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

结合注意力与特征融合网络调制的视觉跟踪算法

许克应，束平，鲍华*

安徽大学电气工程与自动化学院，安徽 合肥 230601

摘要 针对现有网络调制类的跟踪算法忽略高阶特征信息从而在应对大尺度变化、物体形变时易发生漂移的现象，提出了一种结合注意力与特征融合网络调制的目标跟踪算法。首先，将一个高效的选择核注意力模块嵌入在特征提取的主干网络中，使网络更关注于对目标特征信息的提取；其次，对提取的特征采用多尺度交互网络充分挖掘层内多尺度信息，并且融合高阶特征信息来提升对目标的表征能力，以适应跟踪过程中复杂多变的环境；最后，通过金字塔调制网络引导测试分支学习最优交并比预测，实现对目标的精确估计。实验结果表明，在VOT2018、OTB100、GOT10k、TrackingNet和LaSOT视觉跟踪基准上，相比其他算法，所提算法在跟踪精度和成功率上展现了较强的竞争力。

关键词 图像处理；目标跟踪；特征提取；注意力机制；特征融合

中图分类号 TP491.4 **文献标志码** A **DOI**：10.3788/LOP202259.1210013

Visual Tracking Combining Attention and Feature Fusion Network Modulation

Xu Keying, Shu Ping, Bao Hua*

School of Electrical Engineering and Automation, Anhui University, Hefei 230601, *Anhui, China*

Abstract The existing tracking algorithms for network modulation ignore high order feature information, so they are prone to drift when dealing with large scale changes and object deformations. An object tracking algorithm that combines the attention mechanism and feature fusion network modulation is proposed. First, an efficient selective kernel attention module is embedded in the feature extraction backbone network, so that the network pays more attention to the extraction of target feature information; second, a multiscale interactive network is used for the extracted features to fully mine the multiscale information in the layer, and high order feature information is fused to improve the ability of target representation, to adapt to the complex and changeable environment in the tracking process; finally, the pyramid modulation network is used to guide the test branch to learn the optimal intersection over union prediction to achieve an accurate estimation of the targets. Experimental results show that the proposed algorithm achieves more competitive results than other algorithms in tracking accuracy and success rate on VOT2018, OTB100, GOT10k, TrackingNet, and LaSOT visual tracking benchmarks.

Key words image processing; object tracking; feature extraction; attention mechanism; feature fusion

收稿日期：2021-05-08；**修回日期**：2021-06-07；**录用日期**：2021-06-27

基金项目：安徽省自然科学基金(1908085MF217)、安徽省教育厅自然科学重点资助项目(KJ2019A0022)

通信作者：*baohua@ahu.edu.cn

1 引 言

视觉目标跟踪是计算机视觉基本任务之一，在医学诊断、无人驾驶、视频监控、机器人传感等场景中得到了广泛应用[1]。但是由于尺度变化、光照变化、背景干扰等复杂因素，仍然难以开发出快速、准确、鲁棒的跟踪器。因此，提出一种准确且具有较强鲁棒性的目标跟踪算法将具有重要意义。

随着计算机算力的不断提高和相关数据集的建立，基于深度学习的卷积神经网络（CNN）被引入到目标跟踪领域，由于强大的表征能力，吸引了大量国内外学者的不断研究和探索。Bertinetto 等[2]提出的 SiamFC 算法首次将孪生网络（Siamese）框架用于跟踪，利用相同网络结构和参数提取模板特征与候选特征，通过计算两者之间的相似性来估计目标位置，取得了很好的跟踪性能。Li 等[3]基于 SiamFC 提出的 SiamRPN 算法引入了 Faster R-CNN[4]中的区域建议网络（RPN），对网络进行联合分类与回归训练，通过分类部分判别候选框是否包含目标，并利用回归部分预测候选框的偏移量，输出更加准确的目标框。Zhang 等[5]提出的 SiamDW 充分利用最先进的 Siamese 框架，探索了更宽、更深的网络结构，优化了特征提取过程。

准确估计目标尺度对衡量算法的性能至关重要，在孪生网络跟踪算法中，采用区域建议网络和多尺度金字塔搜索的方式在目标尺度估计中得到广泛应用，但存在尺度比例变化不灵活、搜索空间较大、搜索速度慢、计算复杂等问题；其次上述方法大多不考虑背景信息，极少采用在线训练和更新策略。为此，Danelljan 等[6]提出的 ATOM 算法使用类似 Jiang 等[7]提出的区域重叠度估计的策略，使用调制机制将目标特定的外观信息集成到测试网络中，通过离线训练方式最大化标注框与建议框之间的交并比（IoU），得到最优的边界框。同时 ATOM 还结合了相关滤波算法的优点，通过构造分类网络以在线训练的方式对目标前景与背景进行区分，实现对目标的粗略定位，此方法在多个视觉目标跟踪基准上表现出优异性能。然而，基于重叠度最大化的算法在特征提取网络中采用传统的卷积神经网络构建表观模型，不能有效地发挥强大的特征学习和特征表达能力，忽略了高层信息对目标建模的表征能力，在应对尺度变化时易发生漂移现象。因此本文提出了一种结合注意力与特征融合网络调制的目标跟踪算法，通过引入注意力模块关注显著区域并在不同尺度上处理与目标相关的信息，同时采用一种自适应的特征融合策略充分发挥高层特征与低层特征对目标表观的优势，使得所提算法在处理复杂场景时能够有更好的泛化能力。

