融合通道互联空间注意力的 Siamese 网络跟踪算法

崔洲涓[1,2]，安军社[1]，崔天舒[1,2]

(1. 中国科学院国家空间科学中心 复杂航天系统电子信息技术重点实验室，北京 100190；
2. 中国科学院大学，北京 100049)

摘　要：基于 Siamese 网络的跟踪算法在跟踪精度和速度方面展现出巨大的潜力，然而要使离线训练的模型适应在线跟踪仍然面临着挑战。为了提升复杂场景下算法的特征提取以及判别能力，提出了一种融合通道-互联-空间注意力的 Siamese 网络实时跟踪算法。首先构建以深度卷积网络 VGG-Net-16 作为主干网络的 Siamese 跟踪框架，增加特征提取能力；接着设计通道-互联-空间注意力模块，增强模型的适应能力与判别能力；然后加权融合多层响应图，获取更精准的跟踪结果；最后使用大规模数据集对网络进行端到端的训练，在通用数据集 OTB-2015 上进行跟踪测试。实验结果表明：与当前主流算法相比，所提算法具有较强的稳健性，能更好地适应目标外观变化、相似物干扰、目标遮挡等复杂场景，在 NVIDIA RTX 2060 GPU 上，跟踪速度平均达到 37FPS，满足实时性要求。

关键词：目标跟踪；　Siamese 网络；　深度卷积网络；　通道注意力；　互联注意力；　空间注意力
中图分类号：TP391　　**文献标志码：**A　　**DOI：**10.3788/IRLA20200148

Siamese networks tracking algorithm integrating channel-interconnection-spatial attention

Cui Zhoujuan[1,2], An Junshe[1], Cui Tianshu[1,2]

(1. Key Laboratory of Electronics and Information Technology for Space Systems, National Space Science Center, Chinese Academy of Sciences, Beijing 100190, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: The tracking algorithms based on the Siamese networks show great potential in terms of tracking accuracy and speed. However, it is still challenging to adapt the offline trained model to online tracking. In order to improve the feature extraction and discrimination ability of the algorithm in complex scenes, a Siamese network real-time tracking algorithm that combines channel, interconnection and spatial attention mechanisms was proposed. First a Siamese tracking framework with a deep convolutional network VGG-Net-16 as the backbone network was built to increase feature extraction capabilities; then the channel-interconnection-spatial attention module was integrated to enhance the adaptability and discrimination capabilities of the model; then the multi-layer response maps were weighted and fused to obtain more accurate tracking results; and finally the large-scale datasets were used to train the end-to-end network, and tracking test on the benchmark OTB-2015 was completed. The experimental results show that compared with the current mainstream algorithms, the proposed algorithm is more robust and better adapt to complex scenes such as target appearance changes, similar distractors, and occlusion. On the NVIDIA RTX 2060 GPU, the average tracking speed reaches 37FPS, which meets real-time requirements.

Key words: object tracking;　Siamese networks;　deep convolutional networks;　channel attention;　interconnection attention;　spatial attention

收稿日期:2020−04−26;　**修订日期:**2020−06−02

基金项目:中国科学院复杂航天系统电子信息技术重点实验室自主部署基金 (Y42613A32S)

0 引 言

作为计算机视觉的研究方向之一，目标跟踪在视频监控、智能交通、军事制导、航天航空等领域都有广泛的应用。然而由于目标自身形变、旋转、运动模糊，以及外部应用场景光照变化、背景干扰、遮挡等因素的影响，建立一个高效稳健的目标跟踪算法依然面临着巨大的挑战。

随着深度学习方法在图像分类、目标检测等领域的应用取得了突破性的进展，研究者也逐步将其引入目标跟踪领域，利用深度卷积网络出色的特征表达能力提升算法的精度。HCF[1] 利用预训练神经网络提取深层和浅层特征，并根据不同层的特点，分别训练相关滤波器，得到响应图加权融合，提升了跟踪精度。C-COT[2] 通过连续空间域插值转换，将预训练神经网络提取的不同分辨率的特征图插值到连续空域，结合多分辨率的连续卷积滤波器进行训练和检测，取得了更加精确的定位。ECO[3] 从滤波器系数、样本划分和模板更新策略三个方面进行优化，在保持跟踪精度的同时，大幅度提高了跟踪速度。相比传统手工特征，这类算法利用预训练深度特征结合相关滤波框架，可以提升算法的跟踪精度。但预训练深度模型大多是针对图像分类任务而设计，并未能较好地适应目标跟踪。同时，越来越高的特征维度，给在线学习、更新过程带来了巨大的计算开销，直接限制了算法的跟踪速度。

深度学习方法在目标跟踪领域的应用不仅局限于预训练深度特征，为了利用端到端的优势，研究者们开始引入 Siamese 框架，训练专门的端到端跟踪网络。SINT[4] 开创性地引入 Siamese 网络，将目标跟踪任务转化为相似性学习问题，学习一个匹配函数。在跟踪过程中将后续帧的候选框与第一帧目标框进行匹配度计算，得分最高的即为目标。通过使用涵盖了各种目标变化可能情况的大规模训练数据集离线训练模板匹配网络，在不应用任何模型更新，没有遮挡检测，没有跟踪器的组合，没有几何匹配的情况下实现目标跟踪。SiamFC[5] 针对跟踪任务使用大规模视觉图像识别比赛中的视频离线训练一个深度网络，跟踪过程对候选图像进行全卷积，通过计算其两个输入互相关的双线性层完成滑窗在线评估，在保证高准确率的同时，速度也远超实时性。CFNet[6]将相关滤波转换为可微分的神经网络层，联合特征提取网络实现端到端的优化，训练与相关滤波器匹配的卷积特征，学习变换背景矩阵与目标外观矩阵，达到适应目标变化以及抑制背景变化的目的。SiamRPN[7] 引入区域候选网络通过边界框回归代替多尺度检测，网络结构包括特征提取的 Siamese 子网络和产生候选目标区域的 RPN 子网络。通过将 Siamese子网络模板分支、检测分支的特征同时输入到 RPN子网络的分类分支与回归分支，将跟踪问题转换成为单次学习匹配问题。DasiamRPN[8] 从训练数据不平衡、自适应的模型增量学习及长时跟踪等方面对算法进行了优化，在目标遮挡和长时跟踪等情况下表现突出。

尽管基于 Siamese 网络的端到端跟踪算法在通用数据集中取得了卓越的跟踪效果，但仍存在一些问题。离散训练网络只学习到了目标的通用特征，若在线跟踪过程中目标周围出现相似干扰物，则无法进行判断。在线跟踪的目标很可能并未包含在离线训练数据集中，那么相似性度量的结果未必可靠。由于特征提取与判别功能耦合在同一个网络里，缺少了分类器的在线训练环节，导致网络的适应能力、判别能力较弱。文中提出一种端到端的目标跟踪算法，承继 Siamese 网络架构的特点，选取更深的卷积网络模型 VGG-Net-16[9] 作为主干网络，使模型具有更强的特征表达能力；探索多种注意力机制，包括通道注意力、互联注意力以及空间注意力，使离线训练的模型在线跟踪时具有更优越的适应能力以及判别能力；通过加权融合响应得分图，以获取更精准的定位；在大规模数据集上离线训练端到端网络。在目标跟踪通用数据集中，进行算法性能的验证，与当前主流算法相比，文中算法在保证实时性的前提下，能够取得良好的跟踪精度。

