第 41 卷 第 6 期/2021 年 3 月/光学学报 研究论文

光学学报

基于全局感知孪生网络的红外目标跟踪

李畅，杨德东*，宋鹏，郭畅

河北工业大学人工智能与数据科学学院，天津 300130

摘要 目前大多数热红外(TIR)目标跟踪算法都是基于相关滤波或者使用彩色跟踪器的模型进行特征提取。然而，两者都存在适用于彩色目标跟踪却对红外目标特征不敏感的缺陷，导致无法良好地应用到红外目标跟踪。为此，提出一种基于全局感知的孪生神经网络的红外目标跟踪器。将孪生神经网络的后三层特征进行融合优化，得到新的特征，同时加入了由空间转换网络和通道注意力组成的空间感知模块，以得到全局范围内的有效信息，通过引入自注意力机制，使算法更加专注于提取目标的判别信息，最后对结果进行响应融合得到最终的响应图。在PTB-TIR 红外目标跟踪评估基准上的实验结果表明，本文算法能够适应多样的红外环境，同时能够保持良好的跟踪速度(20.2 frame/s)，实现对红外目标有效且稳定的实时跟踪。

关键词 机器视觉；红外图像；目标跟踪；孪生神经网络；注意力机制

中图分类号 TP391.4 **文献标志码** A **doi**：10.3788/AOS202141.0615002

Global-Aware Siamese Network for Thermal Infrared Object Tracking

Li Chang, Yang Dedong*, Song Peng, Guo Chang

School of Artificial Intelligence, Hebei University of Technology, Tianjin 300130, *China*

Abstract At present, most thermal infrared (TIR) tracking methods are based on correlation filters or RGB trackers for feature extraction. However, both of them are only suitable for RGB object tracking but not sensitive to the TIR object features, thus failing to be applied to the TIR object tracking. To this end, a TIR object tracker based on the global-aware siamese neural network was proposed in this paper. First, the features from the last three layers of the siamese neural network were fused to obtain new features. Second, the spatial-aware module composed of the spatial transformer network and channel attention was added to get the global effective information. Simultaneously, the self-attention mechanism was introduced to make the algorithm more focus on extracting the discriminant information of the objects. At last, the final response map was acquired by response fusion of the results. The experimental results on the TIR pedestrian tracking benchmark (PTB-TIR) show that the proposed algorithm can adapt to a variety of TIR environments while maintaining a high tracking speed (20.2 frame/s), achieving effective and stable real-time tracking of TIR objects.

Key words machine vision; infrared image; target tracking; siamese neural network; attention mechanism

OCIS codes 150.0155; 150.1135; 110.3080

1 引 言

在计算机视觉处理技术中，目标跟踪是一项重要组成部分[1]，也是研究的基本任务之一，目标跟踪技术在当今军事、医学、导航和安防等领域都有着十分广泛的应用。红外目标跟踪作为目标跟踪技术中的一个分支，与彩色目标跟踪有着许多相似之处：两者都需要在图像中提取目标物体的外观特征，作为模板运用到后续帧的跟踪之中。红外目标跟踪相比于彩色目标跟踪有着明显的优势和劣势，优势在于：

收稿日期：2020-09-08；**修回日期**：2020-10-29；**录用日期**：2020-11-11

基金项目：河北省自然科学基金(F2017202009)、河北省创新能力提升计划(18961604H)

* **E-mail**：ydd12677@163.com

由于红外图像出自特殊的设备仪器，受外界的光照、噪声等影响较小，外观更加稳定，不会由于光照等因素产生突变等。缺点在于：红外图像采集设备无法达到彩色图像设备的清晰程度，采集数据存在分辨率低，目标不清晰，颜色和纹理特征不明显等特点[2]，使得红外目标跟踪算法和彩色目标跟踪算法不能很好地兼容，一些根据目标颜色特征进行跟踪的算法在红外数据集中无法很好地运用。

目前红外目标跟踪算法大多基于传统非深度学习，杨福才等[3]提出基于稀疏编码直方图特征和扰动感知模型的红外目标跟踪算法，利用红外目标的结构特性有效去除了背景干扰。赵东等[4]使用图像引导滤波和核相关滤波方法，能够有效区分背景边缘并对红外小目标进行跟踪。近年来，深度学习算法已经在目标跟踪领域中广泛应用，受此启发，人们尝试将卷积神经网络（CNN）应用到红外目标跟踪以提高算法性能。基于卷积神经网络的红外目标跟踪算法大体可分为两类，分别是基于深度特征的传统红外跟踪算法和基于匹配的端到端红外跟踪算法[5]。其中，基于深度特征的传统跟踪算法通常使用预先训练的神经网络进行深度特征提取，并将其与常规相关滤波跟踪框架进行结合。唐聪等[6]提出基于深度学习的红外与可见光决策级融合跟踪方法，用现有深度模型提取彩色目标模型，来对红外跟踪算法进行训练。DSST[7]算法采用判别相关滤波器确定目标位置信息，并提出了精确的尺度估计方法来确定目标尺度信息，但没有涉及表征能力更强的深度特征；DSST-TIR[8]使用基于分类的深度特征和相关滤波器进行热红外（TIR）跟踪，并表明深度特征可以表示比手工特征更好的信息。MCFTS[9]算法融合了 VGG-Net 神经网络[10]的多层特征，实现了特征提取上的优化，构建了一个整体 TIR 跟踪器，但是特征的优化大大降低了跟踪器的速度。ECO[11]算法在相关滤波的基础上，优化了滤波器和模板更新策略，并将深度特征和手工特征相结合，大幅提升了跟踪器性能；ECO-TIR[12]利用对抗生成网络（GAN）[13-14]得到了大量的全新红外数据集，并用该数据集训练了一个孪生网络，以提取红外目标的深度特征，将其与 ECO 结合进行跟踪。基于匹配的端到端红外跟踪算法则是一个整体的神经网络框架，输入原始图像数据直接得到预测结果。SiamFC[15]离线训练一个权值共享的卷积神经网络，通过模板匹配方法获得响应最大位置，从而进行目标跟踪。SiamFC-TRI[16]通过构建三重损失公式解决三重关系的任务从而学习到更具有判别性的深层特征，但是 Siam 系列算法仅采用第一帧作为模板，无法解决遮挡等问题。CFNet[17]将相关滤波（CF）模块嵌入孪生神经网络来解决 Siamese 系列的模板更新问题，同时整个网络能够进行端到端训练。RASNet[18]提出将三种注意力机制加入到孪生神经网络来适应在线匹配的模板。TADT[19]通过两个辅助任务来实现目标感知以进行在线跟踪。SiamDW[20]在孪生网络框架上设计了新的内部裁剪残差（CIR）单元构成更深更广的骨干网络框架，以获取更准确的跟踪结果。