本文工作主要体现在几个方面：1）提出了多尺度交互模块，通过在层内利用分组卷积与残差连接的方式产生多尺度的特征组合，在层间采用自适应融合的方式有效聚集了多尺度空间特征与语义特征，其次引入轻量级的注意力模块，聚焦于更相关的特征信息；2）使用金字塔调制模块生成参考目标特征表示，该模块将目标区域的特征集合到多个空间维度上形成特征金字塔，为 IoU 预测网络提供更精确的目标特征表示。

2 所提算法内容

所提算法主要包含在线目标分类（Online Classifier）和离线目标估计（Offline Estimator）两个部分，如图 1 所示。其中，目标估计部分根据大规模数据离线训练一个能准确预测目标框与候选框重叠率的估计器；目标分类部分利用初始帧的目标图像及其标签在线训练一个分类器，实现对目标的粗略定位。

2.1 目标估计与目标分类

目标估计部分由参考分支与测试分支构成，旨在给出粗略的目标位置后确定目标精确边界框。如图 1 所示，采用预训练的 ResNet-18 前四层卷积块作为特征提取网络来构建目标的表观模型。由于主干网络中不同层的卷积特征对目标信息的表征存在差异化，为了能够充分利用不同层的特征，结合多层特征融合的思想，利用多尺度交互（MSI）模块对多级特征进行增强与融合。随后，由金字塔调制模块（PMB）生成参考目标的表观特征，将其作为调制向量并对测试分支进行调制，使得测试帧携带目标特定的外观信息。通过准确的感兴趣池化（PrPool）[7]提取候选区域的特征，利用由全连接层构成的 IoU 预测器（IoU predictor）对候选区域进行评估。在模型训练过程中，采用均方损失函数最小化来优化模型参数，将跟踪器估计误差降至最低。损失函数可表示为

$$L_{\mathrm{IoU}}=\frac{1}{N}\sum_{i=1}^{N}(G_{\mathrm{IoU}}^{(i)}-P_{\mathrm{IoU}}^{(i)})^2, \tag{1}$$

式中：$P_{\mathrm{IoU}}^{(i)}$ 为第 i 个候选框预测得分；N 表示候选框总个数；$G_{\mathrm{IoU}}^{(i)}\in[-1,+1]$ 表示第 i 个候选框与目标

