第 59 卷 第 12 期/2022 年 6 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于特征优选模型的Siamese网络目标跟踪算法

武永强[1]，张宝华[1,3*]，吕晓琪[2,3]，谷宇[1,3]，王月明[1,3]，刘新[1,3]，任彦[1]，李建军[1,3]，张明[1,3]

[1]内蒙古科技大学信息工程学院，内蒙古 包头 014010；

[2]内蒙古工业大学信息工程学院，内蒙古 呼和浩特 010051；

[3]内蒙古模式识别与智能图像处理重点实验室，内蒙古 包头 014010

摘要 针对目标跟踪序列背景复杂、目标大尺度变化等导致目标辨识难度大的问题，提出了基于特征优选模型的Siamese网络目标跟踪算法。首先构建深度网络，有效地提取深度语义信息。再利用沙漏网络对多尺度下的特征图进行全局特征编码，将编码后的特征归一化处理，获取有效目标特征。最后构建特征优选模型，将解码获取的特征作为选择器甄别原特征图的有效特征并增强。为了进一步提高模型的泛化能力，引入注意力机制，对目标特征自适应加权，使其适应场景变化。最终提出算法在OTB100标准跟踪数据集测试成功率达到0.648，预测精度达到0.853，实时性为59.5 frame/s；在VOT2018标准跟踪数据集测试精度为0.536，期望平均覆盖率为0.192，实时性为44.3 frame/s，证明了该算法的有效性

关键词 机器视觉；深度学习；目标跟踪；Siamese网络；特征优选；特征融合

中图分类号 TP391.4 **文献标志码** A **DOI**：10.3788/LOP202259.1215003

Target Tracking Algorithm Based on Siamese Network of Feature Optimization Model

Wu Yongqiang[1], Zhang Baohua[1,3*], Lv Xiaoqi[2,3], Gu Yu[1,3], Wang Yueming[1,3], Liu Xin[1,3], Ren Yan[1], Li Jianjun[1,3], Zhang Ming[1,3]

[1]*School of Information Engineering, Inner Mongolia University of Science and Technology, Baotou* 014010, *Inner Mongolia, China*;

[2]*School of Information Engineering, Mongolia Industrial University, Huhehaote* 010051, *Inner Mongolia, China*;

[3]*Inner Mongolia Key Laboratory of Patten Recognition and Intelligent Image Processing, Baotou* 014010, *Inner Mongolia, China*

Abstract In the target tracking sequences, it is difficult to identify the target because of the complex background and large-scale changes of the target. To solve this problem, a target tracking algorithm based on feature optimization model in the Siamese network is proposed. First, the deep network is constructed to extract the deep semantic information effectively. Then, the hourglass network is used to encode the global features of the multi-scale feature map, and the encoded features are normalized to obtain the effective target features. Finally, a feature optimization model is constructed, and the features obtained by decoding are used as selectors to identify and enhance the effective features of the original feature map. In order to further improve the generalization ability of the model, the attention mechanism is introduced to adaptively weigh the target features to adapt to the scene changes. The proposed algorithm is tested on two standard tracking data sets including OTB100 and VOT2018. The success rate in the OTB100 is 0.648, the prediction accuracy is 0.853, and the real-time performance is 59.5 frame/s; the test

收稿日期：2021-04-25；**修回日期**：2021-05-18；**录用日期**：2021-06-02

通信作者：*zbh_wj2004@imust.cn

accuracy in the VOT2018 is 0.536, the expected average coverage rate is 0.192, and the real-time performance is 44.3 frame/s. The test results prove the effectiveness of the proposed algorithm.

Key words machine vision; deep learning; target tracking; Siamese network; feature optimization; feature fusion

1 引 言

目标跟踪是计算机视觉的重要分支，受到广泛的关注。近年来涌现出众多性能优越的目标跟踪器。但是，目标跟踪算法受到多种因素干扰，例如：目标快速移动、帧图像模糊、光照变化、目标遮挡及大尺度变化、背景杂波等[1-2]。

目前，基于Siamese网络的目标跟踪算法凭借精度高、速度快逐渐成为主流算法。用于目标跟踪的全卷积连体网络(SiamFC)[3]端到端的训练网络实现了高速跟踪，推动了Siamese网络的实时化应用。近年来特征融合、注意力机制被证实应用于目标跟踪是有效的，随之出现了大量探索特征融合以及结合注意力机制的Siamese网络目标跟踪器[4-6]。学习注意：用于高性能在线视觉跟踪的残差注意连体网络(RASNet)[4]、高效的视觉跟踪与堆叠通道-空间注意力机制(SCSAtt)[5]算法通过引入注意力机制来探索目标特征，其中RASNet融合3个注意力模块，通过简单的相加和相乘完成多注意力机制特征加权提高模型的精度；而SCSAtt通过设计线性堆叠的注意力机制模块来提高模型精度。用于实时目标跟踪的双连体网络(SA-Siam)[6]提出训练一个双重Siamses网络分别为外观分支，及嵌入注意力机制的语义分支，两个分支互补提高跟踪精度。

虽然上述Siamese网络目标跟踪算法都有其自身的优越性，但仍存在一定缺陷：

1) 骨干特征提取网络选用浅层网络，并不能有效获取图像深度语义信息，导致模型缺乏对目标特征的表征能力。

2) 忽视了不同尺度下目标特征存在差异的潜在因素，所提取的目标特征区域不充分，致使跟踪器在背景复杂、目标大尺度变化等情况下，无法有效甄别目标信息。

3) 通过简单的加权来融合不同模块所提取的特征优势是存在缺陷的，这在一定程度上影响了跟踪精度。

针对上述问题，本文构建深度网络加强对目标的语义信息提取，同时在模板分支上引入层数较浅的沙漏网络[7]在不同尺度的特征图上捕获目标特征，达到充分利用模板图像中目标信息的目的。为将沙漏网络提取的目标特征有效补充到原特征图上，设计特征优选模型弥补简单加权融合的缺陷。最后，引入线性堆叠的注意力机制进一步提炼目标信息，采用跳跃连接把特征优选模型输出的特征与注意力机制提取的特征进行融合，提高模型对目标信息的表征能力。

