶, 郝应光, 王洪玉

引用本文

朱叶, 郝应光, 王洪玉. 基于深度学习的红外视频显著性目标检测[J]. 计算机科学, 2023, 50(9): 227-234.

ZHU Ye, HAO Yingguang, WANG Hongyu. Deep Learning Based Salient Object Detection in Infrared Video [J]. Computer Science, 2023, 50(9): 227-234.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220700204

基于深度学习的红外视频显著性目标检测

朱 叶 郝应光 王洪玉
大连理工大学信息与通信工程学院 辽宁 大连 116024
(zhuye01020928@163.com)

摘 要 面对背景越来越复杂的海量红外视频图像,传统方法的显著性目标检测性能不断下降。为了提升红外图像的显著性目标检测性能,提出了一种基于深度学习的红外视频显著性目标检测模型。该模型主要由空间特征提取模块、时间特征提取模块、残差连接块以及像素级分类器4个模块组成。首先利用空间特征提取模块获得空间特征,然后利用时间特征提取模块获得时间特征并实现时空一致性,最后将时空特征信息和由残差连接块连接空间模块获得的空间低层特征信息一同送入像素级分类器,生成最终的显著性目标检测结果。训练网络时,使用BCEloss和DICEloss两个损失函数结合的方式,以提高模型训练的稳定性。在红外视频数据集OTCBVS以及背景复杂的红外视频序列上进行测试,结果表明所提模型都能够获得准确的显著性目标检测结果,并且具有鲁棒性及较好的泛化能力。

关键词:红外视频;显著性目标检测;深度学习;卷积神经网络;损失函数
中图法分类号 TP751

Deep Learning Based Salient Object Detection in Infrared Video

ZHU Ye,HAO Yingguang and WANG Hongyu
School of Information and Communication Engineering,Dalian University of Technology,Dalian,Liaoning 116024,China

Abstract In the face of massive infrared video images with more and more complex background,the performance of the traditional methods for salient object detection decreases significantly. In order to improve the performance of salient object detection in infrared images,this paper proposes a deep learning-based salient object detection model for infrared video,which mainly consists of a spatial feature extraction module,a temporal feature extraction module,a residual skip connection module and a pixel-wise classifier. First,the spatial feature extraction module is used to extract spatial saliency features from raw input video frames. Secondly,the temporal feature extraction module is used to obtain temporal saliency features and spatio-temporal coherence modeling. Finally,the spatial-temporal feature information and the spatial low-level feature information obtained by connecting the spatial module with the residual skip connection layer are sent into the pixel-wise classifier to generate the final salient object detection results. To improve the stability of the model,BCEloss and DICEloss are combined to train the network. The test is carried out on infrared video dataset OTCBVS and infrared video sequences with complex background. The proposed model can obtain accurate salient object detection results,and has robustness and good generalization ability.

Keywords Infrared video,Salient object detection,Deep learning,Convolutional neural network,Loss function

1 引言

目前,红外成像技术不断发展,其因环境适应能力强、穿透力高,在许多军用、民用领域中得到广泛应用[1]。红外成像技术是红外搜索与跟踪、红外预警、精确制导等应用的关键技术,红外目标检测任务已经成为红外图像处理领域的研究热点[2]。在目标先验情况未知的条件下,目标检测任务利用目标在场景中的某种特性来实现:一种是检测场景中的运动目标,根据检测背景的不同将其划分为基于静态背景的运动目标检测和基于动态背景的运动目标检测;另一种是检测场景中的显著性目标,显著性目标即图像或者视频中最吸引人注意力的部分。显著性目标检测技术的处理流程包含两个部分,首先从图像或者视频中检测出最显著的目标,然后从图像或者视频中准确地将目标分割出来。显著性目标检测可以作为很多图像处理任务的预处理过程,如目标分割[3]、动作识别[4]和目标追踪[5]等。本文主要针对红外视频的显著性目标检测,开展基于深度学习的红外视频显著性目标检测方法研究。

到稿日期:2022-07-24 返修日期:2022-11-08
基金项目:中央高校基本科研业务费专项基金(DUT21GF204)
This work was supported by the Fundamental Research Funds for the Central Universities of Ministry of Education of China(DUT21GF204).
通信作者:郝应光(yghao@dlut.edu.cn)

传统的显著性检测算法一般都是利用图像的颜色、梯度、纹理等较低层次的空间信息提取显著性目标。之后提出了基于任务驱动的检测算法，这一方法依赖于目标的先验知识、图片本身的结构等，所以通常需要大量的数据。常见的算法包括 Hou 等[6]和 Achanta 等[7]分别基于频谱残差和频率调谐提出的对应的显著性检测算法、Cheng 等[8]和 Rahtu 等[9]利用图像的对比度分别提出的基于图像全局对比度和半局部区域的显著性检测算法、Han 等[10]提出的一种改进后的基于图片局部对比度的显著性检测算法。这些传统的显著性检测算法在简单的显著性任务上能够获得不错的检测效果，但是在面对背景越来越复杂的海量红外图像分析时不再适用。为了克服传统显著性检测方法带来的问题，一些新的理论研究和方法逐渐被提出。