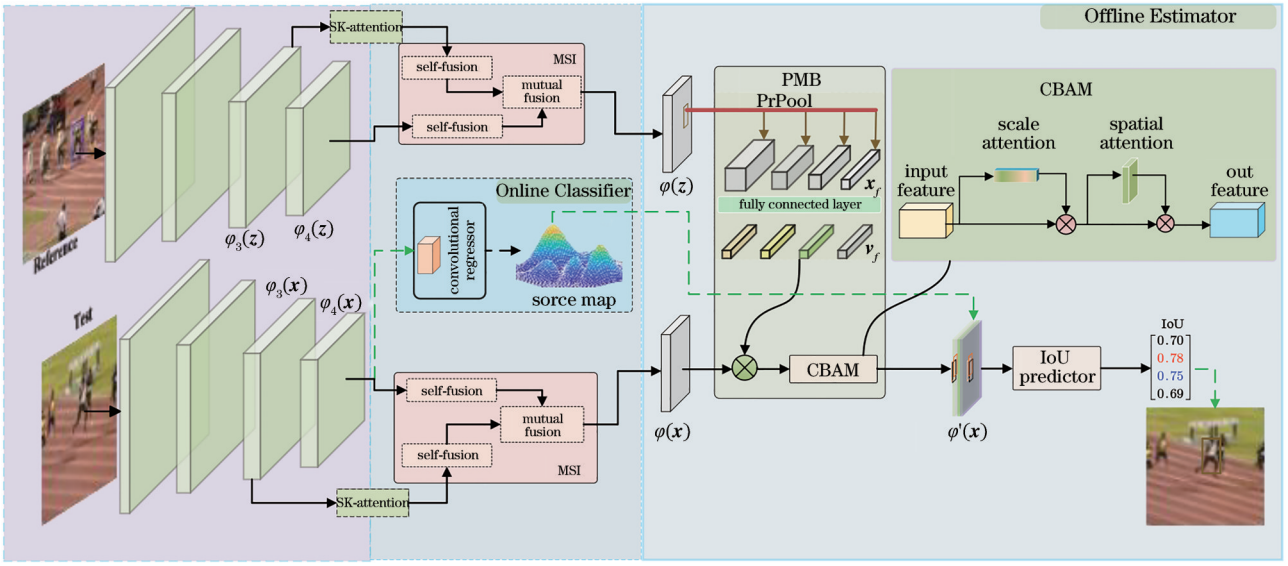

图 1　所提算法的框架

Fig. 1　Framework of proposed algorithm

框重叠率归一化后的值。

虽然目标估计部分能够提供精确的目标框输出，但缺乏对目标和背景干扰物的鲁棒判别能力。借鉴相关滤波器思想，利用分类网络来对目标与背景进行区分。与目标估计部分不同，目标分类部分是针对特定类别进行在线训练的，以预测目标置信分数，提供对象粗略的 2D 位置。目标分类部分由两个全连接层构成，定义为

$$f(\boldsymbol{x};\ \boldsymbol{\omega})=\varphi_2\left[\boldsymbol{\omega}_2*\varphi_1(\boldsymbol{\omega}_1*\boldsymbol{x})\right],\tag{2}$$

式中：$\boldsymbol{x}$ 为测试分支第三层输出特征映射；$\boldsymbol{\omega}_1$、$\boldsymbol{\omega}_2$ 为全连接层参数；φ_1、φ_2 为激活函数；$*$ 表示多通道卷积符号。采用牛顿-高斯[6]下降作为快速收敛的优化策略来优化两层网络参数，得到对目标与背景有高判别性的分类器，其中目标函数表达为

$$L(\boldsymbol{\omega})=\sum_{j=1}^{m}\gamma_j\left\| f\left(\boldsymbol{x}_j;\ \boldsymbol{\omega}\right)-\boldsymbol{y}_j\right\|^2+\sum_k\lambda_k\left\|\boldsymbol{\omega}_k\right\|^2,\tag{3}$$

式中：$\boldsymbol{x}_j$ 为训练样本的特征映射；$\boldsymbol{y}_j$ 为目标的标签信息；每个训练样本的影响程度由 γ_j 表示；$\boldsymbol{\omega}$ 表示分类器参数；$f\left(\boldsymbol{x}_j;\ \boldsymbol{\omega}\right)$ 为卷积操作；λ_k 和 $\boldsymbol{\omega}_k$ 分别表示正则化系数和权重参数。

在对后续视频序列的跟踪过程中，根据分类器得到输出响应最大的位置，在该位置依据前一帧目标框尺度大小产生一些高斯候选框；将其输入到目标估计部分预测候选框质量，选取得分值最高的候选框作为最终的跟踪结果。

2.2　选择核注意力模块

深度学习中的注意力类似于人类的视觉选择性注意力机制[8]，人类视觉皮层神经元感受野会根据刺激来进行调节，选择出对当前任务目标更有利的信息。引入选择核（SK）注意力[9]对主干网络第三层输出特征进行调整，由于其多分支结构为网络提供了动态的感受野，增强了网络对目标的整体感知。

如图 2 所示，SK 注意力模块可分为三个部分：分裂（Split）、融合（Fuse）和选择（Select）。假设输入特征为 $\boldsymbol{X}\in\mathbb{R}^{C\times h\times w}$，首先通过 3×3 和 5×5 的卷积核对其进行两次卷积操作，得到特征 $\tilde{\boldsymbol{U}}$ 和 $\hat{\boldsymbol{U}}$，对承载不同信息量的特征按元素对应位相加的方式进行融合。然后对于每个通道上的特征，使用全局平均池化（GAP）的方式压缩成一系列实数 $\boldsymbol{S}\in\mathbb{R}^c$，其中 $\boldsymbol{S}$ 的第 c 个元素可表示为

$$\boldsymbol{S}_c=f_{\mathrm{GAP}}\left(\boldsymbol{U}_c\right)=\frac{1}{h\times w}\sum_{a=1}^{h}\sum_{b=1}^{w}\boldsymbol{U}_c(a,b),\tag{4}$$