针对上述情况，本文主要的目标是通过引入注意力等机制，使网络在特征提取方面得到优化，能够提取到更为稳定的特征，解决以前的算法容易受外界条件影响的问题。本文以 CFNet 算法为基础，该算法的深度学习框架和相关滤波模块保证了实时性能。将网络增加为两个并联的子网，其中一个子网对特征进行融合优化，并加入全局感知模块，使得提取到的特征具有空间、语义等更多的有效信息，进而提高算法的跟踪准确性；另一子网加入自注意力机制，来提升目标特征的判别能力，使算法能够区分目标和局部干扰。

2 基本原理

基于孪生网络的目标跟踪算法将跟踪转化为模板匹配问题，通过端到端的训练学习目标与模板之间的相似性，从而得到跟踪结果，CFNet 将相关滤波融入孪生网络，进一步提升了基于孪生网络的算法跟踪性能，同时保证了实时性。本文整体结构如图 1 所示，主要包括 CFNet 网络基础框架、特征融合模块、融合空间转换器与通道注意力的全局感知模块以及自注意力模块。算法共分为两个子网络，分别是特征融合与全局感知子网和自注意力子网，红外图像经过两个子网得到的响应图再通过平均的方法进行融合得到最终响应图。

2.1 CFNet 跟踪框架

孪生神经网络系列算法由两个权值共享的分支组成，将第一帧模板和当前帧的搜索区域通过该网络提取到的深度特征进行互相关得到响应图，即相似度得分，响应最高位置为目标位置，传统的全卷积孪生网络相似度函数为

$$f(\boldsymbol{z},\boldsymbol{x})=g[\varphi(\boldsymbol{z}),\varphi(\boldsymbol{x})], \tag{1}$$

式中：$\boldsymbol{z}$ 为第一帧模板图像；$\boldsymbol{x}$ 为当前帧的搜索图像；φ 为网络特征提取过程；g 表示互相关操作。考

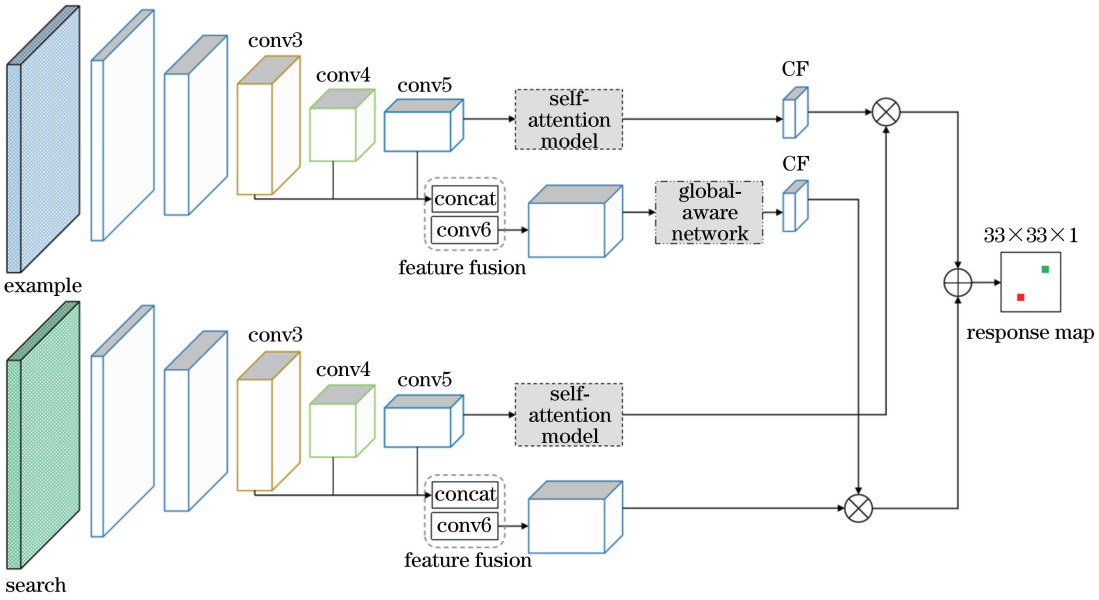

图 1　全局感知目标跟踪整体框架

Fig. 1　Global-aware object tracking framework

虑到计算效率，将相关滤波器(CF)加入模板支路后，得到新的相似度函数为

$$f(\boldsymbol{z},\boldsymbol{x})=g\{\sigma[\varphi(\boldsymbol{z})],\varphi(\boldsymbol{x})\},\tag{2}$$

式中：σ 为相关滤波模块，用于在线更新具有判别能力的模板。训练时使用 logistic 损失进行优化，

$$L(y,v)=\frac{1}{|\boldsymbol{D}|}\sum_{n\in\boldsymbol{D}}\lg\{1+\exp\{-y[u]v[u]\}\},\tag{3}$$

式中：$\boldsymbol{D}\in\mathbb{R}^{M\times M}$ 是由(2)式得到的相似度图；$v[u]$ 表示候选目标的实际数值；$y[u]$ 表示该目标的真实值。