1 基于 Siamese 网络的目标跟踪框架

1.1 Siamese 网络结构

Siamese 网络结构[10] 最早提出用于验证支票上的签名是否与银行预留的签名一致，后来 Sergey Zagoruyko[11] 等将其引入进行图像相似度的计算。

如图 1 所示，Siamese 网络结构通常由结构相似、权值共享的两个分支构成。将两个图片X_1、X_2分别输入两个分支，通过构建一种函数$G_w(X)$将其映射到特征空间，得到两个特征向量$G_w(X_1)$、$G_w(X_2)$，使用一种

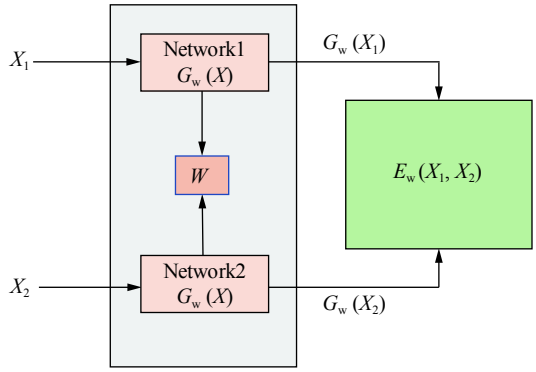


图 1 Siamese 网络结构图

Fig.1 Framework of Siamese network

距离度量$E_w(X_1,X_2)$, 作为两张图片的相似度计算函数, 衡量二者之间的差异。

1.2 基于 Siamese 网络的目标跟踪

近年来, 在深度学习的发展推动下, Siamese 网络被应用于目标跟踪领域, 并取得了令人瞩目的成绩。基于 Siamese 网络的代表性跟踪算法 SiamFC[5] 提出, 将目标跟踪任务转化为相似性度量问题, 利用大量样本数据离线训练一个 Siamese 网络 (由卷积层、非线性激活层和池化层构成), 将视频序列第一帧作为模板分支的输入图像z, 后续帧中的候选区域作为检测分支的输入图像x, 分别通过权值共享的特征提取网络$\varphi(\cdot)$, 将原始图像映射到特定的特征空间, 学习一个度量函数$f(z,x)$, 如公式 (1) 所示, 用于比较模板图像和候选区域搜索图像之间的相似度, 返回响应图, 分数越高, 二者相似度越高。其中$*$代表互相关运算, $b\cdot I$表示在响应图中每个位置的取值。

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

2 融合注意力的 Siamese 网络目标跟踪

2.1 算法框架

算法框架如图 2 所示。

主要包括 Siamese 深度网络、通道-互联-空间注意力模块 (Channel-Interconnection-Spatial Attention Module, CISAM)、互相关运算以及响应得分图融合四部分组成。其中 Siamese 深度网络基于 VGG-Net-16 进行构建, 再融入 CISAM。图像序列经过改进后的 Siamese 网络, 将原始图像映射到特定的特征空间, 得到多层特征图, 分别进行互相关运算, 输出响应得分

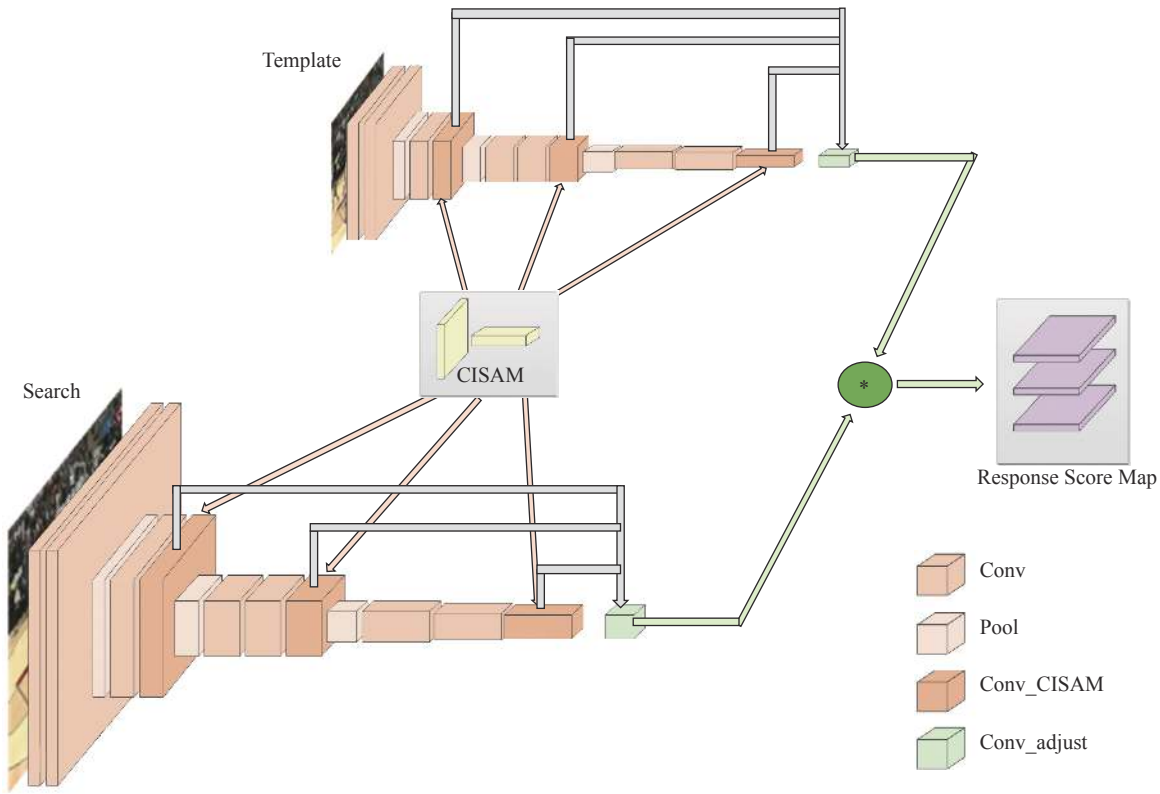


图 2 算法框架图

Fig.2 Framework of algorithm

图 (Response Score Map) 进行融合，得分越高，则相应位置是目标的概率越大。

2.2 Siamese 深度网络

特征提取是决定跟踪算法性能的最关键因素[12]。为了提升网络的特征提取能力，而又不引入填充破坏网络的平移不变性，文中网络基于适应能力更强的 VGG-Net-16 进行构建。

VGG-Net-16 是由 3×3 的小型卷积核以及 2×2 的最大池化层反复堆叠构筑而成的卷积神经网络，通过小型卷积核的深度复用模拟较大卷积核完成对图像的局部感知，提升网络的性能。模型主要由 5 段卷积、2 个全连接特征层以及 1 个全连接分类层组成。VGG-Net-16 能够提取到非常丰富的特征，将其提取到的不同层不同通道的部分特征图可视化，如图 3 所示。

为了将 VGG-Net-16 更好地运用到文中所提算法中，根据 Siamese 网络的特点，同时考虑到后续互相关以及响应图的融合等操作，对 VGG-Net-16 进行修改，具体网络结构如表 1 所示。从表中可以看出，修改主要包括以下四个方面。