深度学习的发展，带动了视频显著性目标检测领域的研究。相较于传统方法，基于深度学习的显著性检测方法不再依赖人为设计特征，可以自动学习有助于显著性检测的特征。一些经典的基于神经网络的算法被提出，例如，Wang 等[11]提出了一个由静态和动态网络两部分组成的全卷积网络，利用静态网络得到静态显著性检测图，然后将静态显著图与视频帧对相结合，经过动态网络生成最终的显著性图，但是这种类型的网络只考虑了相邻视频帧的信息；Simonya 等[12]提出了一个用于实现视频动作识别的网络，利用双流网络提取视频中的时间和空间特征，这一研究启发了人们融合提取到的空间特征和时间特征信息用于生成显著性图。由此，Li 等[13]提出了一个双分支预测网络，这两个分支分别利用视频帧信息和视频帧的光流图进行显著性预测，并通过注意力模块将运动信息补充到显著性分支，获得了较为准确的显著性结果。考虑到对时间信息和空间信息分开建模可能会导致两个信息不一致，针对同时建模时间和空间信息的研究随之展开。Fan 等[14]提出了一个由金字塔扩张卷积模块和显著性转移感知 ConvLSTM 模块组成的模型，先利用金字塔扩张卷积模块提取空间特征信息，再利用显著性转移感知 ConvLSTM 模块捕获时间信息并进行显著性预测。Li 等[15]提出了一个光流引导的递归神经编码器框架，利用光流网络来提取运动信息，并利用 ConvLSTM 来实现视频特征的时间一致性，从而提升视频显著性目标检测的性能。但是目前提出的基于神经网络的显著性检测网络的应用场景大多都是基于可见光条件的，关于红外场景下的应用研究却很少，考虑到红外图像和可见光图像之间差异较大，不能将提出的基于可见光的显著性目标检测网络直接应用于红外视频。

针对上述问题，提出了一种基于深度学习的红外视频显著性目标检测模型。本文的创新点在于：

1）提出了一种基于深度学习的红外视频显著性目标检测模型，该模型在红外视频数据集 OTCBVS 以及背景复杂的真实红外视频序列上都能够获得准确的显著性目标检测结果，且具有鲁棒性和较好的泛化能力。

2）考虑到红外视频对比度低等特点，可能会导致空间特征模块将无关的杂乱背景也当成目标特征提取，因此在空间特征模块中添加注意力模块 CBAM，利用该模块可以使网络更准确地聚焦于目标对象，抑制无关背景带来的影响。

3）训练网络模型时，使用 BCEloss 和 DICEloss 两种损失函数相结合的方式，提高模型的稳定性，也在一定程度上提升了模型的性能。

2 本文方法

本文提出了一个基于深度学习的红外视频显著性目标检测模型，整体模型由空间特征提取模块（Spatial Feature Extractor Module）、时间特征提取模块（Temporal Feature Extraction Module）、残差连接模块（Residual Skip Connection Layer）以及像素级分类器（Pixel-wise Classifier）4 个部分组成。具体来讲，网络输入红外视频帧后，首先利用空间特征提取模块获得红外视频帧的空间特征，该模块包括 ResNet-50，CBAM，ASPP 这 3 个部分；然后利用时间特征提取模块提取时间特征并实现时空一致性，包括 DB-ConvGRU 和 Non-local block 两部分；最后将时空特征信息和由残差连接块连接 ResNet-50 获得的空间低层特征信息一同送入像素级分类器，生成最终的显著性目标检测结果，整体网络框架如图 1 所示。

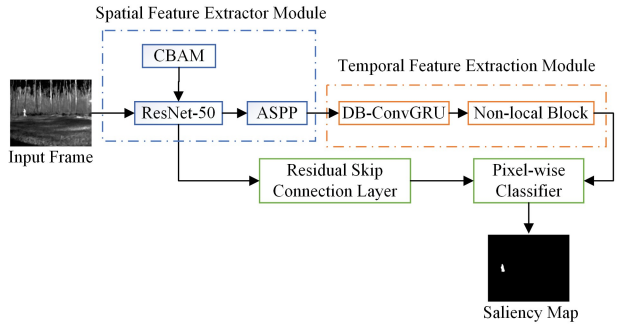


图 1 整体网络框图

Fig. 1 Overall network diagram

2.1 空间特征提取模块

空间特征提取模块由 ResNet-50[16]，CBAM[17]，ASPP[18]这 3 个模块组成，使用 ResNet-50 的前 5 个组层提取特征信息，这 5 个组层可表示为 starting stage，stage1，stage2，stage3 和 stage4。为了减少空间特征的损失，stage4 层不再进行下采样操作。为了使网络更准确地聚焦于目标对象，抑制杂乱背景带来的影响，在 ResNet-50 的每一个 stage 中加入 CBAM 注意力模块。为了在网络中获得更高级的图像特征，将一个空洞卷积空间金字塔池模块（ASPP）附加到 ResNet-50 网络的最后一层，ASPP 模块可以增强感受野，使网络更好地获取多尺度的上下文信息。

其中 CBAM 模块是一种轻量级的卷积神经网络注意力模块，它结合了通道注意力机制及空间注意力机制，通过引入注意力机制可以使网络更准确地聚焦于目标对象，抑制无关背景的干扰，进而提升模型的性能。为了能够使用 ImageNet 预先训练过的 ResNet-50，本文将 CBAM 模块插入到 ResNet-50 的每一个 block 之后，图 2 展示了将 CBAM 模块插入到 ResNet-50 中的具体位置。