式中：$\boldsymbol{U}_c$ 为融合后特征 $\boldsymbol{U}$ 在第 c 个通道上的特征；$f_{\mathrm{GAP}}(\cdot)$ 表示全局平均池化操作。此外，通过一个简单的全连接层降低维数，得到一个更紧凑的特征 $\boldsymbol{z}\in\mathbb{R}^{d\times 1}$，对 $\boldsymbol{z}$ 执行两次矩阵变化后，输出矩阵 $\boldsymbol{a}$ 和其冗余矩阵 $\boldsymbol{b}$。在选择操作过程中，使用 $\boldsymbol{a}$ 与 $\boldsymbol{b}$ 分别对 $\tilde{\boldsymbol{U}}$ 和 $\hat{\boldsymbol{U}}$ 进行加权求和操作，得到最终输出特征：

$$\boldsymbol{V}=\boldsymbol{a}\cdot\tilde{\boldsymbol{U}}+\boldsymbol{b}\cdot\hat{\boldsymbol{U}},\tag{5}$$

式中：$\boldsymbol{V}\in\mathbb{R}^{C\times h\times w}$ 表示输出特征。

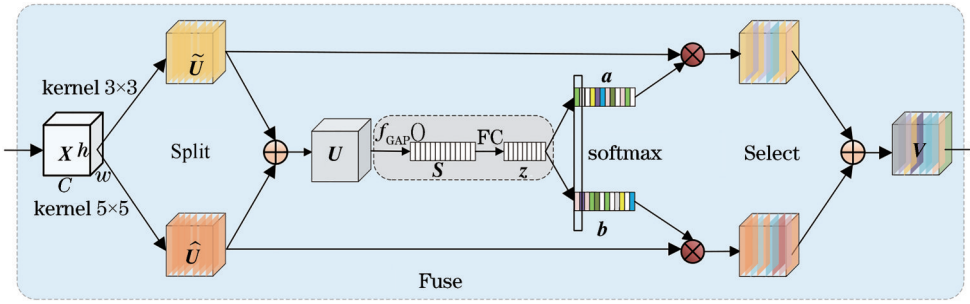

图 2　SK注意力网络结构

Fig. 2　Architecture of the SK attention network

2.3　多尺度交互模块

在深度跟踪网络的背景下，多尺度特征的融合表现出了惊人的性能[10-11]，受这些作品的启发，本文提出了一种融合多尺度、多层次特征信息的MSI模块，在多个尺度上利用学习到的特征编码全局和局部上下文信息。如图3所示，与大多数方法仅在层间进行特征多尺度表示不同，所提出的MSI还通过分组卷积的形式在层内构建特征的多尺度组合。

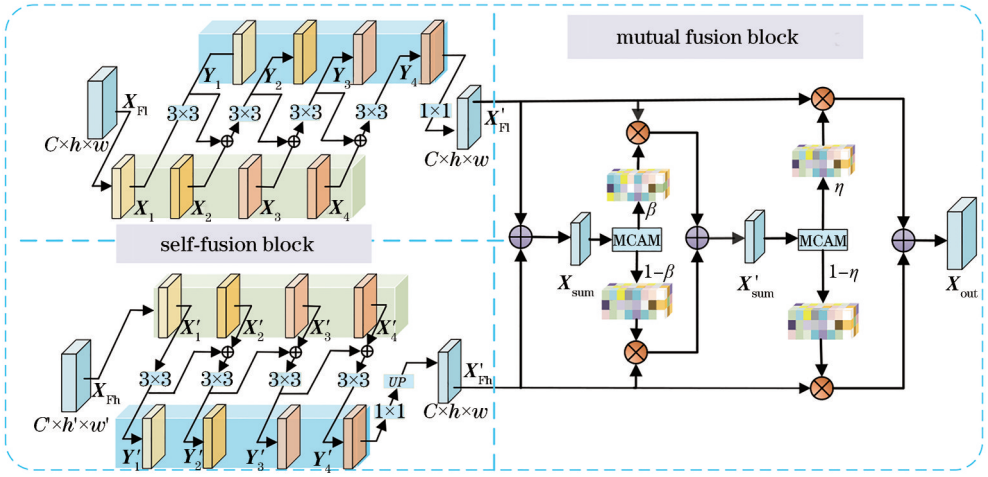

图 3　多尺度交互模块结构

Fig. 3　Architecture of the MSI module

2.3.1　互融合模块

在特征提取网络中对目标基本上下文信息进行提取时，需要对CNN模型不同深度上的卷积特征进行处理。如图4所示，低层特征通常包含目标丰富的纹理和轮廓信息，但语义信息较为模糊；而高层信息则是语义的、抽象的，并且随着网络的深化而有一定程度的信息丢失。因而提出互融合模块(MFB)，通过自适应学习各尺度特征图融合的通道权重，将不同层次的特征结合起来，同时兼顾语义和细节信息，根据网络本身的需要通过调整特征中不同抽象信息的比例来提高对特征的表达能力。