2 算法原理

2.1 Siamese网络目标跟踪算法

Siamese网络一般由两个完全相同的分支构成对称结构，用于便捷共享权重来学习目标的特征。SiamFC搭建了两个参数和形貌完全一样的全卷积网络。通过提取首帧目标特征和后续帧候选区域的特征，然后计算相似度，得分最高的位置表示预测目标的位置。相似度计算公式为

$$f(x,z)=\varphi(x)*\varphi(z)+b, \tag{1}$$

式中：$\varphi(\cdot)$表示卷积操作；$*$表示互相关计算；b表示偏置($b\in\mathbb{R}^{n\times n}$，$\mathbb{R}$为响应图每个位置上取值的实数集)；$z$表示模板分支图像(一般为第一帧图像的目标)；$x$表示搜索分支图像；$f(\cdot,\cdot)$表示模板分支和搜索分支的相似性计算结果。

本文不同于SiamFC，模板分支和搜索分支有显著差异。搜索分支只通过骨干网络提取特征；在模板分支通过骨干网络特征提取之后，再经过沙漏网络以及后续级联的通道注意力机制[8]和空间注意力机制[8]。这样搭建模型是考虑到Siamese网络目标跟踪算法以相似性学习为基础，计算首帧目标模板与后续帧的信息相似度确定目标位置，且训练期间不更新模板。故在模板分支引入特征优选、注意力机制等模块可有效利用首帧图像提取目标模板特征，有助于提高跟踪精度。搜索分支不再引入除骨干网络外的特征提取模块，可以减少计算开销、增加跟踪速度。模型框架图如图1所示，图中：※表示特征优选；×表示相乘；+表示相加；X_{corr}表示互相关。

由图1可知，模板分支的计算可以表示为

$$\dot{\varphi}(z)=\psi\left[\varphi(z)\right], \tag{2}$$

式中：$\varphi(z)$表示对模板分支的图像特征提取；$\psi[\cdot]$表示沙漏网络和注意力机制提取的特征；$\dot{\varphi}(\cdot)$表示

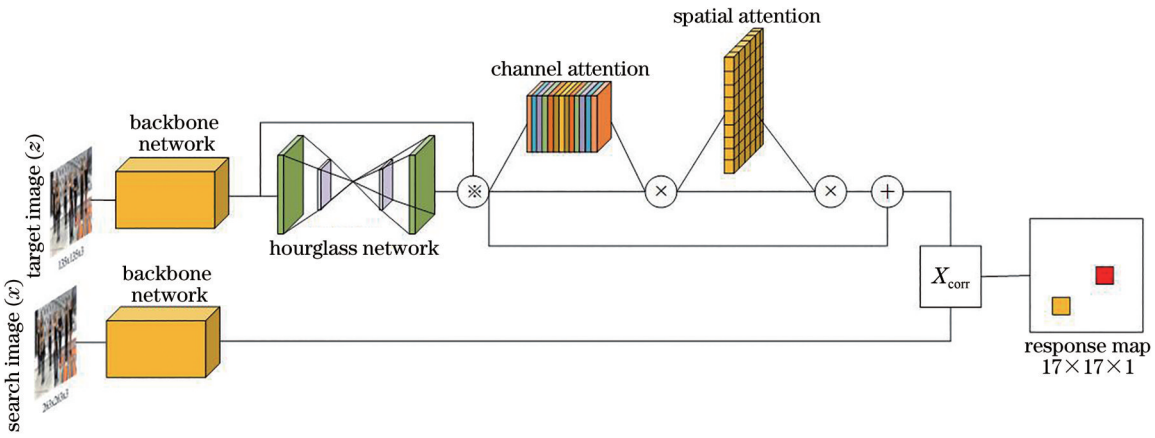

图 1　模型框架图

Fig. 1　Model frame diagram

模板分支最终的特征。

而搜索分支的计算同 SiamFC 一样，最终本文的相似计算公式为

$$f(x,z)=\varphi(x)*\dot{\varphi}(z)+b\ , \tag{3}$$

式中：$\varphi(\cdot)$表示卷积操作；$*$表示互相关操作；b表示偏置$\left(b\in\mathbb{R}^{n\times n}\right)$；$z$表示模板分支图像（一般为第一帧图像）；$x$表示搜索分支图像；$f(\cdot,\cdot)$表示模板分支和搜索分支的相似性响应图。

2.2　骨干网络

考虑到采用浅层网络作为骨干网络，并不能有效获取图像的深度语义信息，故参考 VGG16-Net 网络[9]，基于改进 SiamFC 的实时目标跟踪算法[10]，并结合注意力机制的 Siamese 网络目标跟踪算法[11]的网络结构。设计了由 5 大层结构组成的网络作为骨干网络。网络结构如表 1 所示。

表 1 给出了骨干网络的具体参数，例如：卷积层、最大池化层所在位置、步幅大小、卷积核大小、通道数变化情况、模板和搜索图像的尺寸大小。此外，除 Layer5 外的每一个卷积层后使用 BatchNorm2d 归一化处理之后，再进行 Relu 激活函数完成非线性激活。为了消除 Padding 带来的目标漂移影响，本文没有使用 Padding 去填充卷积层。