2.2　多层特征融合

以往的孪生网络[15-20]都是以网络最后一层得到的特征作为输出特征，这足够表示 RGB 目标。由于深层特征感受野较大，具有判别性的语义特征信息，分辨率低，而浅层特征感受野较小，具有较强的空间结构信息，分辨率高，所以单独使用最后一层特征不足以完美表达红外目标。特征融合是针对特征图的优化操作，将多个卷积层的输出特征统一分辨率后级联，得到新的具有更多通道即更多图像表征信息的融合特征图[21]。在选择特征图进行融合时，考虑到前两层特征虽然包含丰富的空间信息，但尺寸较大，统一的过程会丢失过多的特征信息，而从第三层开始，使用后三层特征进行融合既解决了信息丢失过多的问题，又能够包含丰富的空间信息和语义信息。因此，本文算法将 CFNet 特征提取网络的后三层特征进行融合操作，卷积特征融合过程通过实验可视化如图 2 所示，浅层信息用于定位目标位置，深层信息用于区分不同对象。为了将具有不同分辨率的多层特征有效地融合在一起，对浅层特征进行最大池化以得到与深层特征同样的分辨率，并通过归一化来平衡三层特征的影响后，连接得到融合特征图，考虑到该特征图维度过大，在连接之后，使用 1×1 的卷积对融合特征图降维来减少训练时间，如图 1 中 conv6 所示。可将融合过程表示为

$$f_{\text{fusion}}=\text{concat}\{\text{bn}[\text{mp}(f_{\text{conv3}})],\text{bn}[\text{mp}(f_{\text{conv4}})],\text{bn}(f_{\text{conv5}})\},\tag{4}$$

$$f_{\text{final}}=\text{conv}(f_{\text{fusion}}),\tag{5}$$

式中：f_{conv3}、f_{conv4}、f_{conv5}、f_{fusion} 和 f_{final} 表示对应卷积层的输出特征、融合后的特征以及降维后的最终特征；mp 表示最大池化操作；bn 表示批归一化；concat 表示特征连接，进行特征串联操作，该步骤得到的融合特征图 f_{fusion} 大小为 49×49×832；conv 表示 1×1 的卷积，得到的最终特征图 f_{final} 大小为 49×49×256。

2.3　全局感知模块

在通过(4)式和(5)式得到融合的多层特征后，它已经同时包含了空间信息和语义信息，在这个基础上，我们希望网络模型能够具有全局感知能力，即能够应对目标的形变、旋转等变化，同时对特征通道分配相应权重以获取每个通道的重要程度，使算法对空间和通道信息均表现出鲁棒性。针对该问题，本文建立了全局感知模块，如图 3 所示。

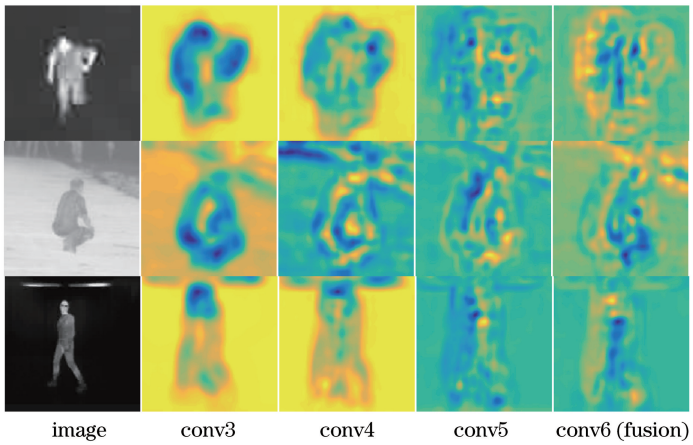

图 2　特征融合的可视化结果

Fig. 2　Visualization of feature fusion

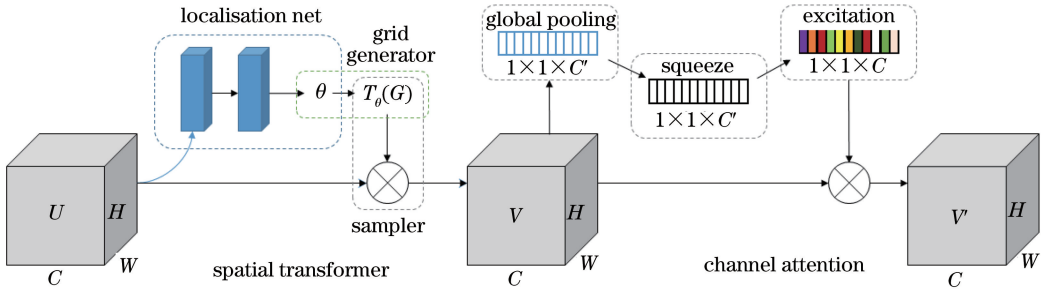

图 3　拟议的全局感知模块结构

Fig. 3　Architecture of proposed global-aware network model

2.3.1　空间转换器

作为一个可以插入到卷积神经网络中的独立模块，空间转换器[22]以特征图作为输入，对其进行仿射变换后以原尺寸输出，达到优化特征图的目的。该模块包含三个子模块，分别是定位网络、网格生成器和采样器，如图 3 中 spatial transformer 部分所示。首先由定位网络接收输入特征图 $\boldsymbol{U}$，经过隐藏层后得到变化参数 $\boldsymbol{\theta}$，隐藏层包含三个卷积层和一个全连接层。得到的 $\boldsymbol{\theta}$ 是一个六维变量，映射输入和输出之间的变换关系。在得到 $\boldsymbol{\theta}$ 后，进一步做矩阵运算，以目标特征图 $\boldsymbol{V}$ 中像素点坐标为自变量，得到原特征图 $\boldsymbol{U}$ 的坐标，从而得到 $\boldsymbol{V}$ 中该点的像素值。该过程表示为

$$\begin{bmatrix} x^s \\ y^s \end{bmatrix} = \boldsymbol{T}_{\boldsymbol{\theta}}(G) = \boldsymbol{A}_{\boldsymbol{\theta}} \begin{bmatrix} x^t \\ y^t \\ 1 \end{bmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{bmatrix} x^t \\ y^t \\ 1 \end{bmatrix}, \tag{6}$$

式中：(x^t, y^t) 为输出特征图 $\boldsymbol{V}$ 中像素坐标；(x^s, y^s) 为输入特征图 $\boldsymbol{U}$ 中对应映射的像素坐标；$\boldsymbol{A}_{\boldsymbol{\theta}}$ 是仿射变换矩阵。在计算出输出特征 $\boldsymbol{V}$ 中对应输入特征 $\boldsymbol{U}$ 的坐标位置后，可以根据 $\boldsymbol{U}$ 中像素值对 $\boldsymbol{V}$ 进行填充。但此时通过计算得到的坐标存在小数的情况，采用双线性插值来代替简单的取整操作，$\boldsymbol{V}$ 中像素点填充过程表示为

$$V_i = \sum_n^H \sum_m^W U_{nm} * \max(0, 1 - |x_i^s - m|) * \max(0, 1 - |y_i^s - n|), \tag{7}$$