(1) 删减：考虑到目标跟踪任务对精细位置的需求，删除第五段卷积，缩小了网络模型的总步长；

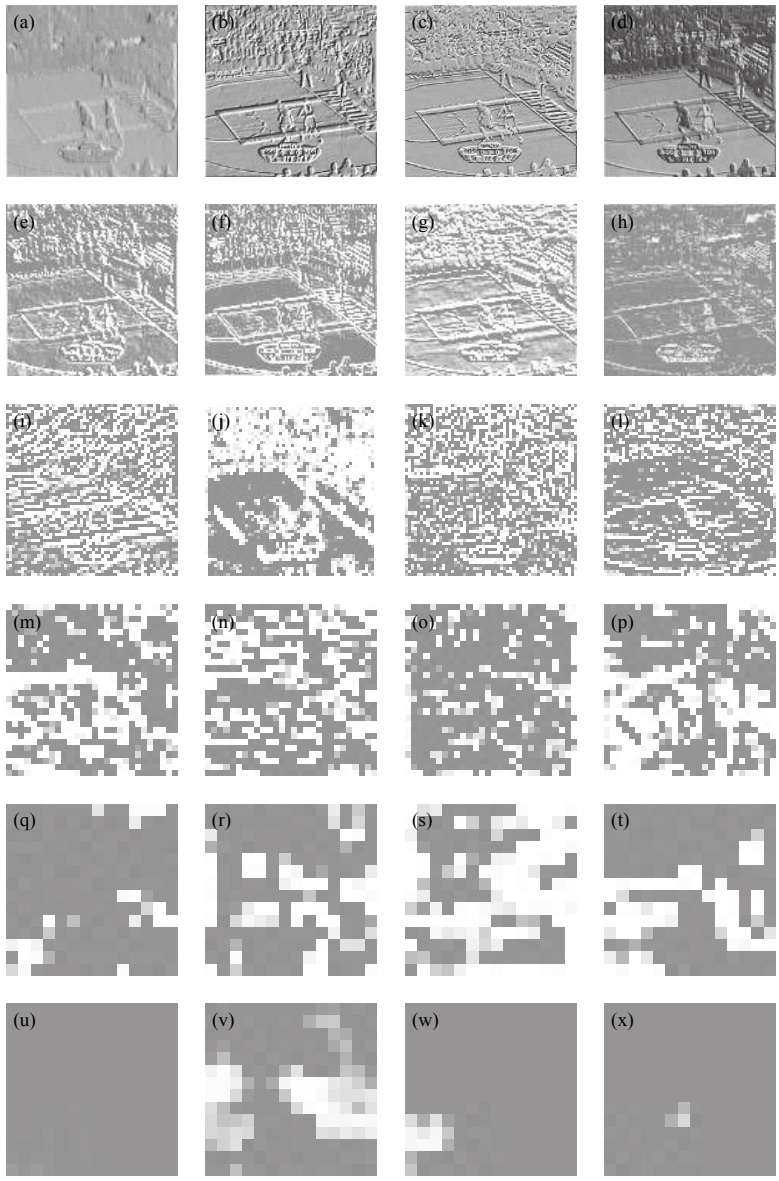


图 3 VGG-Net-16 提取的部分特征图

Fig.3 A set of feature maps extracted by VGG-Net-16

(2) 增加：在网络模型内，在卷积层 Conv2_2、Conv3_3、Conv4_3 后均融入了通道-互联-空间注意力模块 CISAM，形成特征图层 Conv_CISAM，尺寸维持不变，分别记作 Conv2_2_CISAM、Conv3_3_CISAM 以及 Conv4_3_CISAM；

(3) 提取：为了利用多分支特征协同进行定位，将 Conv2_2_CISAM、Conv3_3_CISAM 以及 Conv4_3_CISAM 分别提取；

(4) 调整：为了便于后续响应得分图计算时的融合，在进行互相关运算前，将多分支特征 Conv2_2_CISAM、Conv3_3_CISAM 以及 Conv4_3_CISAM 分别输入调整层 Conv_adjust 进行调节。

表 1 Siamese 网络结构

Tab.1 Architecture of Siamese network

Layer name	Kernel size	Chan. map	Template size	Search size	Channel output	Stride	CISAM
Input			127×127	255×255	3	-	No
Conv1_1	3×3	64×3	125×125	253×253	64	1	No
Conv1_2	3×3	64×64	123×123	251×251	64	1	No
Pool1	2×2		61×61	125×125	64	2	No
Conv2_1	3×3	128×64	59×59	123×123	128	1	No
Conv2_2	3×3	128×128	57×57	121×121	128	1	Yes
Pool2	2×2		28×28	60×60	128	2	No
Conv3_1	3×3	256×128	26×26	58×58	256	1	No
Conv3_2	3×3	256×256	24×24	56×56	256	1	No
Conv3_3	3×3	256×256	22×22	54×54	256	1	Yes
Pool3	2×2		11×11	27×27	256	2	No
Conv4_1	3×3	512×256	9×9	25×25	512	1	No
Conv4_2	3×3	512×512	7×7	23×23	512	1	No
Conv4_3	3×3	512×512	5×5	21×21	512	1	Yes

2.3 通道-互联-空间注意力模块 CISAM

仅仅增加网络的深度只能相对提升特征表达能力，并不能从根本上解决 Siamese 框架存在的问题。因为相对于相关滤波类的跟踪算法，基于 Siamese 网络的跟踪算法缺少了分类器的在线训练环节，在同一个网络中耦合了特征提取与判别功能。这需要网络同时具备两种特质：一是能够稳定地适应目标自身在各种场景中的变化，抽象出目标代表性、本质性的特征；二是能够敏感地区分目标及相似物，显著地提炼出二者之间的差异。

设经过 Siamese 主干网络提取到的特征图分别为 $\varphi(z)\in\mathbb{R}^{C\times H_{\mathrm{T}}\times W_{\mathrm{T}}}$、$\varphi(x)\in\mathbb{R}^{C\times H_{\mathrm{D}}\times W_{\mathrm{D}}}$，其中，$C$为通道数，$H_{\mathrm{T}}$、$H_{\mathrm{D}}$为特征图在垂直方向的尺寸，$W_{\mathrm{T}}$、$W_{\mathrm{D}}$为特征图在水平方向的尺寸，$H_{\mathrm{T}}\leqslant H_{\mathrm{D}}$、$W_{\mathrm{T}}\leqslant W_{\mathrm{D}}$，二者进行互相关运算得到的 $f(z,x)\in\mathbb{R}^{C\times H_f\times W_f}$，其中响应图垂直方向、水平方向的尺寸分别为 $H_f=H_{\mathrm{D}}-H_{\mathrm{T}}+1$、$W_f=W_{\mathrm{D}}-W_{\mathrm{T}}+1$，$b$为响应图中每个位置的权值。将 Siamese 网络的匹配函数公式 (1) 展开如公式 (2) 所示：

$$f(z,x)=\sum_{c=0}^{C-1}\sum_{h=0}^{H_{\mathrm{T}}-1}\sum_{w=0}^{W_{\mathrm{T}}-1}\varphi_{c,h,w}(z)\,\varphi_{c,H_f+h,W_f+w}(x)+b \tag{2}$$

可以看到，互相关的计算过程对特征图的通道层面、空间层面的关注度非常平均，权重系数均为 1。事实上在跟踪任务中，不同通道、不同位置的特征重要性是不同的。研究表明[13]，注意力机制在人类视觉过程中起到重要作用，图像中的高层语义特征能够吸引人类的视觉注意力。因此，可以借鉴性地引入注意力机制弱化贡献度小的特征图，强化贡献度大的特征图，关注目标前景和语义背景的差异特征，实现对目标的增强、对干扰的抑制以及对不同对象间的判别，提高复杂场景下算法的稳健性与实时性。为此，选择在特征提取主干网络的卷积层中融入注意力模块，通过调节权重参数突出或筛选目标的重要信息，抑制无