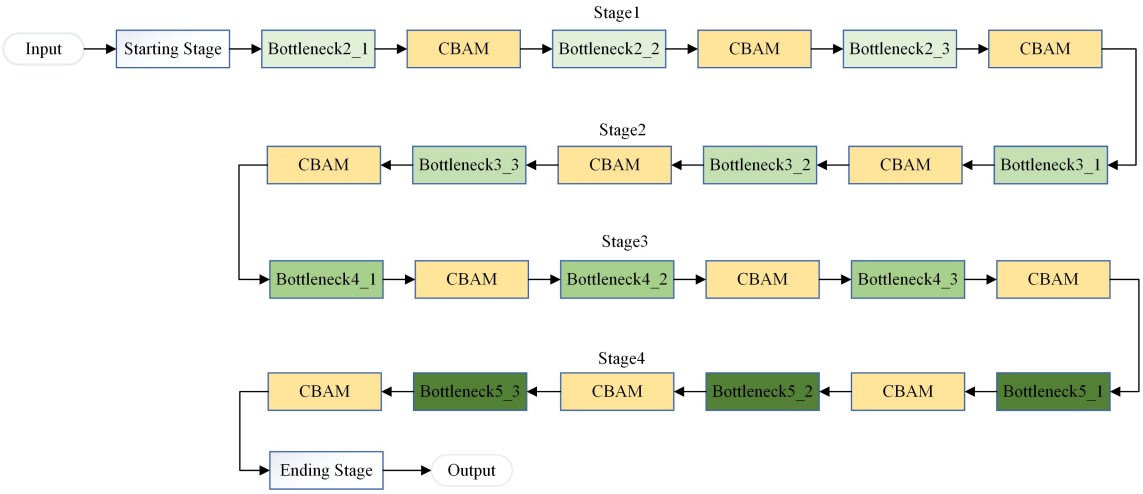


图 2　ResNet-50_CBAM 框图

Fig. 2　ResNet-50_CBAM diagram

2.2　时间特征提取模块

为了充分利用红外视频帧所包含的时间特征信息，本文设计了时间特征提取模块，该模块由深层双向 ConvGRU（DB-ConvGRU）模块和 Non-local block 模块[19]两个部分组成，利用时间模块提取时间特征信息并增强时空一致性。

ConvGRU[20]是在 GRU[21]的基础上改进的，循环神经网络（GRU）是 LSTM[22]的一种变形，它摒弃了 LSTM 中的记忆单元并将输入门和遗忘门结合为更新门，相比 LSTM，能够在提升训练速度的同时保持精度基本不变。LSTM 和 GRU 通常用来处理时序数据，它们无法处理包含丰富空间信息或者与周围的点有着较强相关性的图像，这也意味着 LSTM 和 GRU 可能会丢失较多的空间特征信息。为了利用 GRU 和 LSTM 构建时空序列的预测模型，将 LSTM 和 GRU 的全连接改为卷积，称为 ConvLSTM 和 ConvGRU。ConvGRU 的计算公式如下：

$$Z_t=\sigma(\boldsymbol{W}_{xz}*X_t+\boldsymbol{W}_{hz}*H_{t-1}) \tag{1}$$

$$R_t=\sigma(\boldsymbol{W}_{xr}*X_t+\boldsymbol{W}_{hr}*H_{t-1}) \tag{2}$$

$$H'_t=\tanh(\boldsymbol{W}_{xh}*X_t+R_t\circ(\boldsymbol{W}_{hh}*H_{t-1})) \tag{3}$$

$$H_t=(1-Z_t)\circ H'_t+Z_t\circ H_{t-1} \tag{4}$$

其中‘*’表示卷积运算符，‘∘’表示哈达玛积，$\sigma(\cdot)$表示激活函数，$\boldsymbol{W}$ 表示可学习的权重矩阵，为便于注释，省略了偏差项。

Song 等[23]利用 PDB-ConvLSTM 模型捕获互补的时空特征，本文将两个 ConvGRU 模块按照向前和向后两个方向堆叠起来构成深层双向 ConvGRU 模块，用于加强两个方向之间的时空信息交换。图 3 为深层双向 ConvGRU 模块，该模块可以获取过去和未来的序列特征信息。具体实现公式如下：

$$H_t^f=ConvGRU(H_{t-1}^f,X_t) \tag{5}$$

$$H_t^b=ConvGRU(H_{t+1}^b,H_t^f) \tag{6}$$

$$H_t=\tanh(\boldsymbol{W}_{hf}*H_t^f+\boldsymbol{W}_{hb}*H_t^b) \tag{7}$$

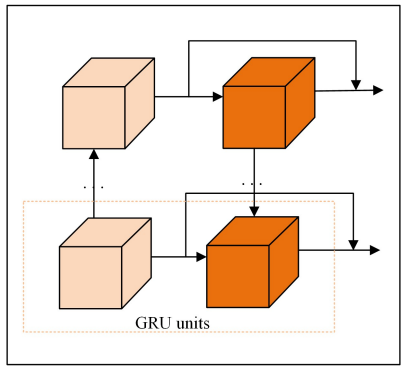


图 3　深层双向 ConvGRU 模块

Fig. 3　DB-ConvGRU module

Non-local block 模块从传统非局部均值方法[24]中获得灵感，该模块将某个位置的响应计算为输入特征映射的所有位置的加权和，在捕获时间（一维时序信号）、空间（图片）和时空（视频序列）的长范围依赖的同时保证输入尺度和输出尺度不变，可以利用该模块在输入的红外视频帧的特征之间建立时空连接。因此，本文将 Non-local block 模块添加到深层双向 ConvGRU 模块后，以增强时间特征提取模块的时空一致性。

2.3　残差连接块及像素级分类器

本文中像素级分类器的输入包含两部分，分别是时间特征模块提取到的特征信息和残差连接块连接 ResNet-50 获得的空间低层特征信息。像素级分类器将输入的特征信息解码生成最终的显著性目标检测结果，具体连接方式如图 4 所示。