表 1　骨干网络结构

Table 1　Structure of backbone network

Layer number	Network structure	Convolution kernels	Stride	Channel number	Template image /pixel	Search image / pixel
Layer1	Input layer			–3	135×135	263×263
	Conv2d	3	1	96–3	133×133	261×261
	Conv2d	3	1	96–96	131×131	259×259
	MaxPool2d	3	2	–	65×65	129×129
Layer2	Conv2d	3	1	128–96	63×63	127×127
	Conv2d	3	1	128–128	61×61	125×125
	MaxPool2d	3	2	–	30×30	62×62
Layer3	Conv2d	3	1	256–128	28×28	60×60
	Conv2d	3	1	256–256	26×26	58×58
	Conv2d	3	1	256–256	24×24	56×56
	MaxPool2d	2	2	–	12×12	28×28
Layer4	Conv2d	3	1	512–256	10×10	26×26
Layer5	Conv2d	3	1	512–512	8×8	24×24

2.3　沙漏网络及特征优选模型

沙漏网络被广泛使用在人体姿态估计中，其基本思想为先进行下采样，再进行上采样，期间引入残差模块[12]。沙漏网络的设计之初就是为获取不

同尺度下图像中所涵盖的信息，下采样可以快速编码得到特征图上的全局信息，上采样可以提取前者的全局高维特征。本文所构建的沙漏网络及特征优选模型的框架图，如图 2 所示，图中：S 表示 Sigmoid 激活函数；+1 表示加 1；×表示相乘。

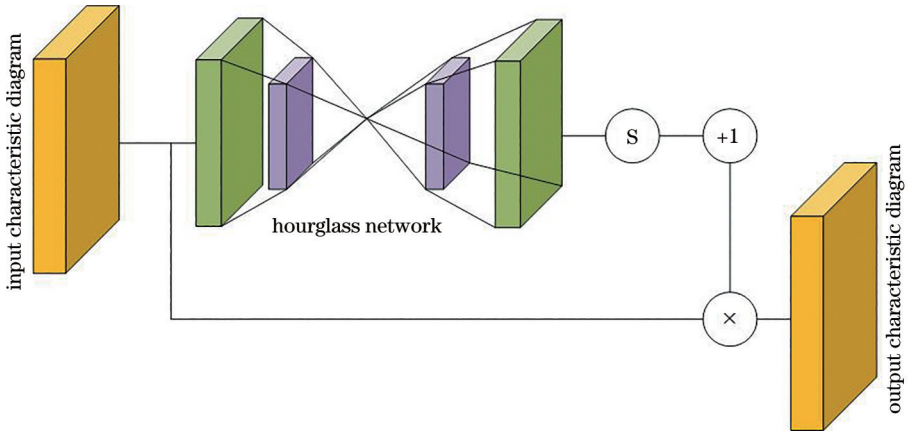

图 2　沙漏网络及特征优选模型

Fig. 2　Hourglass network and feature optimization model

由图 2 可知，该模型特征提取部分由卷积核为 3 和 2 的 Conv2d，以及卷积核分别为 3 和 2 的 ConvTranspose2d 层构成，这样设计保证了特征图经过沙漏网络尺寸不变。为了对原特征图的目标特征进行有效增强，需要使用 Sigmoid 作为激活函数对沙漏网络提取的不同尺度下的目标特征进行归一化处理，但会削弱其输出响应，若直接与原特征图融合会导致目标特征响应值变得很小，故本文参考残差注意力机制[13]，基于目标感知特征筛选的 Siamese 网络跟踪算法[14]，设计特征优选模型，将归一化的输出响应放大，再作为选择器对原特征图的目标特征进行增强。用数学公式表示为

$$\vartheta(v)=\varphi(z)\times\left\{S_{\text{igmoid}}\left\{\rho\left[\varphi(z)\right]\right\}+1\right\},\quad(4)$$

式中：$\varphi(z)$ 表示骨干网络提取的特征；$\rho[\cdot]$ 表示沙漏网络提取特征；$S_{\text{igmoid}}\{\cdot\}$ 表示 Sigmoid 激活函数；$\vartheta(\cdot)$ 表示特征优选模型的输出。

2.4　注意力机制

人类的视觉机制会重点关注场景特定目标，而忽略场景的背景，凭经验放大部分信息来理解场景。受人眼视觉特性的启发，注意力机制通过优先摄取重要物体部位，来提高模型的跟踪精度。

SCSAtt 提出在 Siamese 网络的模板分支堆叠通道-空间注意力机制，这种做法有效地线性结合了两种注意力机制，增强了跟踪模型的判别能力以及表征能力。

通道注意力模块面向不同的卷积核所得到的特征通道也是不同的，每个特征通道都有特定的视觉感受，其表征也不是同一对象，因此增强跟踪目标的特征通道响应是必要的。全局平均池化可以关注到全局特征，全局最大池化注重捕获细节特征，故使用全局平均池化和全局最大池化联合构建通道注意力机制。具体模块构造细节可参考 SCSAtt[5]，通道注意力机制模块如图 3 所示。

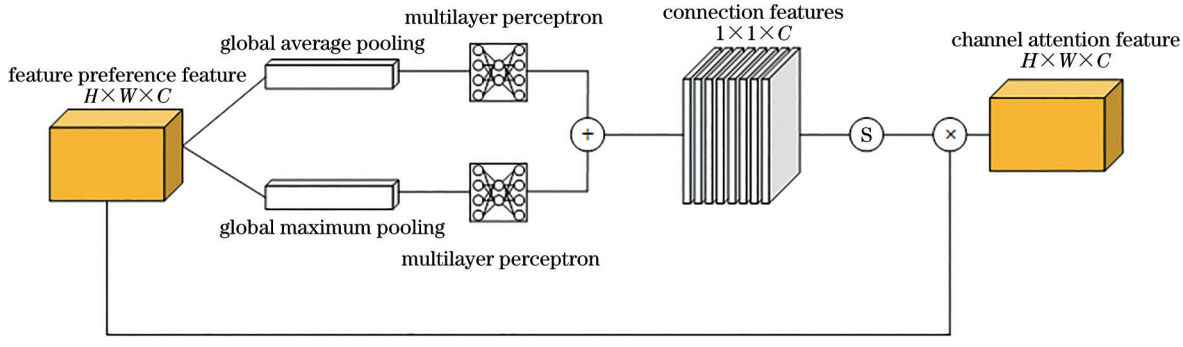

图 3　通道注意力模块[5]