式中：V_i 为输出特征图的像素值；U_{nm} 为输入特征图 (n, m) 处的像素值。可以发现，V_i 由该点上下左右四个点的像素值决定。

2.3.2　通道注意力

由于特征图具有多个通道，每个通道包含的信息不尽相同，单纯地平均各通道的权重并不能更好地提取有效特征[23]。为了确定在跟踪任务中提取特征图通道维度上的重要信息，本文在跟踪框架中加入通道注意力网络[24]来对特征通道进行加权。如图 3 中的 channel attention 部分所示，该模块串联在空间转换器之后，包含压缩和激励两个部分。压缩部分以空间转换器得到的尺寸为 $H \times W \times C$ 的特征图 $\boldsymbol{V}$ 作为输入，将其压缩至 $1 \times 1 \times C$，使其获得原特征图全局感受野，这步操作由全局平均池

化实现，

$$z=\frac{1}{H\times W}\sum_{i}^{H}\sum_{j}^{W}v_{ij}, \tag{8}$$

式中：z 为池化后的特征；V_{ij} 为输入特征在(i,j)处的值。在后续的激励部分，通过全连接(FC)层和Sigmoid层对每个通道进行加权后与原始特征图相乘得到最终结果，这一步骤可以表示为

$$\phi=\sigma[W_2\delta(W_1z)], \tag{9}$$

$$\boldsymbol{V}'=\mathrm{scale}(\boldsymbol{V},\phi), \tag{10}$$

式中：W_1 和 W_2 是两层全连接层，分别负责特征的降维和升维；δ 表示 ReLU 激活函数；σ 表示 Sigmoid 激活函数层；ϕ 为 Sigmoid 层的输出，即特征通道权重，最后将 ϕ 和输入特征 $\boldsymbol{V}$ 进行 scale 操作，即对特征通道重新加权，得到最终特征图 $\boldsymbol{V}'$。

2.4 自注意力机制

红外目标与彩色目标相比，缺少颜色、光照等信息，导致跟踪器难以分辨相似物体的差异。为了提取跟踪目标的判别信息，增强对特征的判别能力，更好地区分图像中的相似目标，本文使用基于自注意力机制[25]的编-解码网络结构，如图 4 所示，该模块的输入为特征提取网络第五层输出特征，此时的特征并不具有判别相似目标的能力，于是加入自注意力模块，专注于让跟踪器学习到目标与相似物体之间的差异，更准确地进行红外目标跟踪。该结构包含两个较大卷积层，用于识别判别区域，得到每个区域的重要程度，然后通过两个反卷积层对该区域进行定位，之后通过 Sigmoid 激活函数对上层输出进行加权，使其专注于目标区域。对特征图加权的方法表示为

$$\varphi(\boldsymbol{X})=\boldsymbol{X}+\mathrm{scale}\{\sigma[W_{\mathrm{C}}\delta(W_{\mathrm{D}}\boldsymbol{X})]\}, \tag{11}$$

式中：$\boldsymbol{X}$ 为输入特征图；$\varphi(\boldsymbol{X})$为输出的具有信息判别能力的特征图；W_{C} 和 W_{D} 分别表示编码层与解码层。

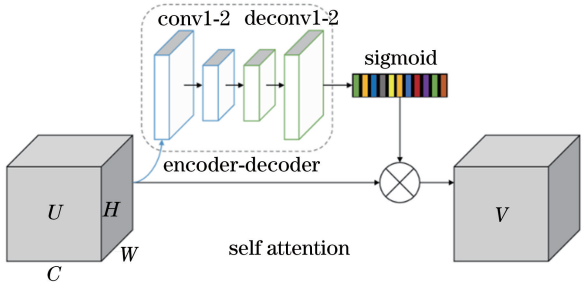

图 4 自注意力机制中的编解码模块

Fig. 4 Encoding-decoding module in self-attention mechanism

3 实验分析

3.1 训练数据和评价标准

本实验硬件环境为 12 核 Inter Core i7 CPU，单块 GTX 1080Ti 显卡，操作系统为 Ubuntu，编程环境采用 MatlabR2018b 结合 Matconvnet 工具箱，训练数据使用热红外视觉训练数据集[26]，该训练集包含 30 个类别的 1100 个视频序列，有超过 450k 张图片和 530k 个边界框，涵盖广泛的拍摄设备和场景，保证了多样性。训练时使用 AlexNet 五层卷积层作为基础特征提取网络，使用随机梯度下降法对网络进行优化训练，批尺寸 batchsize 为 8，动量 momentum 为 0.9，训练周期 epoch 为 70，学习率在整个周期内从 10^{-2} 到 10^{-4} 呈指数衰减。

3.2 通过 PTB-TIR 目标跟踪测试基准评估算法性能

为验证本文算法的有效性，将其与另外 10 种红外目标跟踪方向的流行算法(DSST[7]、UDT[27]、CREST[28]、MLSSNet[26]、MCFTS[9]、SiamFC[15]、HSSNet[29]、SiamFC-tri[16]、KCF[30]以及 baseline 算法 CFNet[17])进行对比，实验的评价标准为 PTB-TIR 数据集[31]，其中包含 60 个序列，具有 9 种不同的挑战：形变(DEF)、遮挡(OCC)、尺度变化(SV)、背景杂波(BC)、低分辨率(LR)、快速运动(FM)、运动模糊(MB)、移出视野(OV)和热交叉(TC)。PTB-TIR 测试基准使用精确度(Pre)和成功率(Suc)来评估跟踪器的整体性能。精确度是根据测量中心位置误差在规定阈值内的帧数的百分比值得出，这里阈值默认为 20 pixel，中心位置误差 E_{CL} 的计算方式可用两点的欧氏距离表示为

$$E_{\mathrm{CL}}=\sqrt{(x-x_{\mathrm{g}})^2+(y-y_{\mathrm{g}})^2}, \tag{12}$$

式中：(x,y)表示预测框的中心位置；$(x_{\mathrm{g}},y_{\mathrm{g}})$表示该帧真实值的中心位置。成功率则是根据测量其重叠率大于规定阈值的帧数的百分比值得出，这里阈值默认为 0.5，重叠率(S_{O})的计算方式可表示为

$$S_{\mathrm{O}}=\frac{|R\cap R_{\mathrm{g}}|}{|R\cup R_{\mathrm{g}}|}, \tag{13}$$