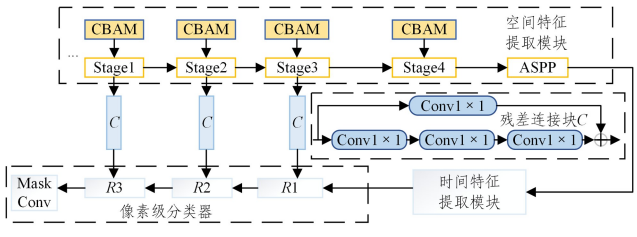


图 4　具体连接图

Fig. 4　Concrete connection diagram

其中像素级分类器由 R1,R2,R3 这 3 个细化块级联组成,每一个细化块通过残差连接块连接 ResNet-50 中的一个层,主要目的是减少细化块的下采样导致的空间特征细节信息的丢失。残差连接块是一种被称为残差跳跃连接层的残差瓶颈体系结构[25],它可以将更多的空间信息带到细化块中,更好地实现像素级显著推理。

2.4 损失函数

在一般的显著性目标检测中,通过计算真值和预测显著图的交叉熵损失函数(BCEloss)计算损失,具体计算公式为:

$$L_B=-\sum(Y_{x,y}\ln(S_{x,y})+(1-Y_{x,y})\ln(1-S_{x,y})) \tag{8}$$

由于卷积神经网络中的尺度变化引起的等级不平衡问题会削弱二值交叉熵的影响,预测的空间不一致,因此考虑引入 DICEloss。该损失函数适用于图像的二值分割,且一定程度上能缓解正负样本在数量上不平衡的问题。该损失函数的计算公式为:

$$Dice=1-\frac{2\sum_{i=1}^{N}y_i\cdot\hat{y}_i}{\sum_{i=1}^{N}y_i+\sum_{i=1}^{N}\hat{y}_i} \tag{9}$$

其中,y_i 与 $\hat{y}_i$ 分别是像素 i 的标签和预测值,N 为像素点的点数。

为了强调前景背景的差异并保持类内一致性,选择使用 BCEloss 和 DICEloss 两个损失函数结合的方式,使模型一致地推动整个显著区域,并更好地处理因各种物体比例不同而出现的前后区域之间的像素不平衡的问题,而无需任何后处理或者额外的参数。最后使用的损失函数为:

$$L=L_{BCEL}(p,g)+L_{CEL}(p,g) \tag{10}$$

其中,p 为图像的预测值,g 为图像的真值。

3 实验结果及分析

3.1 实验配置

3.1.1 数据集

由于红外数据集在军事、国防等领域有特殊用途,目前尚无适用的可用于训练的公开红外数据集,因此本文算法利用公开数据集 DAVIS2016[26] 和 VOS[27] 中的训练集进行训练,利用 VOS 数据集中的验证集进行验证。其中 DAVIS2016 是视频物体分割的数据集,包含 50 个高质量视频序列,有 3455 张密集标注的像素级别的帧;VOS 数据集是一个由 200 个视频组成的基于视频的显著性目标检测数据集。

在验证模型性能时,利用红外数据集进行测试,目前公开的红外数据集有 KAIST 行人检测数据集、FLIR 红外目标识别数据集,以及 OTCBVS 红外数据集。KAIST 数据集共有 95328 个可见光-红外图像对,分为 4 种类型,总共有 103128 个标注、1182 个人,主要用于行人检测任务;FLIR 红外目标识别数据集包含 10000 张可见光-红外图像对,包含 4 种类型,其中训练集有 67618 张、测试集 11696 张;OTCBVS 红外数据集用于计算机视觉算法的研究,共有 13 个子数据集,包含超过 20000 张图像。

其中 KAIST 行人检测数据集适用于目标检测任务,不适用于红外视频的显著性目标检测;FLIR 红外目标识别数据集没有对准,使用前需要自行校正,处理过程较为复杂且容易影响测试结果。因此本文测试时选择使用 OTCBVS 数据集。OCTBVS 数据集包含行人数据库、Terravic 面部红外数据库等不同类型的 14 个小型数据集,选择 Dataset01-Dataset 01-OSU Thermal Pedestrian Database(行人红外数据库)和 Dataset05-Terravic Motion IR Database(运动红外数据库)作为本次测试的数据集,表 1 列出了这两个小型数据集的具体信息。

表 1 红外数据测试集

Table 1 Infrared data test set

红外数据集子序列	序列个数	简要介绍
Dataset01:OSU Thermal Pedestrian Database	10	主要用于红外行人检测任务,共有 10 个相似的序列
Dataset05:Terravic Motion IR Database	18	主要用于红外图像的检测和跟踪任务,共有 18 个序列,包含室外目标(2/1 个行人)跟踪,室内室外监控视频、飞机运动和跟踪、水下和水面运动、背景运动(由于强风云和树木的运动)

考虑到 Dataset01:OSU Thermal Pedestrian Database 中有 10 个相似序列,在验证模型性能时,选择其中的一个序列进行测试,该序列共有 72 帧图像。为了对检测结果进行定量分析,使用 LabelMe 软件为该红外序列标注显著性真值。选择 Dataset 05-Terravic Motion IR Database 数据集中的 irw09 和 irw10 两个红外视频序列进行测试,相比 irw10 的背景,irw09 的背景更为杂乱。

3.1.2 训练环境及训练参数

本次实验基于 Pytorch1.0.1 的框架实现,在 Ubuntu 中用 python 进行实验,训练时初始学习率为 10^{-5},batch size 默认为 1,训练 100 个 epoch,将 BCEloss 和 DICEloss 两个损失函数结合起来,作为本次训练的损失函数,在这种设置下使用 GeForce GTX2080Ti GPU 完成加速训练。