Fig. 3　Channel attention module[5]

通道注意力具体计算过程如下：

$$f_{\max}^{1\times1\times C}=C_{onv2}\left\{R_{elu}\left\{C_{onv1}\left[G_{\max}\left(\vartheta_M^{H\times W\times C}\right)\right]\right\}\right\},\quad(5)$$

$$f_{avg}^{1\times1\times C}=C_{onv2}\left\{R_{elu}\left\{C_{onv1}\left[G_{avg}\left(\vartheta_M^{H\times W\times C}\right)\right]\right\}\right\},\quad(6)$$

式中：$\vartheta_M^{H\times W\times C}$ 表示式(4)所提取的特征图像；C_{onv1}，C_{onv2} 表示卷积，不同的是两者输入输出通道数不同；$f_{avg}^{1\times1\times C}$、$f_{\max}^{1\times1\times C}$ 分别表示全局平均池化和最大池化的输出特征，将其二者融合得

$$\eta^{1\times1\times C}=S_{igmoid}\left(f_{\max}^{1\times1\times C}+f_{avg}^{1\times1\times C}\right),\quad(7)$$

式中：Sigmoid(•)表示 Sigmoid 激活函数；$\eta^{1\times1\times C}$ 表示连接之后的特征，最终输出表示为

$$C_A^{H\times W\times C}=\eta^{1\times1\times C}\times\vartheta_M^{H\times W\times C},\quad(8)$$

式中，$C_A^{H\times W\times C}$ 表示通道注意力机制的输出。

不同于通道注意力机制，空间注意力机制关注图像中目标的位置信息特征，通过把特征映射上的最大池化和最小池化融入信道之中，增强通道特征的位置，补全了通道注意力机制缺失的位置特征。空间注意力机制模块如图 4 所示。

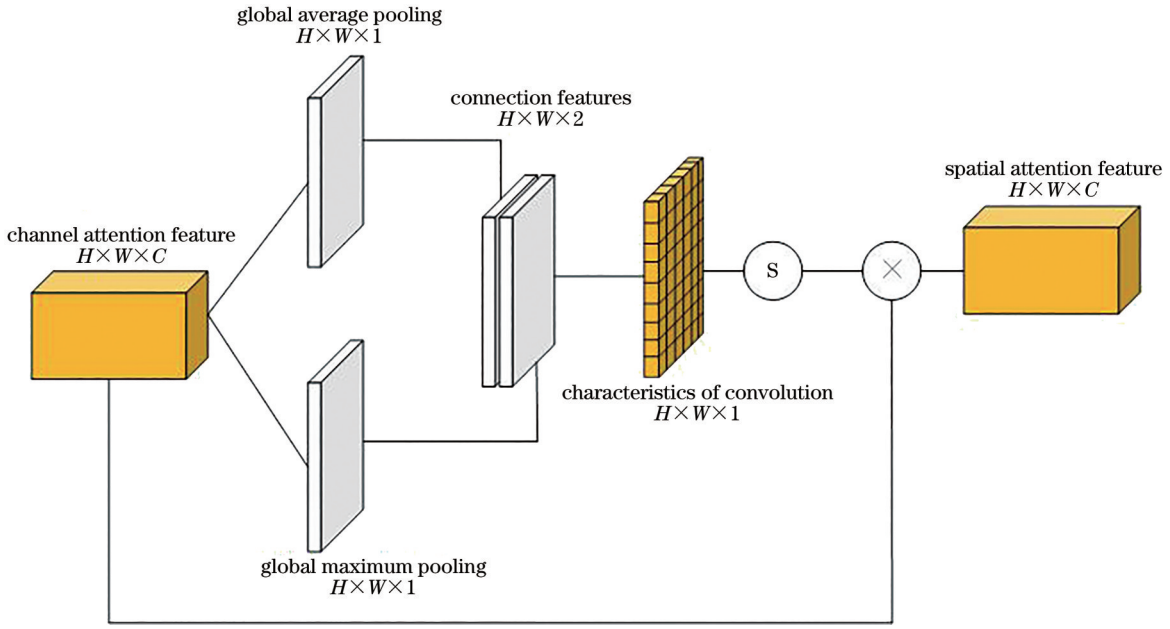

图 4　空间注意力模块[5]

Fig. 4　Spatial attention module[5]

空间注意力机制计算过程如下：

$$S_{\max}^{H\times W\times1}=G_{\max}\left(C_A^{H\times W\times C}\right),\quad(9)$$

$$S_{avg}^{H\times W\times1}=G_{avg}\left(C_A^{H\times W\times C}\right)。\quad(10)$$

连接空间全局最大池化和全局平均池化过程如下所示：

$$\xi_s^{H\times W\times1}=S_{igmoid}\left\{C_{onv2}\left[C_{oncat}\left(S_{\max}^{H\times W\times1},S_{avg}^{H\times W\times1}\right)\right]\right\},\quad(11)$$

式中：C_{onv2} 表示填充为 1、步幅为 1、核为 3 的卷积；$C_{oncat}(•,•)$ 表示连接最大池化和平均池化；$\xi_s^{H\times W\times1}$ 表示经过上述处理的特征输出，最终空间注意力模块输出为

$$S_A^{H\times W\times C}=\xi_S^{H\times W\times1}\times C_A^{H\times W\times C},\quad(12)$$