关的细节信息，提升网络的判别能力，获得相当于分类器在线学习的功能。

CBAM[14] 提出将注意力机制同时运用在通道和空间两个层面，可以嵌入目前大部分主流深度网络，在不显著增加计算量和参数量的前提下能提升网络模型的特征提取能力。考虑到 Siamese 网络的结构特点，文中在此基础上，加入了互联注意力模块，形成通道-互联-空间注意力模块 CISAM，具体结构如图 4 所示。

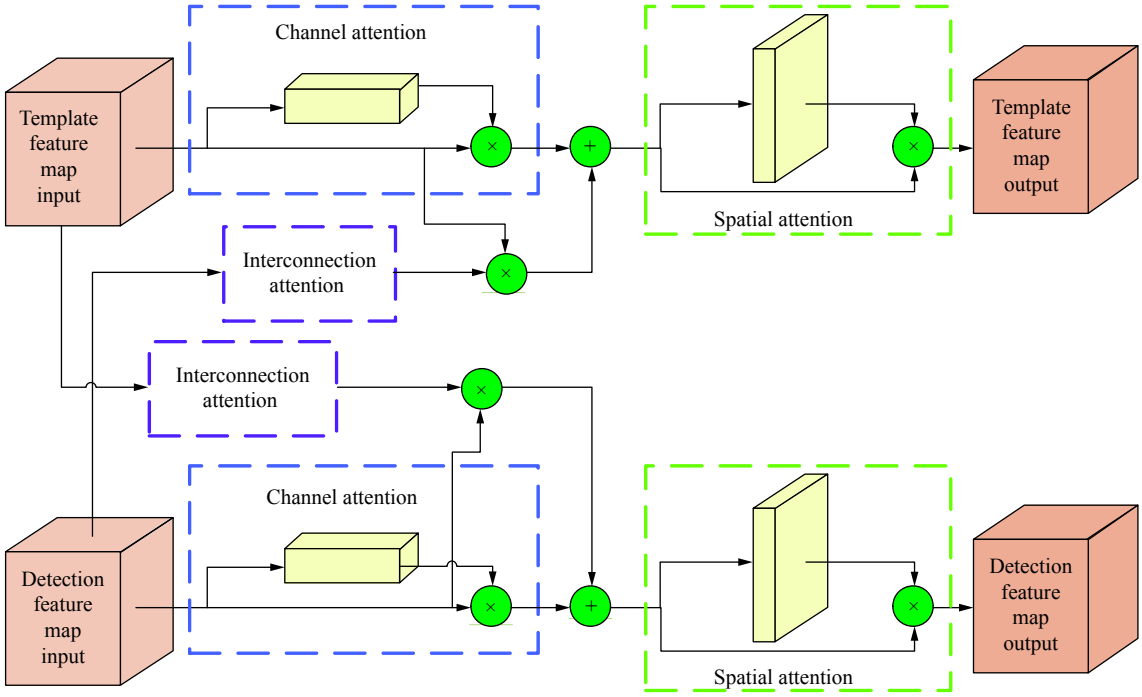


图 4 CISAM 结构图

Fig.4 Structure of CISAM

模板分支、检测分支卷积层输出的特征图 $\varphi(z)$、$\varphi(x)$，经过通道注意力模块得到注意图 $A_{\mathrm{C}}\{\cdot\}\in\mathbb{R}^{C\times1\times1}$，通过公式 (3)、(4) 分别输出通道特征图 $\varphi_{\mathrm{C}}(z)$、$\varphi_{\mathrm{C}}(x)$，⊗表示元素点乘。

$$\varphi_{\mathrm{C}}(z)=A_{\mathrm{C}}\{\varphi(z)\}\otimes\varphi(z) \tag{3}$$

$$\varphi_{\mathrm{C}}(x)=A_{\mathrm{C}}\{\varphi(x)\}\otimes\varphi(x) \tag{4}$$

通道注意力模块对输入特征图在空间维度进行压缩，同时使用全局平均池化与全局最大池化，前者对特征图上的每一个像素点都进行反馈，而后者对全局平均池化进行补充。每个通道等价为一个不同类型的特征判别器，从语义层面对特征优化选择，激活与目标相关程度更高的，删除冗余的通道特征。

同时，模板分支、检测分支卷积层输出的特征图 $\varphi(z)$、$\varphi(x)$，经过互联注意力模块得到注意图 $A_{\mathrm{I}}\{\cdot\}\in\mathbb{R}^{C\times C}$，由于两个分支的空间尺寸并不相同，进行互联时需要先调整维度，得到 $\varphi_{\mathrm{RS}}(z)\in\mathbb{R}^{C\times P_{\mathrm{T}}}$、$\varphi_{\mathrm{RS}}(x)\in\mathbb{R}^{C\times P_{\mathrm{D}}}$，其中 $P_{\mathrm{T}}=H_{\mathrm{T}}\times W_{\mathrm{T}}$、$P_{\mathrm{D}}=H_{\mathrm{D}}\times W_{\mathrm{D}}$，通过公式 (5)、(6) 分别输出互联特征图 $\varphi_{\mathrm{IRS}}(z)$、$\varphi_{\mathrm{IRS}}(x)$。再调整恢复维度后输出互联特征图 $\varphi_{\mathrm{I}}(z)$、$\varphi_{\mathrm{I}}(x)$。

$$\varphi_{\mathrm{IRS}}(z)=A_{\mathrm{I}}\{\varphi_{\mathrm{RS}}(x)\}\otimes\varphi_{\mathrm{RS}}(z) \tag{5}$$

$$\varphi_{\mathrm{IRS}}(x)=A_{\mathrm{I}}\{\varphi_{\mathrm{RS}}(z)\}\otimes\varphi_{\mathrm{RS}}(x) \tag{6}$$

互联注意力模块将两个分支分别编码至另外一个分支，可以帮助结合目标及其背景之间的相关性。两个分支的通道注意力与互联注意力分别汇合后，再经由空间注意力模块得到注意图 $A_{\mathrm{S}}\{\cdot\}\in\mathbb{R}^{1\times H_{\mathrm{T}}\times W_{\mathrm{T}}}$、$A_{\mathrm{S}}\{\cdot\}\in\mathbb{R}^{1\times H_{\mathrm{D}}\times W_{\mathrm{D}}}$，通过公式 (7)、(8) 融合提取出特征图 $\varphi_{\mathrm{S}}(z)$、$\varphi_{\mathrm{S}}(x)$：

$$\varphi_{\mathrm{S}}(z)=A_{\mathrm{S}}\{[\varphi_{\mathrm{C}}(z)+\varphi_{\mathrm{I}}(z)]\}\otimes[\varphi_{\mathrm{C}}(z)+\varphi_{\mathrm{I}}(z)] \tag{7}$$

$$\varphi_{\mathrm{S}}(x)=A_{\mathrm{S}}\{[\varphi_{\mathrm{C}}(x)+\varphi_{\mathrm{I}}(x)]\}\otimes[\varphi_{\mathrm{C}}(x)+\varphi_{\mathrm{I}}(x)] \tag{8}$$

空间注意力模块使用全局平均池化与最大池化对输入特征图在通道层面进行压缩，得到两个二维的特征图，并拼接得到一个通道数为 2 的特征图，经由隐藏层对其进行卷积操作，保证特征图在空间维度上