式中：R 表示跟踪得到的预测目标区域；R_{g} 表示目标的真实区域。

3.2.1 消融实验

为了验证两个子网络对算法产生的影响，设计了消融实验如表 1 所示。表中 GA 表示全局感知子网络，ST 表示自注意力子网络，可以看出，在单独加入全局感知子网络后，精确度和成功率分别提升了

5.9%和 8.6%，在单独加入自注意力子网后，精确度和成功率分别提升了 4.9%和 7.5%，两个子网同时加入后，精确度和成功率分别提升了 11.2%和 8.6%。说明两个子网对算法都有较大贡献。在响应融合后，得到了更高的分数，说明两个子网拥有相互促进的作用。

3.2.2　定量分析

图 5 为本文算法与其他算法的总体精确度和成功率图，可以看出，本文算法总精确度为 0.735，较第二名算法 MLSSNet 高出 0.8%，较基准算法 CFNet 高出 11.2%，本文算法总成功率为 0.530，较第二名算法 DSST 高出 0.4%，较基准算法 CFNet 高出 8.6%，证明本文算法在 CFNet 基础上，提取了更好的深度特征，通过加入注意力和感知模块，进一步优化了提取到的特征，更适应红外跟踪算法，提升了算法的整体性能。

表 1　在 PTB-TIR 数据集上消融实验结果对比

Table 1　Comparison of results of ablation experiments on PTB-TIR dataset

Tracker	PTB-TIR	
	Pre ↑	Suc ↑
CFNet	0.623	0.444
CFNet+GA	0.682	0.493
CFNet+ST	0.709	0.519
CFNet+GA+ST(proposed)	**0.735**	**0.530**

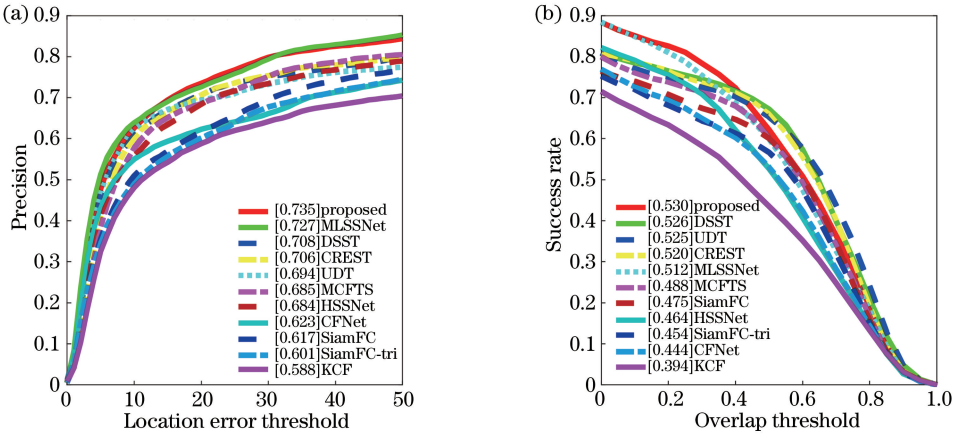

图 5　不同算法的总体精确度和成功率。(a)精确度；(b)成功率

Fig. 5　Overall accuracy and success rate of different algorithms. (a) Precision; (b) success rate

为了使实验结果更全面，利用 11 种算法针对不同的视频属性进行对比，图 6 为所有算法精确度定量对比结果。本文算法在形变、尺度变化、背景杂波、低分辨率四种属性中排名第一，在快速运动、运动模糊、移除视野三种属性中排名前三。

图 7 为所有算法成功率的定量对比结果。本文算法在形变、背景杂波、低分辨率三种属性中排名第一，在遮挡、尺度变化、快速运动、运动模糊、移除视野五种属性中排名前三。

为了验证算法的实时性能，利用 PTB-TIR 红外跟踪评估基准来记录算法测试时的平均速度，11 种算法的平均速度对比结果如表 2 所示，算法包含四种类型，分别是使用手工特征的相关滤波算法、使用深度特征的相关滤波算法、其他类型的深度学习算法和基于模板匹配的深度学习算法，其中相关滤波类算法因其不具有深度结构，获得了较高的速度，但相应的精确度和成功率略低于深度学习算法。本文算法测试速度达到 20.2 frame/s，基本满足实时性的需求。

3.2.3　定性分析

为了更直观地评估算法性能，在四类算法中各选取一个具有代表性的算法：DSST、MCFTS、CREST、MLSSNet，以及本文算法及其基准 CFNet 共 6 个算法，对各算法在部分具有多个跟踪属性的视频序列上的表现进行定性分析，直观跟踪效果如图 8 所示，每个测试的视频序列都包含 3～5 个视频属性。

1）平面外旋转情况。在 airplane 序列第 145 帧中，登机乘客在拐角处发生平面外旋转，此时基于模板匹配的跟踪算法(CFNet、MCFTS、MLSSNet)均发生不同程度的漂移，本文算法因加入了空间转换网络，对目标的旋转平移等变化表现出鲁棒性，所以能够精确跟踪目标。

2）低分辨率情况。在 campus 和 conversation 序列中，因拍摄距离等问题，导致图像分辨率和清晰度较低，此时跟踪算法难以精准预测目标位置，MCFTS、CREST 算法因仅使用深度特征但缺乏模板匹配策略而导致跟踪失败。本文算法由于采用基于模板匹配的端到端深度框架，并加入了特征融合