式中，$S_A^{H\times W\times1}$ 表示空间注意力特征。

3　实验结果与分析

实验环境为 Linux(Ubuntu16.04)系统，使用 python 语言在 PyTorch 框架编写程序。实施细节如下：

1) 训练阶段：为了更好地训练模型，从数据集序列里随机抽取训练图像。参数设置如下：批尺寸(batch-size)为 32；动量为 0.9；开始学习率为 0.01；最终学习率为 10^{-5}。训练损失函数，参考结合缓冲区与三元组损失的 Siamese 网络目标跟踪[15]等，本文逻辑损失函数如下：

$$\iota\left[f(z,x),g\right]=\frac{1}{|M|}\sum_{m\in M}\log\left\{1+\exp\left[-f(z,x)[m]\cdot g[m]\right]\right\},\quad(13)$$

式中：M 表示响应图上待跟踪位置集合；$f(z,x)[m]$ 表示模板分支 z 与搜索分支 x 的相似性得分；$g[m]$ 表示真实标签范围为 $\{-1,+1\}$。

为确保外观变化的鲁棒性，训练过程中采用动量随机梯度下降法(SGD)对损失函数进行优化，用公式表示为

$$\mathrm{Arg}_{\theta}\frac{1}{P}\sum_{i=1}^{P}\iota\left[f\left(z_i,x_i\right),g_i\right],\tag{14}$$

式中，P表示样本。

损失函数训练过程图如图5所示。

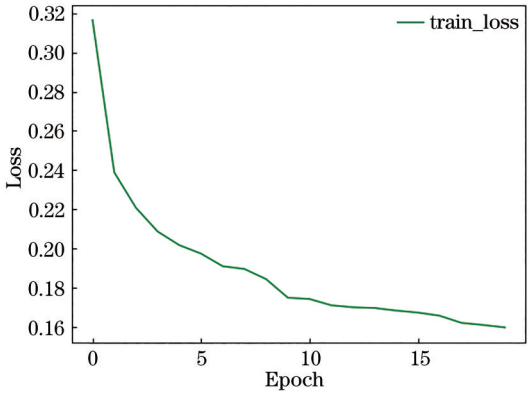

图5　损失函数训练过程图

Fig. 5　Training process diagram of loss function

2）测试阶段：视频序列中首帧图像为模板图像；后续帧为搜索图像。为了更精确地估计目标位置，使用双三次插值预测目标位置，本文参考SCSAtt[5]、SiamFC[3]，采用同样的方法，设置多个搜索比例因子应对序列中目标比例变化问题，比例因子具体值为$1.0375^{\{-1,0,+1\}}$。同时，为了避免模型过度相信比例因子，造成精度损失，再设置惩罚项(0.9745)与之相乘对其进行约束，以减小误差。

3.1　训练及测试数据集

本文使用GOT-10k[16]和ILSVR2015_VID[17]作为训练数据集，其中，GOT-10k涵盖了563个类别，超过10000个视频序列，多达87种运动模式，150多万个人工手动标记的物体真实边界框。GOT-10k包含了5大类别：人物、人造物体、自然物体、动物、其他类别。

使用OTB数据集和VOT数据集作为测试数据集。OTB数据集的评判标准为准确率(precision rate)和成功率(success rate)。准确率表示为真实中心位置与预测中心位置之间的欧氏距离，计算过程如下：

$$\varepsilon=\sqrt{\left(x_{\mathrm{a}}-x_{\mathrm{b}}\right)^2+\left(y_{\mathrm{a}}-y_{\mathrm{b}}\right)^2},\tag{15}$$

式中：$\left(x_{\mathrm{a}},y_{\mathrm{a}}\right)$为真实的中心距离；$\left(x_{\mathrm{b}},y_{\mathrm{b}}\right)$为预测的中心距离。

成功率表示真实标注框面积和预测框面积的重合程度，计算公式为

$$S=\frac{S_{\mathrm{t}}\cap S_{\mathrm{gt}}}{S_{\mathrm{t}}\cup S_{\mathrm{gt}}},\tag{16}$$

式中：S_{t}表示预测的目标框面积；S_{gt}表示真实标注的目标框面积。

VOT数据集采用Accuracy和预期平均重叠(EAO)对跟踪算法进行评估。Accuracy为单个视频下预测跟踪框与真实标注框交并比大小。EAO为将跟踪成功的视频拆分出来，计算短期序列上重叠曲线的平均值。EAO计算公式为

$$\rho_{\mathrm{o}}=\frac{1}{N}\sum_{t=1}^{N}\nu_t,\tag{17}$$

式中：N表示跟踪有效的帧数；ν_t表示t帧的跟踪准确度。

3.2　实验结果定量分析

跟踪器的实验平台为linux(Ubuntu16.04)系统，配备NVIDIA RTX2080Ti的计算机截取训练的前20个轮次，与SCSAtt相比减少了训练时间，并且成功率略高于SCSAtt。具体数据结果如表2所示。