3.2 性能指标

1)平均绝对误差 MAE[28]

MAE 通过计算显著性预测图与真实图之间的平均绝对误差获得,表示显著性预测图与真实图之间的差别,MAE 越小说明该算法性能越好。其中显著预测图和真实图都需要归一化,具体计算公式如下:

$$MAE=\frac{1}{W\times H}\sum_{x=1}^{w}\sum_{y=1}^{H}\|S(x,y)-G(x,y)\| \tag{11}$$

2)Max F-measure

综合评价指标(F-Measure)[29]用来计算准确率和召回率的加权调和平均值。但是准确率和召回率指标有时会出现互相矛盾的情况,这时就需要综合考虑。最常用的方法就是计算 F-Measure 的值,具体公式如下:

$$F=\frac{(\beta^2+1)Precision*Recall}{\beta^2 Precision+Recall} \tag{12}$$

其中,参数β^2一般取值为 0.3,即增加了 Precision 的权重值,因为通常认为准确率(Precision)更重要。

本次研究中选择最大的 F-measure 作为评估指标,即 Max F-measure,该指标越大代表模型性能越好。

3)Max E-measure

E-measure[30]用来计算图像和局部像素匹配的全局平均值,具体计算公式如下:

$$Q_S=\frac{1}{W\times H}\sum_{i=1}^{w}\sum_{j=1}^{H}\boldsymbol{\varphi}_S(i,j) \tag{13}$$

其中，$\boldsymbol{\emptyset}_s$ 是增强的对齐矩阵，分别反映了显著性预测图和真实值减去全局平均值之后的相关性。Max E-measure 即计算的所有 E-measure 中的最大值，因此该指标值越大，代表模型的性能越好。

4）S-measure[31]

S-measure 用来评估真实值显著性映射与真实值之间的结构相似性，其中 S_o 与 S_r 分别指对象感知和区域感知结构的相似性，具体计算公式如下：

$$S=\propto x\,S_o+(1-\propto)x\,S_r \tag{14}$$

其中，$\propto$ 一般设置为 0.5。

5）检测准确率

Dice 系数是一种集合相似度度量函数，用于计算两个样本之间的相似度，本质上是衡量两个样本之间的重叠部分，取值为[0,1]，当取值为 1 时代表两个样本完全一致。选择合适的 Dice 值，视频帧大于该值即检测到目标，Dice 系数用于计算显著性目标检测准确率，具体计算公式为：

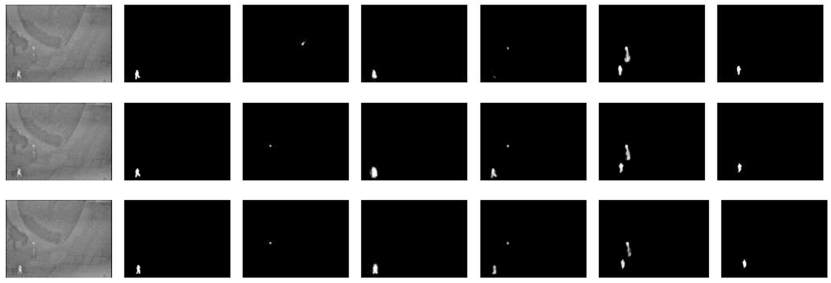
$$DICE=\frac{2|X\cap Y|}{|X|+|Y|} \tag{15}$$

$$\text{检测准确率}=\frac{\text{检测到目标的帧数}}{\text{视频总帧数}} \tag{16}$$

3.3　模型测试结果及分析

1）Dataset 01-OSU Thermal Pedestrian Database 测试结果

为了验证本文模型的有效性，将其与目前已有的针对可见光图像的显著性目标检测算法的测试结果进行对比。如图 5 依次给出了红外数据序列的原始视频帧、真值以及不同模型的测试结果的第 20－22 帧，表 2 列出了不同模型测试结果的定量指标对比。



（a）原图　（b）真值　（c）CPD　（d）MGA　（e）Ds-Net　（f）RcrNet　（g）本研究

图 5　Dataset 01 测试结果

Fig. 5　Dataset 01 test results

表 2　不同算法的评估指标

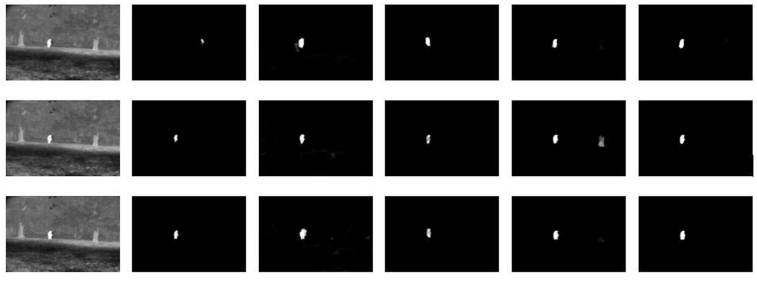
Table 2　Evaluation indicators of different algorithms

模型	MAE	Max F-measure	Max E-measure	S-measure	检测准确率/%
CPD	0.0025	0.3257	0.5450	0.6320	4.11
MGA	0.0028	0.6366	0.8676	0.7551	73.61
Ds-Net	0.0022	0.6588	0.8388	0.7184	75.00
RcrNet	0.0046	0.4715	0.7773	0.6505	45.21
ours	**0.0015**	**0.7466**	**0.9118**	**0.8156**	**89.04**