保持不变。空间注意力更集中于位置的描述，与通道注意力的相互补充，通过构建特征图中不同位置之间的联系，从空间层面学习特征图的哪些部分应该会有更高的响应，针对位置进行加权融合。

2.4 响应得分图融合

充分利用不同层次的特征图，能够提升跟踪性能。浅层特征空间分辨率较高，纹理信息全面，易于定位目标，能够防止表观相似的干扰背景。深层特征语义信息丰富，能够稳定适应目标形变、遮挡等表观变化。为了降低复杂场景中干扰源的影响，获取高质量的响应得分图，采用多层特征融合的方法。首先将 CISAM 输出的特征图通过调整层将通道数统一调整为 256；接着分别计算模板分支与检测分支中各调整层输出的第q层特征进行互相关运算；最后将响应得分图以加权的方式进行融合，如公式 (9)，其中α_q为第q层的权重参数。综合分析响应得分图，确定参数的具体值。

$$S_{\mathrm{RM}}=\sum_{q=2,3,4}\alpha_q S_{\mathrm{RM}}^{(q)} \tag{9}$$

2.5 网络训练

(1) 数据集的准备

训练集中的数据选自 ImageNet VID[15]、COCO[16] 和 Youtube-bb[17]，通过特定的比例组合。从相同的视频序列中随机选取两帧，并将它们组合成一对模板图像 (127×127) 和搜索图像 (255×255)，作为 Siamese 网络的输入。

为保证所有的输入图像均为正方形，按照公式 (10) 调整，其中(w,h)为原始目标框尺寸，p为上下文边距，s为尺度变换因子，$A=127^2$。

$$s(w+2p)\times s(h+2p)=A \tag{10}$$

正样本定义为响应得分图中的元素u在响应得分图中心c的搜索区域半径R内，其余的视作负样本，k为全卷积网络的总步长。文中设定$R=2$，$k=1$。

$$\begin{cases}+1 & \text{if } k\|u-c\|\leqslant R\\ -1 & \text{otherwise}\end{cases} \tag{11}$$

(2) 网络模型的搭建

按照表 1 定义的网络结构进行网络模型的搭建，按照 VGG-Net-16 预训练模型进行初始化。

(3) 损失函数与优化器的选择

损失函数采用逻辑回归损失函数，如公式 (12) 所示：

$$\ell(y,v)=\log(1+\exp(-yv)) \tag{12}$$

式中：$y\in\{+1,-1\}$代表候选样本对对应的标签值；v代表单个图像样本对的实际相似计算得分值。整体响应图的损失函数定义为全部点的损失函数的均值，可表示为：

$$L(y,v)=\frac{1}{|D|}\sum_{u\in D}\ell(y[u],v[u]) \tag{13}$$

式中：D为响应得分图中像素数目；$u\in D$为响应得分图各个位置；$v[u]$为每个位置的互相关卷积计算值各个搜索位置u的响应得分；$y[u]$代表响应得分图各个位置的真值标签正负样本的划分。

最小化目标函数如公式 (14) 所示，模型参数θ指代的是网络的各层参数，采用随机梯度下降 (SGD) 对网络参数进行优化，从而得到网络参数的最优值，使其具有识别输入样本对特征共通性的能力。

$$\arg\min_{\theta}=\underset{z,x,y}{E}L(y,f(z,x;\theta)) \tag{14}$$

(4) 网络训练

训练时学习率设置为 10^{-3}~10^{-6}。整个训练过程包含 100 多个阶段，每个阶段由 6000 对样本组成。每次计算 8 对样本的平均损失值。

3 实验与分析

3.1 注意力可视化分析

为分析文中通道-互联-空间注意力模块 CISAM 的作用，利用类激活热力图 Grad-CAM[18] 对不同的网络进行可视化，如图 5 所示。越敏感的位置温度越高，越不敏感的位置温度越低。第一排是未加 CISAM 的网络模型可视化结果，第二排是加入 CISAM 后的网络模型可视化结果。实验结果可以看出添加 CISAM 模块的网络的注意范围更广，能够更好地覆盖到所要识别的物体。当周围有相似物出现时，能更好地进行区分，将注意力集中在目标上，并未受到相似物体的干扰。

3.2 实验配置

文中所提跟踪算法的实验平台硬件配置如表 2 所示。

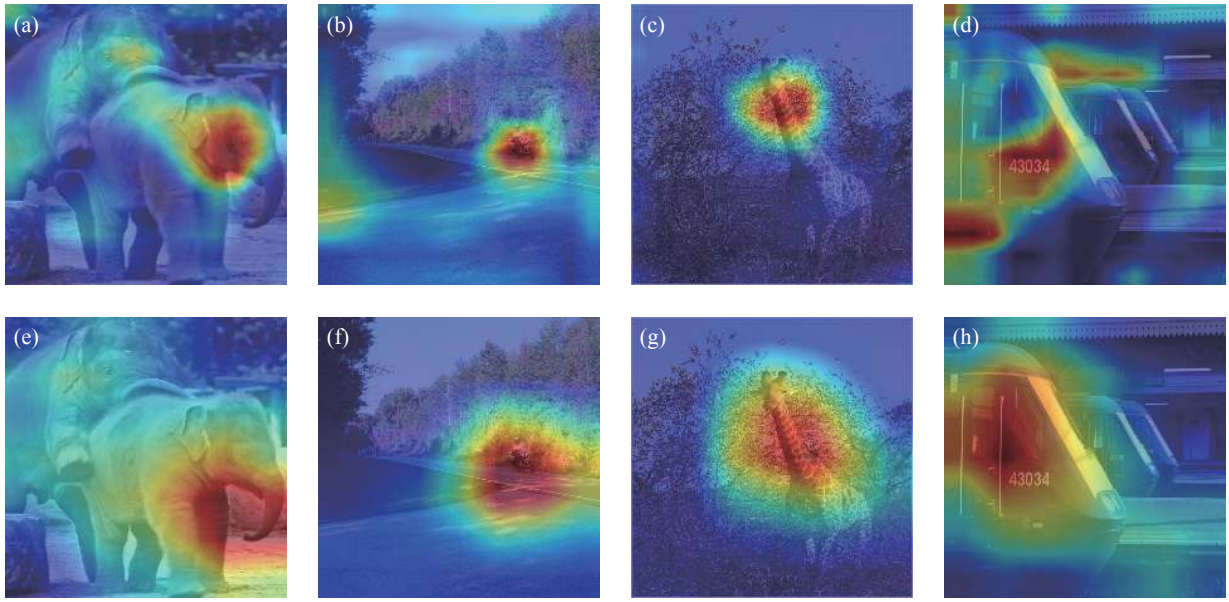


图 5 Grad-CAM 网络可视化结果

Fig.5 Grad-CAM network visualization results

表 2 实验平台参数

Tab.2 Experimental platform parameters

Device	Product model	Memory
CPU	Intel(R)Core(TM)i7-9 700 Basic frequency 3.0 GHz	16G
GPU	NVIDIA GeForce RTX-2 060	6G

算法基于 Pytorch 实现，选择当前具有代表性的六种算法，包括文中所提算法、ECO[3]、SiamFC[5]、CFNet[6]、SiamRPN[7]、DasiamRPN[8]，在目标跟踪通用数据集 OTB-2015[19](涵盖 11 个属性，共 100 组视频) 上测试评估算法的性能。以一次通过评估 OPE (One Pass Evaluation) 作为跟踪算法准确性能的评价标准。