MFB网络结构如图3所示，对于给定的低级特征和高级特征，MFB的目标是对它们进行融合以获得更有用的信息。由于两个层次的特征具有不同的分辨率和不同的通道，在融合前，首先应用一个卷积层将特征的通道数压缩到相同的维数，并通过线性插值提高分辨率，得到对应调整后的低层特征与高层特征$\boldsymbol{X}'_{\mathrm{Fl}}$、$\boldsymbol{X}'_{\mathrm{Fh}}$，随后按对应元素求和得到$\boldsymbol{X}_{\mathrm{sum}}$。在$\boldsymbol{X}_{\mathrm{sum}}$中每个通道上都包含了更多的特征信息，但是整个通道存在着未融合且存在较多的冗余信息。因此，引入多尺度通道注意力模块(MCAM)进行两次通道权重的再分配，对多尺度特征上下文进行聚合。该模块以融合特征为输入，输出一组通道权重系数，通过张量乘法分别对$\boldsymbol{X}'_{\mathrm{Fl}}$、$\boldsymbol{X}'_{\mathrm{Fh}}$进行加权。两次权重分配操作可归纳为

$$\begin{cases}\boldsymbol{X}'_{\mathrm{sum}}=\beta\cdot\boldsymbol{X}'_{\mathrm{Fl}}+(1-\beta)\cdot\boldsymbol{X}'_{\mathrm{Fh}}\\ \boldsymbol{X}_{\mathrm{out}}=\eta\cdot\boldsymbol{X}'_{\mathrm{Fl}}+(1-\eta)\cdot\boldsymbol{X}'_{\mathrm{Fh}}\end{cases},\tag{6}$$

式中：β和η分别表示$\boldsymbol{X}_{\mathrm{sum}}$和$\boldsymbol{X}'_{\mathrm{sum}}$经MCAM输出的

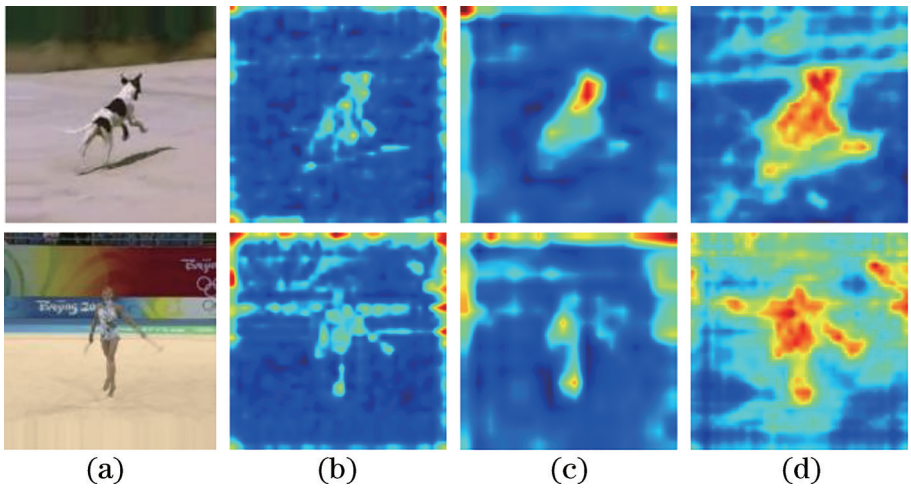

图 4　原始图像与相应层的特征可视化结果。(a)以目标为中心的原始图像;(b)低层特征;(c)高层特征;(d) MFB生成的融合特征

Fig. 4 Visualization results of originalimage and features from corresponding layer. (a) Original image centered on the target; (b) low level features; (c) high level features; (d) fused version generated by MFB

一组权重系数;$\boldsymbol{X}_{\text{out}}$表示最终的融合特征,它包含了低层特征与高层特征。为了直观理解,融合后的特征如图 4(d)所示,显然融合后的目标区域在空间与抽象信息方面与以前相比都更加精确。

2.3.2　自融合模块

MFB的目的是实现对层间卷积特征的有效利用,同时提出的自融合模块(SFB)则在层内通过分组卷积的方式产生不同感受野的特征组合,进一步地提高对不同尺寸对象的处理能力,SFB的细节如图 3所示。具体来说,将输入特征映射按通道方向划分为4组,每组特征具有相同的通道数;然后利用不同的卷积层对每组输入特征进行特征提取,特别地,除第一组外,其余每组输入特征都与前一组卷积层的输出相结合;最后将所有组的输出特征沿通道方向串联起来,得到融合特征。以特征$\boldsymbol{X}_{\text{FI}}\in\mathbb{R}^{C\times h\times w}$为例,整个过程可归纳为

$$\boldsymbol{Y}_d=\begin{cases}C_d(\boldsymbol{X}_d), d=1\\ C_d(\boldsymbol{X}_d+\boldsymbol{Y}_{d-1}), d=2,3,4\end{cases},\tag{7}$$