同时还对比了包括提出算法在内的8种跟踪算法，提出算法的跟踪成功率及准确率表现较优，如图6所示。

从图6(b)成功率分析可以得出，提出算法的成功率比SiamFC提高了0.061，比SiamRPN提高了

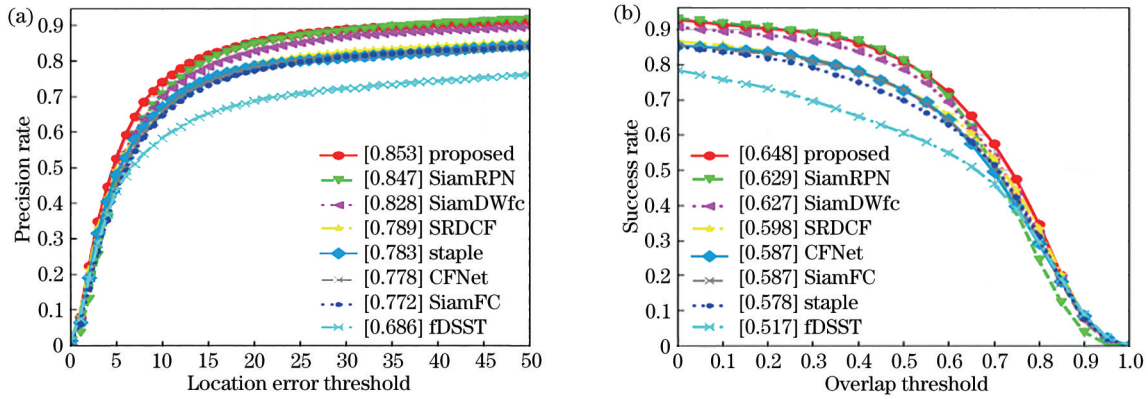

图6　各种算法在OTB100数据集的测试结果对比图。(a)准确率；(b)成功率

Fig. 6　Comparison of test results of various algorithms in OTB100 data set. (a) Precision rate; (b) success rate

表 2　对比实验数据

Table 2　Comparison of experimental data

Name	Epoch	Got-10k	ILSVR2015_VID	Precision rate	Success rate	Speed/(frame·s^{-1})
SCSAtt	50	√	√	0.855	0.641	59.871
Proposed	20	√	√	0.853	0.648	59.497

0.019，比 CFNet 提高了 0.061。从图 6(a)可看出，提出算法也位于前列。此外，提出算法还在 VOT2018 数据集与其他 5 种算法进行了对比，具体数据如表 3 所示。

表 3　各种算法在 VOT2018 数据集的测试结果对比

Table 3　Comparison of test results of various algorithms in VOT2018 data set

Name	Accuracy	EAO	Speed /(frame·s^{-1})
Proposed	0.5360	0.1920	44.33
SiamFC	0.4943	0.1875	31.89
LSART	0.4932	0.3230	1.00
CSRDCF	0.4910	0.2562	10.20
DeepSRDCF	0.4896	0.2753	65.30
ECO-HC	0.4842	0.2486	75.60

6 种算法中，提出算法精确率表现最优，其中比 SiamFC 精确率提高了 0.0417；而且使用了比 SiamFC 更深的网络，实时性没有降反而提高了，体现了提出算法的综合性能较强。

3.3　实验结果定性分析

OTB 数据集中对跟踪挑战分别进行了标注，便于对算法的性能进行更加全面的评估及分析。其中标注包括：尺度变化(SV)、平面内旋转(IPR)、背景杂波(BC)、形变(DEF)、快速移动(FM)、遮挡(OCC)、光照变化(IV)、低分辨率(LR)、快速移动(FM)、完全遮挡(OV)、运动模糊(MB)。对 SCSAtt 以外的 8 个算法就上述难点进行了实验，提出算法的实验结果多数位于前列，体现了提出算法的稳定性。具体结果如表 4 所示(其中 Suc 表示成功率，Pre 表示准确率)。

可视化 9 个算法在 OTB100 数据集跟踪结果如图 7～9 所示。

表 4　各种算法在 OTB100 数据集的挑战表现结果

Table 4　Challenge performance results of various algorithms in OTB100 data set

Name		IPR	IV	BC	OCC	DEF	SV	LR	FM	OPR	OV	MB
Proposed	Suc	0.624	0.646	0.609	0.613	0.609	0.636	0.682	0.616	0.630	0.545	0.628
	Pre	0.842	0.844	0.808	0.807	0.831	0.846	0.998	0.797	0.854	0.715	0.800
Siam RPN	Suc	0.628	0.649	0.591	0.585	0.617	0.615	0.639	0.599	0.625	0.542	0.622
	Pre	0.854	0.859	0.799	0.780	0.825	0.838	0.978	0.789	0.851	0.726	0.816
Siam DWfc	Suc	0.606	0.622	0.574	0.601	0.560	0.613	0.596	0.630	0.612	0.590	0.654
	Pre	0.824	0.794	0.762	0.798	0.763	0.819	0.901	0.808	0.829	0.781	0.841
CFNet	Suc	0.567	0.541	0.561	0.527	0.526	0.546	0.614	0.554	0.553	0.454	0.540
	Pre	0.786	0.707	0.756	0.699	0.714	0.731	0.888	0.705	0.759	0.601	0.680
Siam FC	Suc	0.559	0.572	0.527	0.549	0.512	0.556	0.618	0.571	0.561	0.509	0.554
	Pre	0.743	0.736	0.692	0.723	0.691	0.736	0.900	0.744	0.758	0.673	0.707
Staple	Suc	0.548	0.529	0.560	0.543	0.551	0.521	0.394	0.540	0.533	0.475	0.541
	Pre	0.768	0.783	0.749	0.726	0.752	0.726	0.690	0.708	0.737	0.664	0.698
SRDCF	Suc	0.544	0.613	0.583	0.559	0.544	0.561	0.514	0.597	0.550	0.460	0.594
	Pre	0.745	0.792	0.775	0.734	0.734	0.745	0.760	0.768	0.741	0.594	0.765
fDSST	Suc	0.505	0.559	0.523	0.460	0.427	0.475	0.382	0.458	0.477	0.386	0.469
	Pre	0.698	0.722	0.704	0.602	0.550	0.648	0.678	0.570	0.654	0.474	0.566