3.3 定性分析

六种算法在数据集部分视频序列中的跟踪结果，如图 6 所示。不同算法的目标框使用不同的颜色表示，其中文中算法用红色表示。

图 6(a) CarDark 序列中，目标时而变模糊，时而被遮挡，周围存在相似类干扰，同时光照情况也不断发生变化，文中算法由于可以筛选出适应性更强的目标特征，区分背景，所以始终未发生漂移。

图 6(b) MotorRolling 序列中，目标自身发生超过 360°的旋转，对特征表达能力提出了极大的挑战，除文中所提算法、SiamRPN 以及 DasiamRPN，其他算法在跟踪初始已经偏离了目标。

图 6(c) Skating1 序列中，光照不断发生变化，且目标周围不断出现相似类干扰，多数算法已经远离目标，只有文中算法、SiamRPN 以及 ECO 准确定位。

图 6(d) Skiing 序列中，目标尺寸较小且速度较快，这对特征提取提出了更严格的要求，特征提取能力或者匹配能力低的算法在跟踪初始已经无法定位目标。

图 6(e) Suv 序列中，目标清晰度低，偶尔超出视野范围，文中算法表现稳健，能够保持跟住目标。

图 6(f) Walking2 序列中，目标尺度变化，中间偶有被遮挡，且有相似类干扰出现，文中算法能够避免遮挡出现时，跟踪漂移。

通过对以上涵盖属性较为丰富的视频序列进行分析，文中所提算法在光照变化、尺度变化、旋转、相似物干扰、快速运动、分辨率低、超出视野等情况均有极佳的表现。

3.4 定量分析

跟踪算法主要通过对中心位置误差和覆盖率两个指标进行评估。前者是指跟踪结果与真实目标的中心位置之间的欧式距离；后者是指跟踪结果与真实目标的重叠率，均通过一定的阈值判定跟踪是否成功。二者分别体现在精度图以及成功率图中。

(1) 六种算法在 OTB-2015 上的跟踪精度曲线以及成功率曲线分别如图 7、图 8 所示。文中算法跟踪

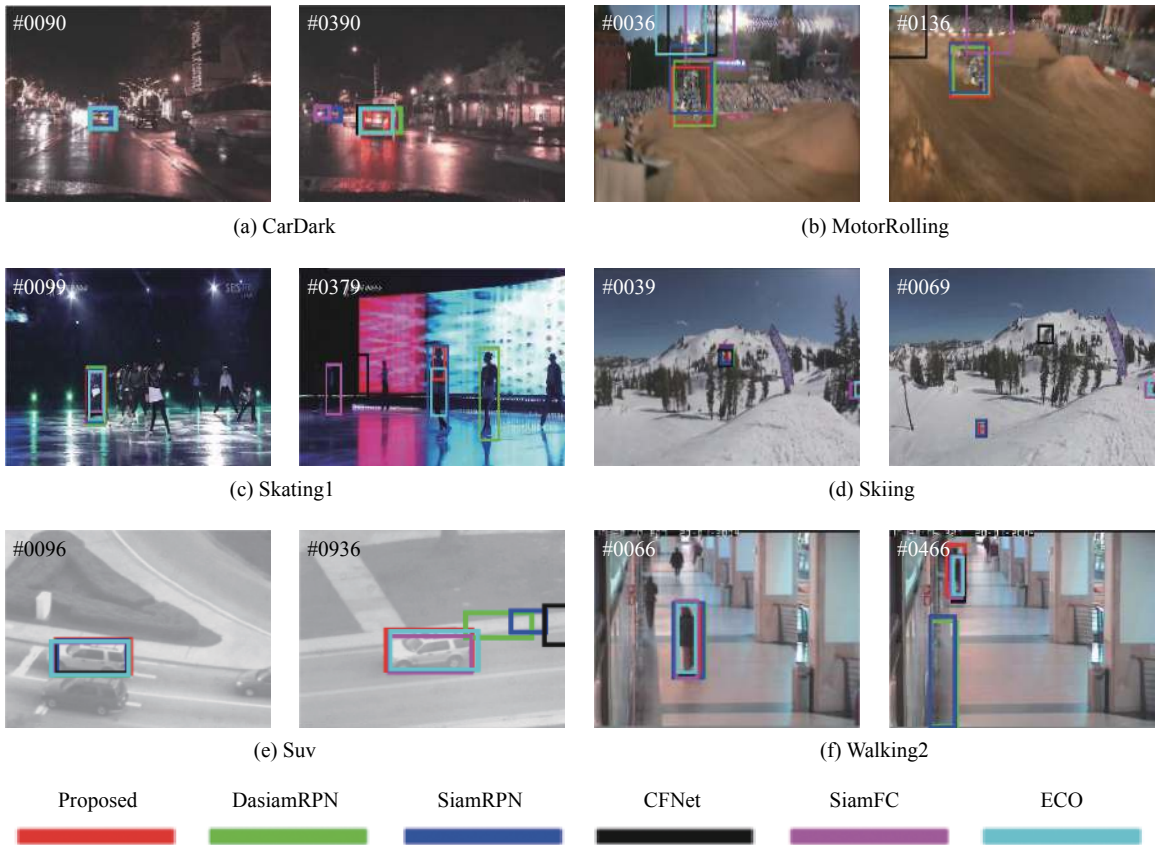


图 6 六种跟踪算法在部分视频序列上的定性结果显示

Fig.6 Qualitative results display of the 6 tracking algorithms over partial video sequences

精度达到 0.896, 成功率为 0.689, 相对于 SiamFC 算法分别提高了 12.4%、10.3%。证明文中算法在 Siamese 网络框架的基础上, 提取了更深的网络特征, 同时融入了注意力模块, 使得网络提取到适应能力更强的特征, 提升了算法的总体精度与稳健性。

(2) 六种算法在 OTB-2015 中, 11 种不同属性视频的跟踪精度定量分析结果如图 9 所示, 文中算法在尺度变化、形变、平面内旋转属性中排名第一, 在光照变化、平面外旋转、遮挡、运动模糊、快速运动、超出视野、低分辨率属性中排名第二, 仅次于 ECO。

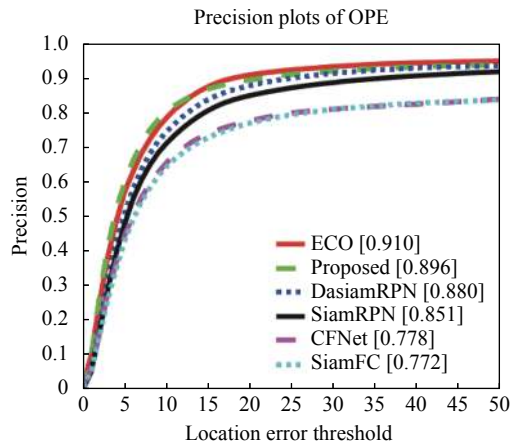


图 7 基于 OTB-2015 OPE 的跟踪精度曲线图

Fig.7 Tracking precision plots of OPE on OTB-2015 dataset

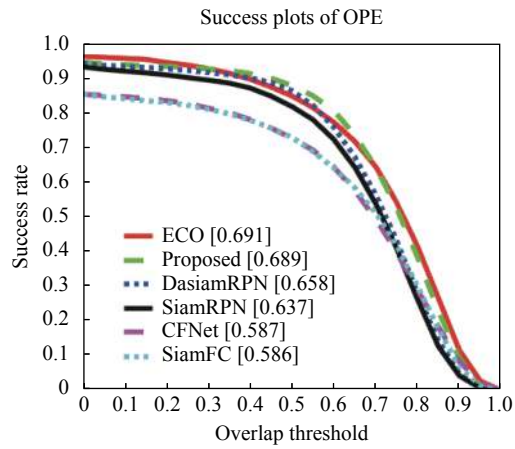


图 8 基于 OTB-2015 OPE 的成功率曲线图

Fig.8 Success plots of OPE on OTB-2015 dataset

(3) 六种算法在 OTB-2015 中不同视频属性的成功率定量分析结果如图 10 所示, 文中算法在尺度变化、形变、平面内旋转、低分辨率属性中排名第一, 在光照变化、平面外旋转、遮挡、运动模糊、快速运动、超出视野、背景复杂属性中排名第二, 仅次于 ECO。