式中:$C_d(\cdot)$代表3×3卷积操作;$\boldsymbol{X}_d$为$\boldsymbol{X}_{\text{FI}}$分裂后对应的特征子集,$\boldsymbol{Y}_d$为对应输出特征。由于采用分组卷积的方式代替常规卷积,在不额外增加计算量的情况下,每个输出特征$\boldsymbol{Y}_d$都包含了不同数量、不同感受野大小的特征,有效提高了网络对不同尺度目标的感知能力。

2.3.3　多尺度通道注意力模块

在高层卷积特征中,每个通道能够当作对特定物体类别的响应,跟踪对象任意性使得部分通道在跟踪某些对象时发挥着重要作用,而其他通道则是可有可无的,因此对所有通道进行同等对待是不必要的。已有的方法通常采用全局池化获取重要通道的信息,然而这种方式倾向于强调全局分布的大对象,可能会削弱小目标存在的大部分信号。在全局上下文注意力[12]的启发下,设计了多尺度通道注意力模块,如图 5所示,其主要思想是从全局和局部

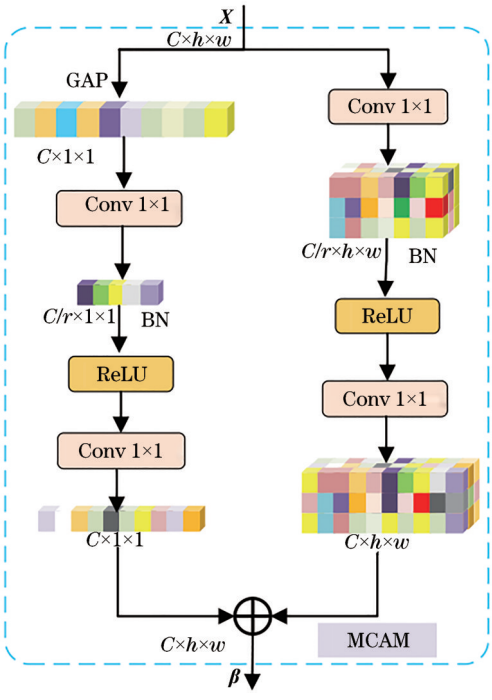

图 5　多尺度通道注意力网络结构

Fig. 5　Architecture of the multi-scale channel attention network

两个方面实现通道关注，以缓解尺度变化和小对象实例引起的问题。

假设输入特征映射为 $\boldsymbol{X}$，为了模拟通道之间的全局上下文关系，采用与压缩和激励网络[8]类似的操作获取全局通道上下文：

$$G(\boldsymbol{X})=\boldsymbol{W}_2\left\{\delta\left\{\mathbb{B}\left\{\boldsymbol{W}_1\left[\left(g(\boldsymbol{X})\right]\right]\right\}\right\}\right\},\tag{8}$$

式中：$\boldsymbol{W}_1\in\mathbb{R}^{C/r\times C}$ 表示降维卷积层；$\boldsymbol{W}_2\in\mathbb{R}^{C\times C/r}$ 代表增维卷积层；r 为通道缩减比；$\mathbb{B}$ 表示归一化；δ 指 ReLU 激活函数；全局平均池化 $g(\boldsymbol{X})=\frac{1}{h\times w}\sum_{a=1}^{h}\sum_{b=1}^{w}\boldsymbol{X}_{[:,a,b]}$。

对于局部上下文关系的获取，选择卷积核大小为 1×1 点向卷积作为局部通道上下文聚合器，在通道方向上对每个空间位置进行加权组合来突出特征中细节部分。同时为了减少参数，通过类似于 SENet 的瓶颈结构计算局部通道上下文：

$$L(\boldsymbol{X})=\boldsymbol{P}_1\left\{\delta\left\{\mathbb{B}\left[\boldsymbol{P}_2(\boldsymbol{X})\right]\right\}\right\},\tag{9}$$

式中：$\boldsymbol{P}_1$ 和 $\boldsymbol{P}_2$ 分别表示 $C/r\times C\times 1\times 1$、$C\times C/r\times 1\times 1$ 的卷积核。利用全局通道上下文 $G(\boldsymbol{X})$ 和局部通道上下文 $L(\boldsymbol{X})$，MCAM 输出可表示为

$$\boldsymbol{\beta}=\sigma\left[G(\boldsymbol{X})\oplus L(\boldsymbol{X})\right],\tag{10}$$

式中：$\boldsymbol{\beta}\in\mathbb{R}^{C\times h\times w}$；$\sigma$ 指 Sigmoid 激活函数；$\oplus$ 表示广播加法。