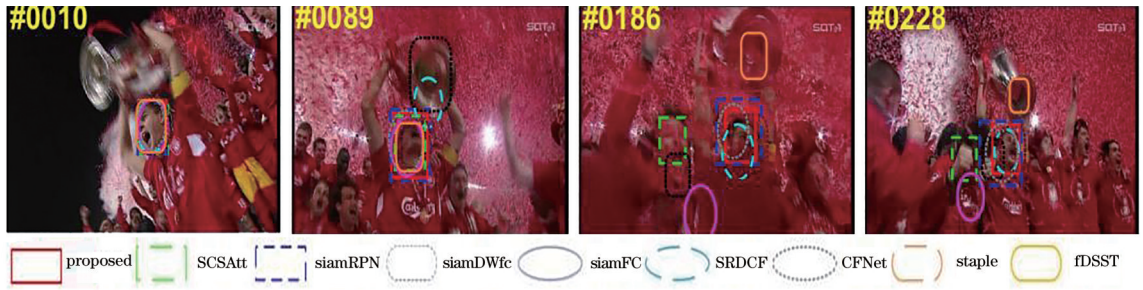

图 7　各种算法在 Soccer 序列上的测试结果

Fig. 7　Test results of various algorithms on Soccer sequence

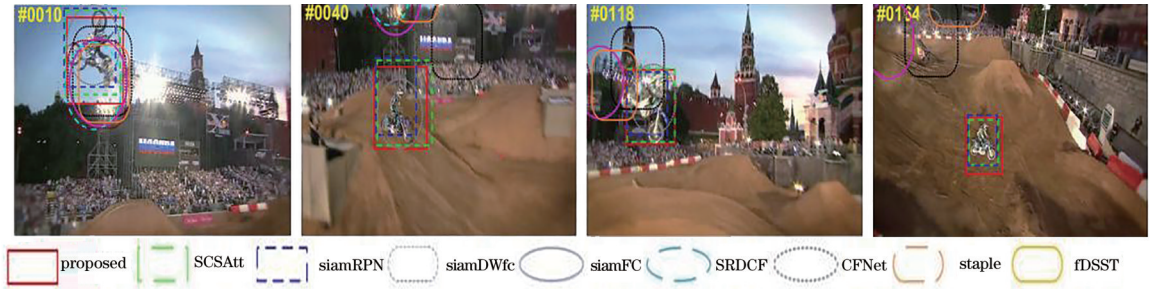

图 8　各种算法在 MotorRolling 序列上的测试结果

Fig. 8　Test results of various algorithms on MotorRolling sequence

图 9　各种算法在 Jogging 序列上的测试结果

Fig. 9　Test results of various algorithms on Jogging sequence

如图 7 所示，跟踪序列从 89 帧到 228 帧开始背景环境逐渐变得复杂，期间伴随着物体运动、相似物体干扰等因素，除了提出算法和 SiamRPN 外，其他算法都出现了跟踪漂移、丢失等情况，直到 228 帧 CFNet、SRDCF 才消除跟踪漂移，重新准确定位到目标。

如图 8 所示，山地摩托参赛者由 10 帧开始到 40 帧结束，完成翻转动作。翻转过程中目标存在高速移动和目标形变。除了提出算法、SCSAtt、CFNet，以及 SiamRPN 外，其他算法都丢失了目标。118 帧目标再次翻转之后在坡道快速俯冲，提出算法仍可准确地对目标进行跟踪。

如图 9 所示，行人从 48 帧被遮挡到 62 帧再次出现，SiamRPN、fDSST、SRDCF 出现跟踪漂移，到 73 帧行人完全脱离遮挡时，fDSST、SRDCF 已经丢失了目标。而提出算法始终精准锁定目标位置。

综上所述，提出算法能够应对目标遮挡、复杂背景、目标形变等复杂场景下的目标跟踪，且对复杂背景、目标旋转表现出更强的目标建模能力。

3.4　消融实验

为了证明本文构建的深度网络以及通过特征优选模型改进沙漏网络提高目标跟踪精度的有效性，以及探索不同深度沙漏网络的效果，故做此消融实验。实验平台在 linux(Ubuntu16.04)系统，在配备 Teas V100 的 GPU 对 GOT-10k 的第一个 split 训练 10 个轮次，使用 OTB100 作为测试集。结果如表 5～7 所示。

表 6 为本文构建的深度网络与浅层网络在 OTB100 标准数据集中目标旋转、复杂背景等场景下的实验结果，结果显示本文构建的深度网络在所有场景下结果都优于浅层网络。

表 5　深度网络消融实验总体数据

Table 5　Overall data of deep network ablation experiment

Name	Improved VGG-Net	Improved Hourglass	AlexNet	Precision rate	Success rate
Proposed-2	√	√	—	0.560	0.724
Proposed-A	—	√	√	0.538	0.703

表 6　深度网络在 OTB100 挑战消融实验数据

Table 6　Experimental data of OTB100 challenge ablation in deep network

Name		IPR	IV	BC	OCC	DEF	SV	LR	FM	OPR	OV	MB
Proposed-2	Suc	0.522	0.510	0.479	0.501	0.490	0.543	0.604	0.559	0.529	0.439	0.544
	Pre	0.677	0.642	0.619	0.643	0.647	0.709	0.872	0.695	0.700	0.569	0.664
Proposed-A	Suc	0.511	0.480	0.458	0.487	0.462	0.520	0.554	0.536	0.522	0.431	0.532
	Pre	0.659	0.608	0.614	0.624	0.622	0.686	0.812	0.673	0.691	0.571	0.663

通过表 7 的实验结果证明，引入沙漏网络增强目标特征确实可以提高目标跟踪的精度。同时发现加入两层沙漏网络效果最佳。具体结果如图 10 所示。

表 7　消融实验数据

Table 7　Ablation experiment data

Name	Improved VGG-Net	Improved Hourglass	Hourglass	Layer	Precision rate	Success rate
SCSAtt	√	—	—	—	0.687	0.529
Proposed-1	√	√	—	1	0.698	0.538
Proposed-2	√	√	—	2	0.724	0.560
Proposed-3	√	√	—	3	0.704	0.539
Proposed-No	√	—	√	2	0.694	0.530