(4) 相较于 ECO, 文中算法跟踪精度在部分场景仍有差距。然而基于深度网络特征的 ECO 算法在 GPU 上的速度为 8 FPS[3], 文中算法的平均速度达到

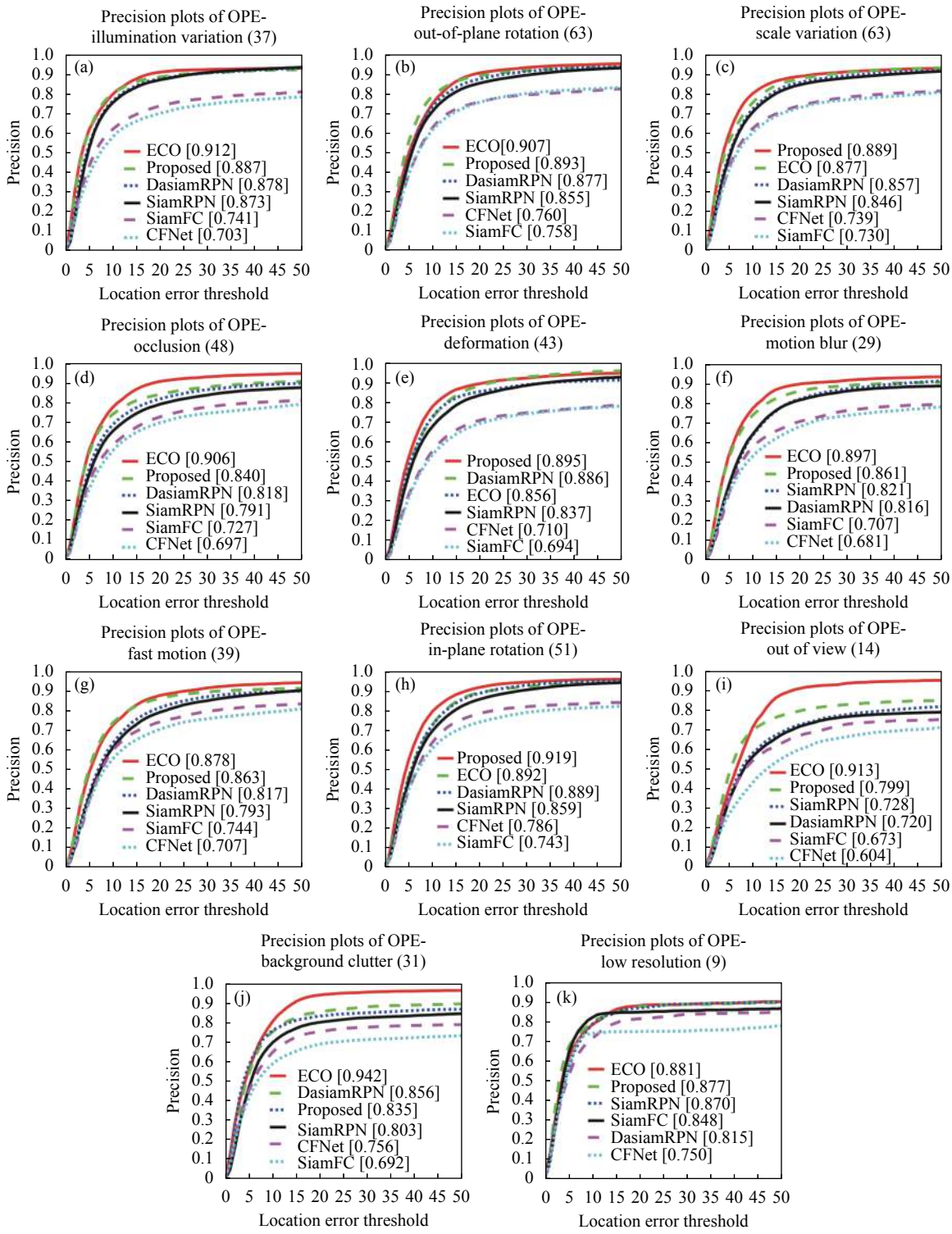


图 9 OTB-2015 11 种不同属性视频序列跟踪精度曲线

Fig.9 Tracking precision plots of 11 different attributes video sequences on OTB-2015 dataset

37 FPS, 在跟踪速度上有了提升。

综上, 文中所提算法在 OTB-2015 共 11 种属性的视频序列中表现稳定, 通过融合注意力模块, 嵌入更深的网络, 取得了很好的跟踪效果, 在提高跟踪精度与稳健性的同时, 还保证了实时的跟踪速度, 较好地适应跟踪场景的变化。