对不同模型的测试结果进行分析，由于 CPD[32] 模型只利用了红外视频帧的空间特征，忽略了时间信息，因此只能检测出少数视频帧中的显著性目标，在后续红外视频帧中虽然检测到了显著性目标，但是却不能将无关的路灯背景过滤掉。MGA 和 Ds-Net 两个模型都通过光流获取运动信息，两个模型都会出现检测不到目标以及背景抑制效果不好的情况。RcrNet[33] 模型利用了空间和时间特征，能够检测出显著性目标，但也无法滤除无关的路灯背景。本文提出的模型不仅可以将显著性目标准确检测出来（对比不同模型的检测准确率，该模型的检测准确率最高），还可以将无关的背景抑制掉，且实现了最低的 MAE、最高的 Max F-measure，Max E-measure 和 S-measure，验证了本文提出的模型的性能最优。

2）Dataset 05-Terravic Motion IR Database 测试结果

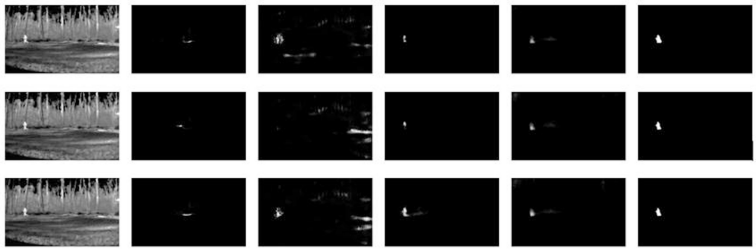
对比两组红外视频序列可知，irw09 相较于 irw10 背景更为杂乱，利用本组实验验证本文模型适用于背景复杂的红外视频序列。为了验证本文模型的有效性，与目前已有的针对可见光图像的显著性目标检测算法的测试结果进行对比，图 6、图 7 为红外视频序列原图以及不同模型的测试结果，分别给出了 irw10 的第 323－325 帧以及 irw09 的第 1468－1470 帧对比图像。



（a）原图　（b）CPD　（c）MGA　（d）Ds-Net　（e）RcrNet　（f）本研究

图 6　irw10 测试结果

Fig. 6　irw10 test results



(a)原图 (b)CPD (c)MGA (d)Ds-Net (e)RcrNet (f)本研究

图 7 irw09 测试结果

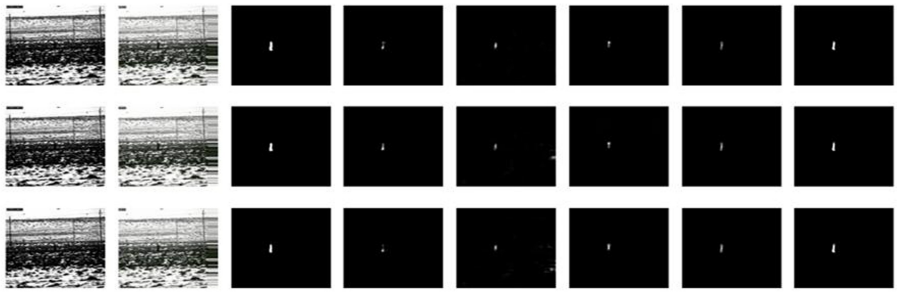
Fig. 7 irw09 test results

由测试结果可以看出,在背景越来越复杂的红外数据集上,CPD 模型很难将目标检测出来;MGA 和 Ds-Net 模型在背景复杂的 irw09 红外视频序列上表现欠佳;RcrNet 模型能够将部分显著性目标检测出来,但检测出的目标不完整;本文模型检测效果最好,能够完整地将显著性目标检测出来,且抑制掉了无关背景。这组实验也证明了本文模型适用于复杂背景的红外视频显著性目标检测。

3.4 真实红外视频序列

为了验证本文提出的模型在复杂背景下也能达到较好的显著性检测结果,除了在公开的红外数据集上进行测试,还在一组背景复杂且目标不明显的真实红外视频数据序列上进行了测试,该红外视频序列共有 245 帧。由于该红外视频序列质量较差,因此需要进行一定的预处理。首先将视频帧进行图像增强,利用 Gamma 变换提高图像的对比度,由于原视频摄像机跟随目标一起移动,因此运动不明显,为了更好地利用视频的运动信息,对视频进行了稳像处理。

图 8 为原图、预处理后的图片、真值以及不同模型的测试结果。取测试结果中的第 222－224 帧进行对比,表 3 列出了不同模型测试结果的评估指标对比。



(a)原图 (b)预处理 (c)真值 (d)CPD (e)MGA (f)Ds-Net (g)RcrNet (h)本研究

图 8 真实红外视频序列测试结果

Fig. 8 Real infrared video sequence test results

表 3 不同算法的评估指标

Table 3 Evaluation indicators of different algorithms

模型	MAE	Max F-measure	Max E-measure	S-measure	检测准确率/%
CPD	0.0024	0.6206	0.7437	0.6635	32.38
MGA	0.0050	0.4599	0.7841	0.6596	25.41
Ds-Net	0.0033	0.5532	0.7062	0.6389	30.34
RcrNet	0.0022	**0.8242**	0.8182	0.7415	46.94
ours	**0.0014**	0.7944	**0.8526**	**0.7926**	**82.86**