2.4 特征金字塔调制模块

特征金字塔调制模块将目标区域的特征集合到多个空间维度中形成特征金字塔，以调制的形式将目标不同尺度的外观信息注入到测试帧中，进一步增强对目标尺度的适应性。具体来说，即给定经 MSI 后的输入特征图 $\varphi(\boldsymbol{z})$、$\varphi(\boldsymbol{x})$ 和初始目标包围框坐标。金字塔调制模块在多个空间维度对模板图像特征映射 $\varphi(\boldsymbol{z})$ 进行 PrPool 操作，以获得目标特征表示 $\boldsymbol{z}_f\in\mathbb{R}^{C\times k_f\times k_f}$，并通过全连接层得到归一化后的调制矢量 $\boldsymbol{v}_f\in\mathbb{R}^{C\times 1\times 1}$，然后利用调制矢量 $\boldsymbol{v}_f$ 对测试特征映射 $\varphi(\boldsymbol{x})$ 进行调制。

$$\boldsymbol{x}_f=\varphi(\boldsymbol{x})\cdot\boldsymbol{v}_f,\tag{11}$$

式中：$\boldsymbol{x}_f\in\mathbb{R}^{C\times h\times w}$，沿着通道方向级联，形成具有不同参考外观信息的特征映射 $\left[\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n\right]=\boldsymbol{x}\in\mathbb{R}^{nC\times h\times w}$，同时引入全卷积注意力模块[13]对特征映射 $\boldsymbol{x}$ 进行精修。在该模块中，使用自适应的通道和空间注意力对特定的位置和目标尺度进行关注，以突出目标主体部分。对于精修后的特征图 $\varphi'(\boldsymbol{x})$，利用 IoU 预测器对给定候选框进行评估，得到最佳的预测输出框。

3 实验设置与结果分析

3.1 实验设置

本实验使用大规模单目标跟踪数据集 LaSOT[14]、TrackingNet[15]和 GOT10k[16]对目标估计网络进行离线训练，同时采用 COCO 数据集[17]进行数据类别扩充。从最大帧间距为 50 帧的视频序列中选取图像样本对。对于参考图像，采样以目标为中心的正方形区域，面积约为目标的 5^2 倍，同时采用颜色抖动和翻转用于数据增强；在搜索图像中，采用类似操作生成补丁并在位置和尺度上进行一些扰动。对网络进行 50 个周期迭代训练，采用均方误差损失函数，使用 Adam 优化器，其中初始学习速率为 0.01，每 15 个周期的衰减系数比为 0.2，在线分类网络的参数设置参照 ATOM[6]。本实验基于 CUDA 10.0 和 PyTorch 1.1.0 编程语言实现，CPU 为 Intel i7-8700，GPU 为 NVIDIA RTX2070S，内存 16 GB。

3.2 定量分析

为验证所提模型的有效性和泛化能力，在 OTB100[18]、VOT2018[19]、LaSOT[14]、TrackingNet[15] 和 GOT10k[16]5 个具有挑战性的测试集上，对所提跟踪算法与主流的跟踪算法进行比较。

3.2.1 在 OTB100 数据集上测试

OTB100 数据集是最常用的视觉对象跟踪基准之一，包含了 100 个具有代表性的视频序列，根据不同的挑战，这些序列具有 11 种属性，包括遮挡、尺度变换、快速运动等。采用一次性评估的方式测试平均重叠率与中心位置误差，得到成功率与精度曲线图，成功率图显示不同重叠率阈值情况下成功跟踪帧的比率，精度图衡量的是跟踪算法预测的目标框与真实框之间的中心欧氏距离。

对所提算法与主流跟踪算法 SiamRPN[3]、ATOM[6]、DeepSRDCF[20]、SRDCF[21]等进行对比实验，结果如图 6 所示。所提算法展现出更优的跟踪精度和成功率，分别达到 0.887 和 0.682，与之前最佳的 ATOM 算法相比，所提算法在成功率和精确度上分别增加 1.9 个百分点和 1.3 个百分点。这验证了所提算法可以提供更准确的预测结果。