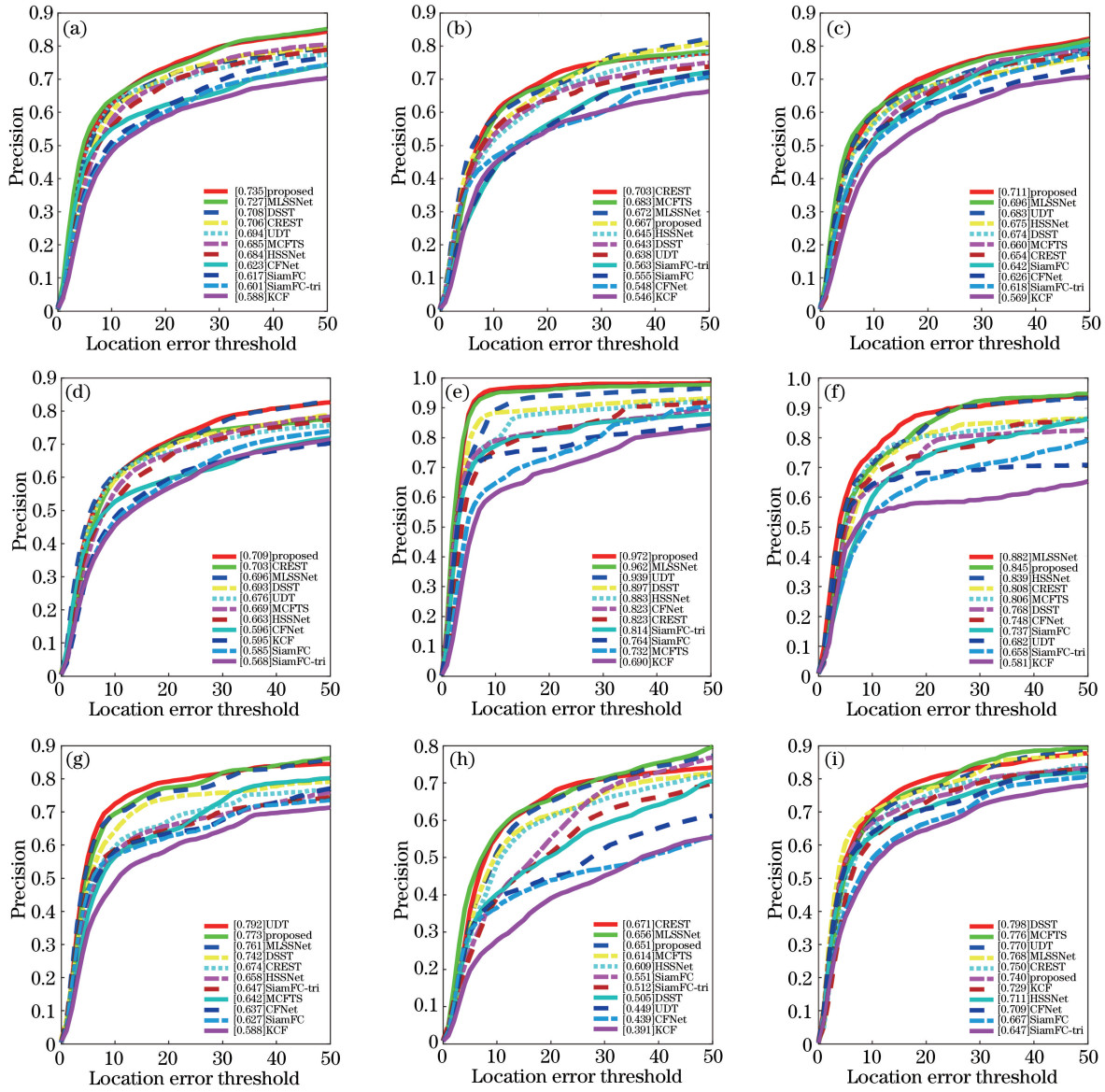

图 6　PTB-TIR 数据集中不同视频属性跟踪精确度。(a)形变;(b)遮挡;(c)尺度变化;(d)背景杂波;(e)低分辨率;(f)快速运动;(g)运动模糊;(h)移出视野;(i)热交叉

Fig. 6　Precision of different attributes videos on PTB-TIR dataset. (a) Deformation; (b) occlusion; (c) scale variation; (d) background clutter; (e) low resolution; (f) fast motion; (g) motion blur; (h) out of view; (i) thermal crossover

表 2　在 PTB-TIR 数据集上 11 种跟踪器的实验结果对比

Table 2　Comparison of experimental results of 11 trackers on PTB-TIR dataset

Category	Tracker	PTB-TIR		Speed / (frame·s^{-1})
		Pre ↑	Suc ↑	
Hand-crafted Feature based CF	DSST	0.708	0.526	45.4
	KCF	0.588	0.394	301.3
Deep feature based CF	MCFTS	0.685	0.488	4.8
Other deep tracker	CREST	0.706	0.520	0.7
Matching based deep tracker	SiamFC	0.617	0.475	66.7
	CFNet	0.623	0.444	37.0
	SiamFC-tri	0.601	0.454	60.0
	UDT	0.694	0.525	82.8
	HSSNet	0.684	0.464	18.0
	MLSSNet	0.727	0.512	19.8
	Proposed	0.735	0.530	20.2