对比不同的模型测试结果,CPD 和 RcrNet 模型在测试时都会出现某些帧中未检测出目标的情况,MGA 和 Ds-Net 模型检测到的目标不完整,而本文提出的模型能够准确地检测出到显著性目标。由表 3 可知,本文模型的检测准确率也是最高的,且还可以将杂乱背景抑制掉。对比不同模型测试结果得到的评估指标发现,RcrNet 模型中的 Max F-measure 指标优于本文模型,这是因为 Max F-measure 这一指标与准确率和召回率有关,但这一指标更偏重于准确率(目标是否找对)而忽略了检测到的目标是否找得全,因此仅有一个指标最高并不能证明该模型性能优于其他模型。从整体来看,本文模型整体评估指标是最优的,证明其性能最好,本组实验也证明了本文模型具有鲁棒性和较好的泛化能力。

3.5 消融实验

为了验证空间模块中引入的注意力模块 CBAM 和时间模块中 Non-local block 的有效性,本文分别在红外视频序列 OCTVS-dataset01 以及自己构建的真实红外视频序列上做了两组消融实验,分别表示为消融实验 1 和消融实验 2,具体结果如表 4、表 5 所列。

表 4 列出了在红外数据集的视频序列上进行消融实验的结果。在本次消融实验中,除了本文提出的模块组合不同,其他一切设置都相同,其中空间模块包含 ResNet-50 和 ASPP 模块,时间模块为 DB-ConvGRU 模块,空间模块中引入了注意力机制 CBAM 模块后,MAE 下降了 0.01%,Max F-measure 提升了 4.7%,Max E-measure 提升了 30.59%,S-measure 提升了 9.29%,由此证明了模型中引入的注意力模块 CBAM 的有效性。在时间模块中引入 Non-local block 后,MAE 下降了 0.02%,Max F-measure 提升了 7.36%,Max E-measure 下降了 1.18%,S-measure 提升了 4.59%。整体对比来看,引入了 Non-local block 模块后的评估指标更好,

证明了引入该模块的有效性。本文提出的模型整体评估指标是最好的，这也证明了本文提出的空间特征提取模块和时间特征提取模块的有效性。

表 4　消融实验 1

Table 4　Ablation experiment 1

模块	MAE	Max F-measure	Max E-measure	S-measure
空间模块	0.0017	0.6137	0.6171	0.6861
空间模块＋CBAM	**0.0016**	**0.6607**	**0.9230**	**0.7790**
时间模块	0.0021	0.6162	**0.9224**	0.7429
时间模块加＋Non local block	**0.0019**	**0.6898**	0.9106	**0.7888**
本文模型	0.0015	0.7466	0.9118	0.81656

表 5　消融实验 2

Table 5　Ablation experiment 2

模块	MAE	Max F-measure	Max E-measure	S-measure
空间模块	0.0026	0.2609	0.3029	0.5192
空间模块＋CBAM	**0.0019**	**0.7442**	**0.5937**	**0.6556**
时间模块	**0.0018**	**0.6309**	**0.8990**	**0.7666**
时间模块加＋Non local block	0.0028	0.2828	0.6686	0.5762
本文模型	0.0015	0.7466	0.9118	0.8156

表 5 列出了在真实构建的红外视频序列上进行消融实验的结果。该红外视频序列目标运动不明显且背景较为复杂，由表 5 可知，空间模块中引入了注意力机制 CBAM 模块之后，MAE 下降了 0.07%，Max F-measure 提升了 48%，Max E-measure 提升了 29.08%，S-measure 提升了 13.64%，由此证明了模型中引入的注意力模块 CBAM 的有效性。在时间模块中引入 Non-local block 后，MAE 反而提升了 1%，Max F-measure 下降了 34.81%，Max E-measure 下降了 23.04%，S-measure 下降了 19.04%，可以看出在真实构建的红外视频序列中，引入的时间注意力模块没有发挥出预想的效果，对比数据集中的红外视频序列与本文构建的真实红外视频序列可知，真实的红外视频序列视频质量太差且目标运动不明显，无法提取到太多的运动信息，导致时间模块提取效果不好，此时整体结果基本依靠空间特征模块得到，这也是后续可以改进的一个方向。

综合来看，空间模块中引入的注意力模块以及时间模块中引入的 Non-local block 对模型的性能提升都有所帮助。另外将空间、时间特征模块结合在一起使用后，模型性能也得到了较大提升，说明本文设计的模块都是必要的，减少任何一个模块都会对模型的性能造成影响。

结束语　面对背景越来越复杂的海量红外视频图像，本文提出了一种基于深度学习的红外视频显著性目标检测模型，首先基于 ResNet-50 构建空间特征提取模块，为了使模型更好地聚焦于目标对象，引入了注意力模块 CBAM，利用该模块获取红外视频帧的空间特征信息；然后利用 DB-ConvGRU 和 Non-local block 两个模块构建时间特征模块，用于获取时间特征信息并实现时空一致性，并引入残差连接块，用于提取低层空间特征送入像素级分类器的细化模块中，以避免下采样过程中的信息丢失；最后将时空特征信息和由残差连接块连接 ResNet-50 获得的空间低层特征信息一同送入像素级分类器，生成最终的显著性目标检测结果。在训练整个网络模型时，使用 BCEloss 和 DICEloss 两个损失函数结合的方式，来提高模型训练的稳定性。本文在红外数据集中的多组红外视频序列上进行测试对比，证明了该算法能够在抑制复杂背景的同时准确地将显著性目标检测出来，性能明显优于其他显著性目标检测模型。另外，本文模型在真实构建的红外视频序列上的表现也优于其他模型，证明了所提模型具有鲁棒性和较好的泛化能力。为了验证空间模块中引入的注意力模块 CBAM 和时间模块中 Non-local block 的有效性，本文分别在红外视频序列以及自己构建的真实红外视频序列上做了消融实验，实验结果证明了本文设计的模块都是有效的。本文模型仍然存在不足之处，当目标运动较为缓慢时，模型中的时间模块无法发挥出原有的作用，后续将会对时间模块进行相应的改进。

参 考 文 献

[1] ZHANG B H, JIAO D D, PEI H Q, et al. Infrared moving object detection based on local saliency and sparserepresentation[J]. Infrared Physics & Technology, 2017, 86(12): 187-193.