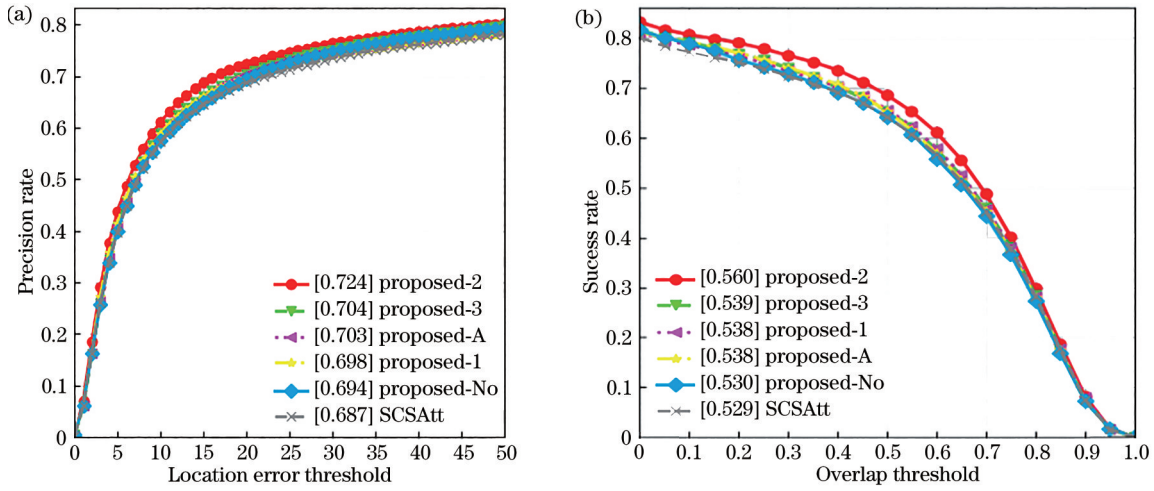

图 10　消融实验结果。(a)准确率；(b)成功率

Fig. 10　Results of ablation experiment. (a) Precision rate; (b) Success rate

4　结　　论

为实现通过增强目标特征信息、提高模型跟踪精度的目的，提出了基于特征优选模型的 Siamese 网络跟踪算法。该算法在 SCSAtt 以及 SiamFC 的基础上以改进 VGG16-Net 作为特征提取网络；模板分支结合特征优选沙漏网络，以及线性堆叠的注意力机制之后，与搜索分支计算得分响应图。使用 GOT-10k、ILSVRC2015_VID 数据集进行端到端的训练，最终以 OTB100、VOT2018 数据集作为测试数据集，提出算法在跟踪精度上取得了较优的结果，并且时效性远高于 24 frame/s 的实时性要求。下一步工作将对目标跟踪特征融合进行更深度挖掘，同时致力于解决相似物体干扰，进一步提高算

法的跟踪性能，以及模型的泛化性、鲁棒性。

参 考 文 献

[1] Wu Y, Lim J, Yang M H. Online object tracking: a benchmark[C]//2013 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2013, Portland, OR, USA. New York: IEEE Press, 2013: 2411-2418.

[2] Zhang K H, Zhang L, Yang M H. Fast compressive tracking[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(10): 2002-2015.

[3] Bertinetto L, Valmadre J, Henriques J F, et al. Fully-convolutional Siamese networks for object tracking[M]//Hua G, Jégou H. Computer vision-ECCV 2016 workshops. Lecture notes in computer science. Cham: Springer, 2016, 9914: 850-865.

[4] Li B, Yan J J, Wu W, et al. High performance visual tracking with Siamese region proposal network [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 8971-8980.

[5] Rahman M M, Fiaz M, Jung S K. Efficient visual tracking with stacked channel-spatial attention learning[J]. IEEE Access, 2020, 8: 100857-100869.

[6] He A F, Luo C, Tian X M, et al. A twofold Siamese network for real-time object tracking[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 4834-4843.

[7] Newell A, Yang K Y, Deng J. Stacked hourglass networks for human pose estimation[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9912: 483-499.

[8] Li Y, Liu Y, Cui W G, et al. Epileptic seizure detection in EEG signals using a unified temporal-spectral squeeze-and-excitation network[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering: A Publication of the IEEE Engineering in Medicine and Biology Society, 2020, 28(4): 782-794.

[9] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-04) [2021-02-04]. https://arxiv. org/abs/1409.1556.

[10] 张红颖, 贺鹏艺, 王汇三. 基于改进SiamFC的实时目标跟踪算法[J]. 激光与光电子学进展, 2021, 58(6): 0615003.
Zhang H Y, He P Y, Wang H S. A real-time target-tracking algorithm based on improved SiamFC[J]. Laser & Optoelectronics Progress, 2021, 58(6): 0615003.

[11] 张丹璐. 结合注意力机制的孪生网络目标跟踪算法研究[D]. 北京: 北京建筑大学, 2020: 24-32.
Zhang D L. Siamese network combined with attention mechanism for object tracking[D]. Beijing: Beijing University of Civil Engineering and Architecture, 2020: 24-32.

[12] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[13] Wang F, Jiang M Q, Qian C, et al. Residual attention network for image classification[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6450-6458.

[14] 陈志旺, 张忠新, 宋娟, 等. 基于目标感知特征筛选的孪生网络跟踪算法[J]. 光学学报, 2020, 40(9): 0915003.
Chen Z W, Zhang Z X, Song J, et al. Tracking algorithm for Siamese network based on target-aware feature selection[J]. Acta Optica Sinica, 2020, 40(9): 0915003.

[15] 李勇, 杨德东, 韩亚君, 等. 融合扰动感知模型的孪生神经网络目标跟踪[J]. 光学学报, 2020, 40(4): 0415002.
Li Y, Yang D D, Han Y J, et al. Siamese neural network object tracking with distractor-aware model [J]. Acta Optica Sinica, 2020, 40(4): 0415002.

[16] Huang L H, Zhao X, Huang K Q. GOT-10k: a large high-diversity benchmark for generic object tracking in the wild[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(5): 1562-1577.

[17] Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.