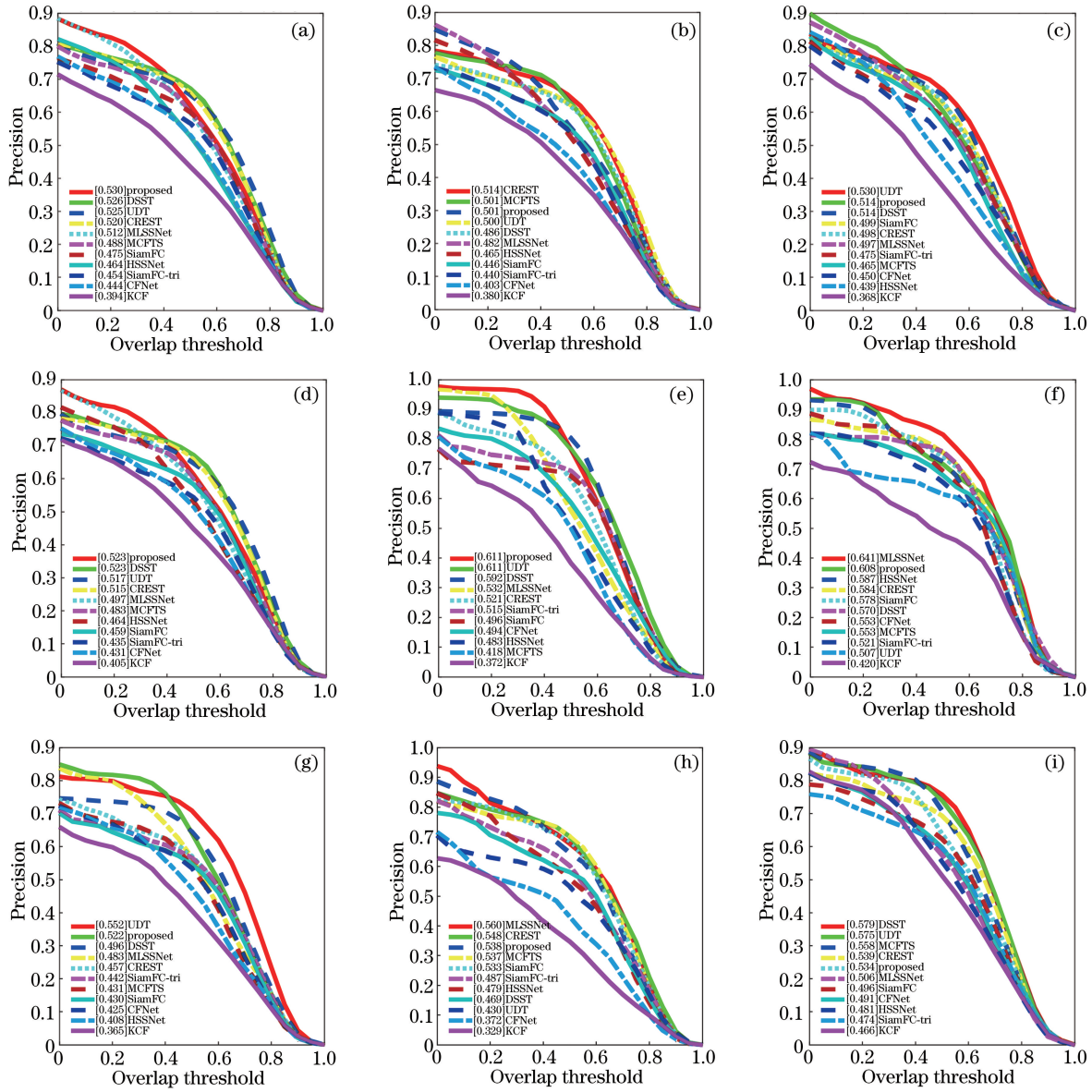

图 7 PTB-TIR 数据集中不同视频属性跟踪成功率。(a)形变;(b)遮挡;(c)尺度变化;(d)背景杂波;(e)低分辨率;(f)快速运动;(g)运动模糊;(h)移出视野;(i)热交叉

Fig. 7 Success rates of different attributes videos on PTB-TIR dataset. (a) Deformation; (b) occlusion; (c) scale variation; (d) background clutter; (e) low resolution; (f) fast motion; (g) motion blur; (h) out of view; (i) thermal crossover

方法,能够提取包含更多信息的目标特征,所以在分辨率较低的情况下依然能够对小目标进行准确跟踪。

3) 背景杂乱情况。由于背景中包含多个与目标相似的人物,可能导致跟踪过程出现背景杂乱。在 conversation 序列第 100 帧和 meetion 序列第 196 帧中,跟踪目标周围出现背景干扰项,常规模板算法缺少模板更新策略,导致 CREST、MLSSNet 等算法出现了不同程度的跟踪漂移,预测结果成为了旁边的干扰目标,而本文算法由于加入了 CF 层,可以实时进行模板更新,与此同时,全局感知模块能够得到目标的空间位置以及语义特征信息,所以能够准确预测目标位置并与背景中的干扰项区分开来。

4) 遮挡情况。在 road 序列第 615 帧和 crowd 序列第 122 帧中,目标在正常运动过程中发生遮挡,road 序列遮挡情况轻微,此时 CFNet 算法跟踪结果完全漂移;crowd 序列发生完全遮挡,此时 CFNet、MCFTS 和 CREST 算法完全漂移,说明此类算法特征的表达能力较弱,面对遮挡这种突变情况的应对能力差,本文算法加入了通道注意力和自注意力机制,能够提取判别能力强的特征,并且对特征的变化具有较强的鲁棒性,能够预测目标的整体位置和尺

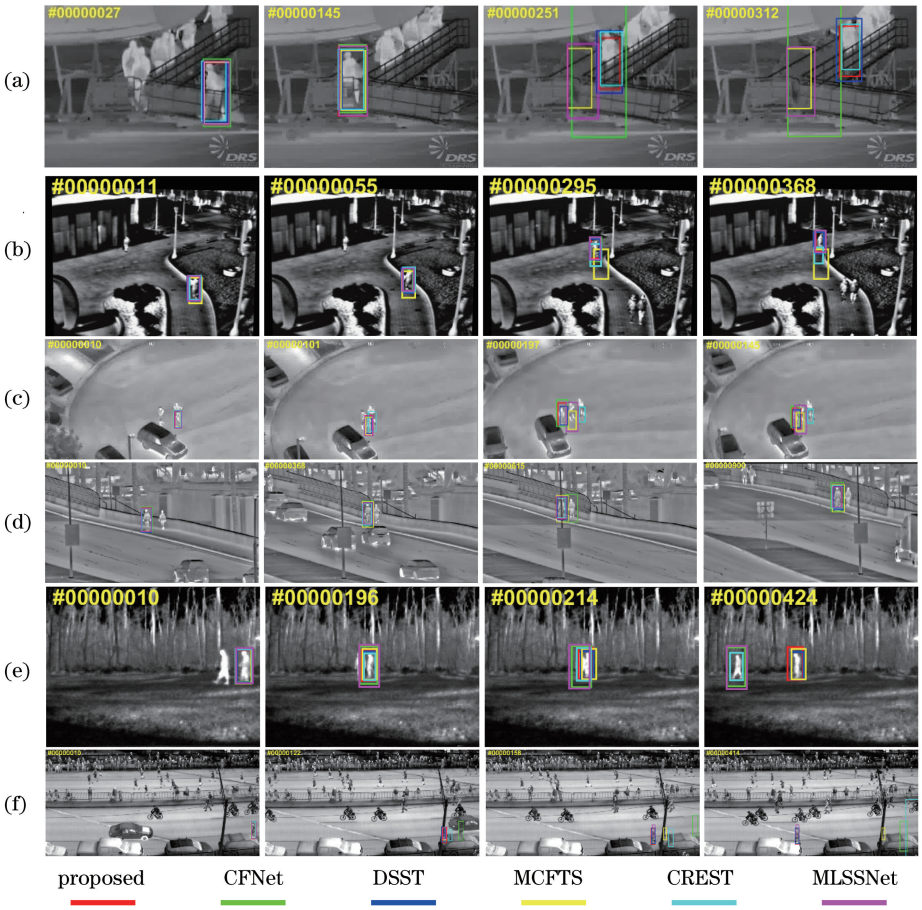

图 8 六种算法的实际跟踪效果图。(a) Airplane;(b) campus;(c) conversation;(d) road;(e) meetion;(f) crowd