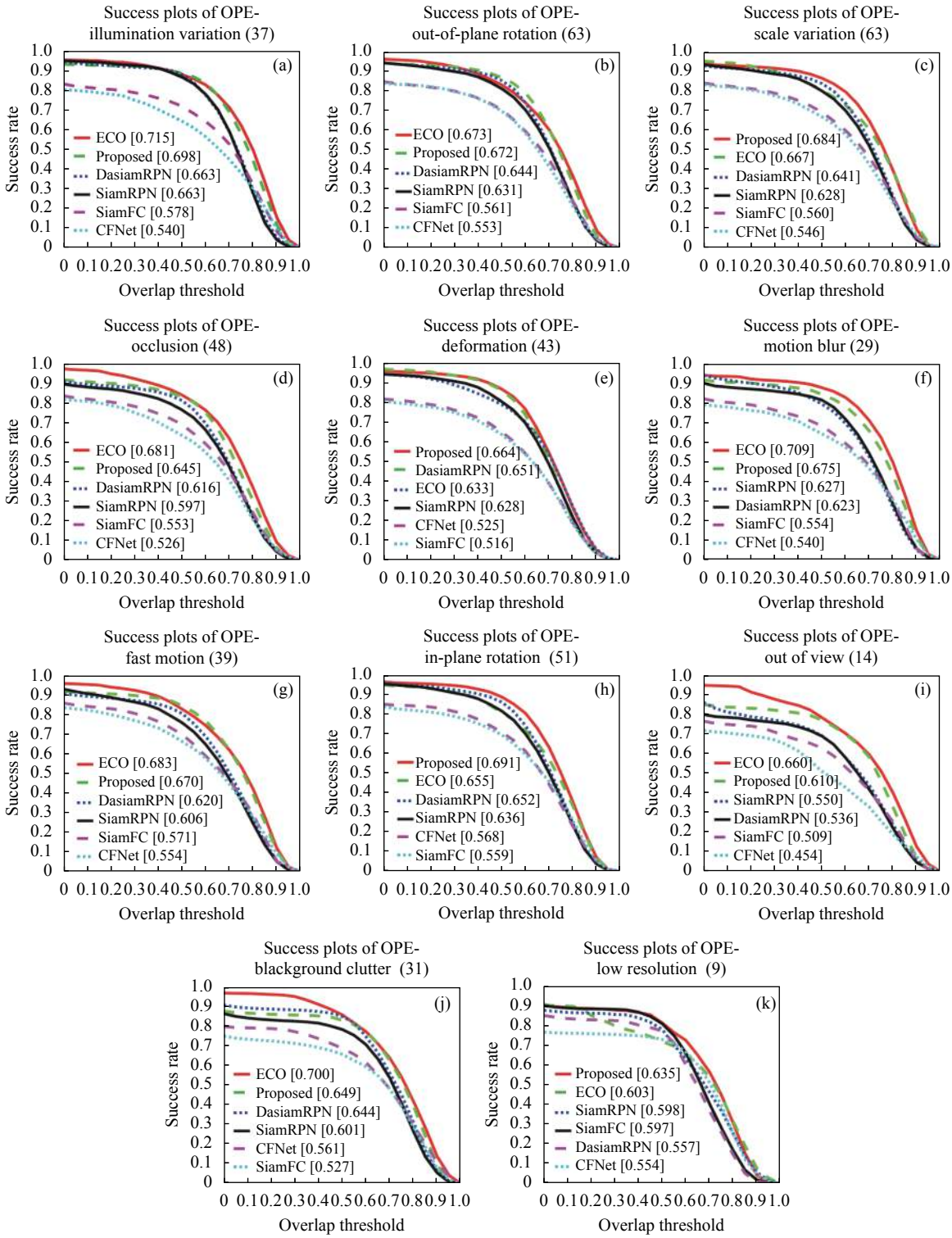


图 10 OTB-2015 11 种不同属性视频序列成功率曲线

Fig.10 Success plots of 11 different attributes video sequences on OTB-2015 dataset

4 结 论

文中对基于 Siamese 网络的目标跟踪算法进行了研究，提出一种嵌入深度卷积网络 VGG-Net-16 作为特征提取主干网络、融合注意力机制的端到端跟踪算法，并通过实验对其进行了验证，得到以下结论：

(1) 通道-互联-空间注意力模块的融入，显著提升了网络模型的特征提取能力、适应能力与判别能力。可视化分析实验表明，注意力模块的存在使得网络覆盖到目标更多的部位，学习更具代表性的特征，同时还可以排除周围相似物的干扰，区分性更强。

(2) 在通用目标跟踪数据集 OTB-2015 上的实验表明，与当前主流算法相比，文中算法能够达到较高的跟踪精度，在目标外观变化、相似物干扰、目标遮挡等复杂场景下，有稳健的表现。

(3) 在 NVIDIA RTX 2060 GPU 下的平均跟踪速度可达到 37 FPS，满足实时应用的要求。

注意力机制与目标跟踪算法的融合还处于探索阶段，后续工作会集中于注意力模块在网络中嵌套方式的研究，更进一步提升算法的性能。

参考文献：

[1] Ma C, Huang J B, Yang X, et al. Hierarchical convolutional features for visual tracking[C]//Proceedings of the IEEE International Conference on Computer Vision, 2015: 3074−3082.

[2] Danelljan M, Robinson A, Khan F S, et al. Beyond correlation filters: Learning continuous convolution operators for visual tracking[C]//ECCV, 2016.

[3] Danelljan M, Bhat G, Khan F S, et al. Eco: Efficient convolution operators for tracking[C]//CVPR, 2017.

[4] Tao Ran, Gavves E, Smeulders A W M. Siamese instance search for tracking[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016: 1420−1429.

[5] Bertinetto L, Valmadre J, Henriques J F, et al. Fully-Convolutional Siamese Networks for Object Tracking[M]//Hua G, Jegou H. Computer Version ECCV 2016 Workshops. Cham: Springer, 2016, 9914: 850−865.

[6] Valmadre J, Bertinetto L, Henriques J F, et al. End-to-end representation learning for correlation filter based tracking[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017: 5000−5008.

[7] Li Bo, Yan Junjie, Wu Wei, et al. High performance visual tracking with Siamese region proposal network[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018: 8971−8980.

[8] Zhu Zheng, Wang Qiang, Li Bo, et al. Distractor-aware Siamese networks for visual object tracking[C]//The 15th European Conference on Computer Vision, 2018: 103−119.

[9] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10)[2018-12-15]. https://arxiv.org/abs/1409.1556.

[10] Bromley J, Guyon I, LeCun Y, et al. Signature verification using a “Siamese” time delay neural network[C]//Advances in Neural Information Processing Systems, 1994: 737−744.

[11] Zagoruyko S, Komodakis N. Learning to compare image patches via convolutional neural networks[C]//CVPR, 2015.

[12] Wang N, Shi J, Yeung D, et al. Understanding and Diagnosing Visual Tracking Systems[C]//2015 IEEE International Conference on Computer Vision (ICCV), 2015: 3101−3109.

[13] Li Wanyi, Wang Peng, Qiao Hong. A survey of visual attention based methods for object tracking [J]. ***Acta automatica sinica***, 2014, 40(4): 561-576. (in Chinese)
黎万义, 王鹏, 乔红. 引入视觉注意机制的目标跟踪方法综述[J]. 自动化学报, 2014, 40(4): 561-576.

[14] Woo S, Park J, Lee J Y, etal. CBAM: Convolutional Block Attention Module[C]// European Conference on Computer Vision, 2018: 3−19.

[15] Russakovsky O, Deng J, Su H, et al. Image net large scale visual recognition challenges[C]//IJCV, 2015.

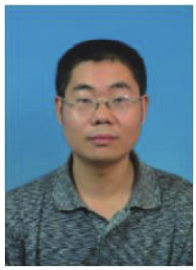
[16] Lin T-Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//ECCV, 2014: 740−755.

[17] Real E, Shlens J, Mazzocchi S, et al. Youtube boundingboxes: A large high-precision human-annotated data set for object detection in video[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017: 7464−7473.

[18] Selvaraju R R, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 618−626.

[19] Wu Y, Lim J, Yang M H. Object tracking benchmark [J]. ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, 2015, 37(9): 1834-1848.

第一作者简介： 崔洲涓 (1986–), 女, 河南安阳人, 计算机应用技术专业博士生, 主要从事计算机视觉、深度学习、目标跟踪等方面的研究。Email: constance669@126.com



导师简介： 安军社 (1969–), 男, 陕西富平人, 研究员、博士生导师, 计算机软件与理论专业博士。以第一作者在核心刊物上发表论文 8 篇, 其他刊物和会议论文 6 篇。申请专利 16 项, 其中以第一发明人获得发明专利 4 项, 已获专利 4 项, 作为合作者获得专利超过 30 项。参加载人航天、探月工程、中俄联合火星探测、北斗导航和自主火星探测等多项国家重点项目。主要研究领域为空间飞行器综合电子系统, 主要研究方向为星载计算平台、星载数据网络技术等。获部委奖励 3 次。曾获科学技术进步奖三等奖, 获部委科技进步一等奖; 获科工委颁发的"绕月探测工程初样研制建设先进个人"称号, 中国科学院颁发的"参加载人航天工程优秀工作者"荣誉称号, 中国科学院空间科学与应用总体部颁发的"神舟 2 号飞船应用系统任务先进个人"; 1999 年 11 月, 载人航天第一次飞行试验突出贡献奖成员。主要从事航天计算器等方面的研究。Email: anjunshe@nssc.ac.cn