[2] ZHAO J, FENG C, SHAO F Q, et al. Moving object detection and segmentation based on adaptive frame difference and level set [J]. Information and Control, 2012, 41(2): 153-158.

[3] LEE M, CHO S, LEE S, et al. Unsupervised Video Object Segmentation via Prototype Memory Network[C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023: 5924-5934.

[4] AHN D, KIM S, HONG H, et al. STAR-Transformer: A Spatio-temporal Cross Attention Transformer for Human Action Recognition[C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023: 3330-3339.

[5] ZHOU F, KANG S B, COHEN M F. Time-Mapping Using Space-Time Saliency[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 3358-3365.

[6] HOU X D, ZHANG L Q. Saliency detection: A spectral residual approach[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2007: 1-8.

[7] ACHANTA R, HEMAMI S, ESTRADA F, et al. Frequency-tuned salient region detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2009: 1597-1604.

[8] CHENG M M, MITRA N J, HUANG X L, et al. Global Contrast Based Salient Region Detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 37(3): 569-572.

[9] RAHTU E, HEIKKILA J. A simple and efficient saliency detector for background subtraction [C]// Proceedings Eighth IEEE International Conferenceon Computer Vision(ICCV 2009). IEEE, 2009: 1137-1144.

[10] HAN J H, MA Y, ZHOU B, et al. A robust infrared small target detection algorithm based on human visual system[J]. IEEE Geoscience and Remote Sensing Letters, 2014, 11(12): 2168-2172.

[11] WANG W, SHEN J, SHAO L. Video salient object detection via

fully convolutional networks[J]. IEEE Transactions on Image Processing, 2018, 27(1): 38-49.
[12] SIMONYAN K, ZISSERMAN A. Two-Stream Convolutional Networks for Action Recognition in Videos[C]// Proceedings of the 27th International Conference on Neural Information Processing Systems. 2014: 568-576.
[13] LI H F, CHEN G Q, LI G B, et al. Motion guided attention for video salient object detection[C]// Proceedings Eighth IEEE International Conference on Computer Vision(ICCV 2019). IEEE, 2019: 7273-7282.
[14] FAN D P, WANG W, CHENG M M, et al. Shifting More Attention to Video Salient Object Detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 8546-8556.
[15] LI G B, XIE Y, WEI T H, et al. Flow Guided Recurrent Neural Encoder for Video Salient Object Detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 3243-3252.
[16] HE K M, ZHANG X Y, REN S H, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.
[17] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]// Proceedings of the European Conference on Computer Vision(ECCV). 2018: 3-19.
[18] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.
[19] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7794-7803.
[20] BALLAS N, YAO L, PAL C, et al. Delving deeper into convolutional networks for learning video representations[J]. arXiv: 2016.06432, 2022.
[21] KYUNGHYUN C, BART V, CAGLAR G, et al. Learning phrase representations usingrnn encoder-decoder for statistical machine translation[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2014: 1724-1734.
[22] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
[23] SONG H M, WANG W G, ZHAO S Y, et al. Pyramid dilated deeper convlstm for video salient object detection[C]// Proceedings of the European Conference on Computer Vision(ECCV). 2018: 715-731.
[24] BUADES A, COLL B, MOREL M. A non-local algorithm for image denoising[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2005: 60-65.
[25] HE K M, ZHANG XY, REN S H, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-780.
[26] FEDERICO P, PONT-TUSET J, MCWILLIAMS B, et al. A benchmark dataset and evaluation methodology for video object segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 724-732.
[27] LI J, CHEN X. A benchmark dataset and saliency-guided stacked autoencoders for video based salient object detection [J]. IEEE Transactions on Image Processing, 2018, 27(1): 349-364.
[28] PERAZZI F, KRÄHENBÜHL P, PRITCH Y, et al. Saliency filters: Contrast based filtering for salient region detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2012: 733-740.
[29] MARGOLIN R, ZELNIK-MANOR L, TAL A. How to evaluate foreground maps[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 248-255.
[30] FAN D P, CHENG M M, CHENG G, et al. Enhanced-alignment measure for binary foreground map evaluation[C]// International Joint Conferences on Artificial Intelligence. 2018: 698-704.
[31] FAN D P, CHENG M M, LIU Y, et al. Structure-measure: A new way to evaluate foreground maps[C]// Proceedings Eighth IEEE International Conference on Computer Vision(ICCV 2017). IEEE, 2017: 4558-4567.
[32] WU Z, SU L, HUANG Q. Cascaded partial decoder for fast and accurate salient object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 3902-3911.
[33] YAN P X, LI G B, XIE Y, et al. Semi-supervised video salient object detection using pseudo-labels[C]// Proceedings Eighth IEEE International Conference on Computer Vision (ICCV 2019). IEEE, 2019: 7283-7292.

ZHU Ye, born in 2000, postgraduate. Her main research interests include salient object detection in infrared videos and so on.

HAO Yingguang, born in 1968, associate professor. His main research interests include modeling complex time-varying systems and image processing algorithm.

(责任编辑:何杨)