Fig. 8 Actual tracking results of six algorithms. (a) Airplane; (b) campus; (c) conversation; (d) road; (e) meetion; (f) crowd

寸,进行遮挡情况下的有效跟踪。

4 结 论

本文提出了基于全局感知的孪生网络红外目标跟踪算法,以基于相关滤波的端到端跟踪网络为框架,对红外目标进行特征层次上的优化。作为优化的初步阶段,特征融合保证了其信息的完整性,融合后的特征同时具有浅层的位置信息和深层的语义信息;融入空间转换和通道注意力机制,为特征提取赋予了全局的信息感知能力;加入自注意力机制,使特征拥有判别性信息,优化类内差异,让跟踪器能够对目标更加鲁棒。在 PTB-TIR 评价基准上的实验结果表明,本文算法较好地解决了红外环境下的跟踪问题,能够适应目标形变、背景杂乱和遮挡等问题,提高了算法对红外目标的跟踪成功率和准确率,具有一定的研究价值。

参 考 文 献

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

[2] He Y J, Li M, Zhang J L, et al. Infrared target tracking via weighted correlation filter [J]. Infrared Physics & Technology, 2015, 73: 103-114.

[3] Yang F C, Yang D D, Mao N, et al. Robust infrared target tracking based on histograms of sparse coding [J]. Acta Optica Sinica, 2017, 37(11): 1115002.
杨福才, 杨德东, 毛宁, 等. 基于稀疏编码直方图的稳健红外目标跟踪[J]. 光学学报, 2017, 37(11): 1115002.

[4] Zhao D, Zhou H X, Qin H L, et al. Infrared dim-small target tracking based on guided image filtering and kernelized correlation filtering [J]. Acta Optica Sinica, 2018, 38(2): 0204004.
赵东, 周慧鑫, 秦翰林, 等. 基于引导滤波和核相关滤波的红外弱小目标跟踪[J]. 光学学报, 2018, 38(2): 0204004.

[5] Liu Q, Li X, He Z Y, et al. Multi-task driven feature models for thermal infrared tracking [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 11604-11611.

[6] Tang C, Ling Y S, Yang H, et al. Decision-level fusion tracking for infrared and visible spectra based on deep learning [J]. Laser & Optoelectronics Progress, 2019, 56(7): 071502.
唐聪, 凌永顺, 杨华, 等. 基于深度学习的红外与可见光决策级融合跟踪[J]. 激光与光电子学进展, 2019, 56(7): 071502.

[7] Danelljan M, Häger G, Shahbaz Khan F, et al. Accurate scale estimation for robust visual tracking [C] // Proceedings of the British Machine Vision Conference 2014, Nottingham. British Machine Vision Association, 2014: 1-11.

[8] Gundogdu E, Koc A, Solmaz B, et al. Evaluation of feature channels for correlation-filter-based visual object tracking in infrared spectrum[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), June 26 - July 1, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 290-298.

[9] Liu Q, Lu X H, He Z Y, et al. Deep convolutional neural networks for thermal infrared object tracking [J]. Knowledge-Based Systems, 2017, 134: 189-198.

[10] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10)[2020-09-01]. https://arxiv.org/abs/1409.1556.

[11] Danelljan M, Bhat G, Khan F S, et al. ECO: efficient convolution operators for tracking[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6931-6939.

[12] Zhang L C, Gonzalez-Garcia A, van de Weijer J, et al. Synthetic data generation for end-to-end thermal infrared tracking [J]. IEEE Transactions on Image Processing, 2019, 28(4): 1837-1850.

[13] Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 2242-2251.

[14] Isola P, Zhu J Y, Zhou T H, et al. Image-to-image translation with conditional adversarial networks [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 5967-5976.

[15] Bertinetto L, Valmadre J, Henriques J F, et al. Fully-convolutional siamese networks for object tracking [M]. Cham: Springer International Publishing, 2016: 850-865.

[16] Dong X P, Shen J B. Triplet loss in Siamese network for object tracking [M]. Cham: Springer International Publishing, 2018: 472-488.

[17] Valmadre J, Bertinetto L, Henriques J, et al. End-to-end representation learning for correlation filter based tracking [C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 5000-5008.

[18] Wang Q, Teng Z, Xing J L, et al. Learning attentions: residual attentional Siamese network for high performance online visual 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: 4854-4863.

[19] Li X, Ma C, Wu B Y, et al. Target-aware deep tracking [C] // 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 1369-1378.

[20] Zhang Z P, Peng H W. Deeper and wider Siamese networks for real-time visual tracking [C] // 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 4586-4595.

[21] Bell S, Zitnick C L, Bala K, et al. Inside-outside net: detecting objects in context with skip pooling and recurrent neural networks[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 2874-2883.

[22] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial transformer networks [C] // NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2. 2015: 2017-2025.

[23] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module [M]. Cham: Springer International Publishing, 2018: 3-19.

[24] Hu J, Shen L, Albanie S, et al. Squeeze-and-excitation networks [C] // IEEE Transactions on Pattern Analysis and Machine Intelligence, New York: IEEE Press: 2011-2023.

[25] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems, 2017: 5998-6008.

[26] Liu Q, Li X, He Z Y, et al. Learning deep multi-

level similarity for thermal infrared object tracking [J]. IEEE Transactions on Multimedia, 2020, (99): 1.

[27] Wang N, Song Y B, Ma C, et al. Unsupervised deep tracking [C] // 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 1308-1317.

[28] Song Y B, Ma C, Gong L J, et al. CREST: convolutional residual learning for visual tracking [C]// 2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 2574-2583.

[29] Li X, Liu Q, Fan N N, et al. Hierarchical spatial-aware siamese network for thermal infrared object tracking[J]. Knowledge-Based Systems, 2019, 166: 71-81.

[30] Henriques J F, Caseiro R, Martins P, et al. High-speed tracking with kernelized correlation filters[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(3): 583-596.

[31] Liu Q, He Z Y, Li X, et al. PTB-TIR: a thermal infrared pedestrian tracking benchmark [J]. IEEE Transactions on Multimedia, 2020, 22(3): 666-675.