3.2.2 在 VOT2018 数据集上测试

VOT2018 由 60 个视频序列组成，所有序列都由以下挑战属性注释：光照变化、尺度变化、遮挡、

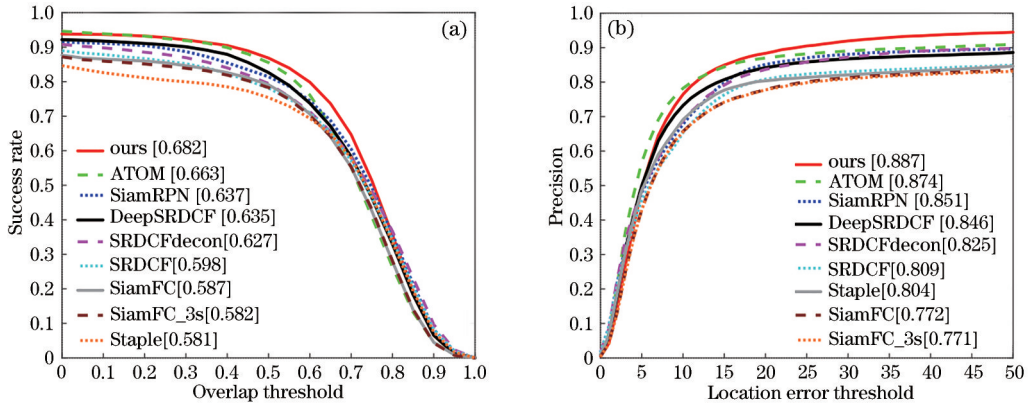

图 6　OTB100 数据集上的成功率与精度对比。(a)成功率;(b)精度

Fig. 6　Comparison of success rate and precision on OTB100 dataset. (a) Success rate; (b) precision

摄像机运动和运动变化。相对于其他数据集，VOT2018 具有目标尺度小、非刚性形变大等特点。为提高对数据集的有效利用，采取跟踪失败后再重启的机制来全方位评估算法性能。VOT2018 采用鲁棒性(robustness, Rob.)、准确性(accuracy, Acc.)以及预期平均重叠率(EAO)来评估跟踪器的性能。鲁棒性反映跟踪器的稳定程度，失败重启次数越少，指标越低，跟踪器稳定性越高。准确性反映预测框与标注框之间的平均重合率。其中 EAO 评分可以综合反映准确性和鲁棒性，当精度越高，鲁棒性越低，EAO 值则越高，跟踪器综合性能越好。

如图 7 所示，在 VOT2018 上对所提算法进行评估，并与包括基线算法 ATOM[6] 在内的其他 9 个跟踪

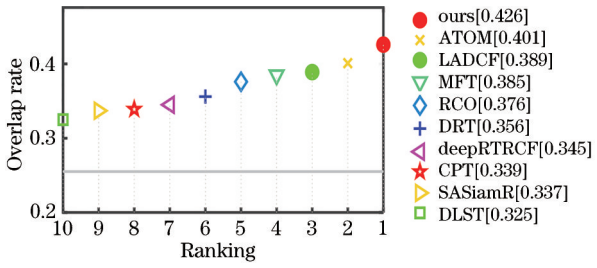

图 7　VOT2018 数据集上 EAO 性能

Fig. 7　EAO performance on VOT2018 dataset

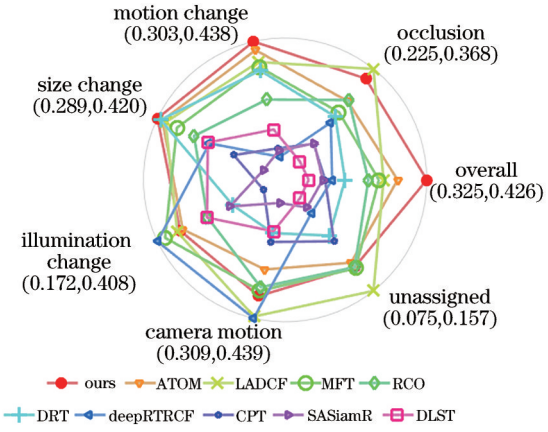

图 8　在 VOT2018 不同视觉属性下的 EAO

Fig. 8　EAO under different visual attributes on VOT2018

算法进行比较，在 EAO 排名中，所提算法取得了最好的结果。同时表 1 列出了不同跟踪器在 VOT2018 上的比较结果，所提算法的 EAO 评分为 0.426，相对于 ATOM，提升了 2.5 个百分点。此外，还在 VOT2018 上对跟踪器进行不同挑战属性的 EAO 值比较，如图 8 所示，所提算法在尺寸变化和运动变化属性上具有更好的优势，这得益于 MSI 模块融合多层次特征，同时金字塔调制模块聚合目标不同外观信息，使得测试分支更具有判别性。

表 1　在 VOT2018 数据集上的实验结果对比

Table 1　Comparison of experimental results on the VOT2018 dataset

Parameter	DLST	SASiamR	CPT	DRT	RCO	MFT	LADCF	ATOM	Ours
EAO	0.325	0.337	0.339	0.356	0.376	0.385	0.389	0.401	0.426
Acc.	0.543	0.566	0.507	0.519	0.507	0.505	0.503	0.590	0.597
Rob.	0.224	0.258	0.239	0.201	0.155	0.140	0.159	0.204	0.183

3.2.3　在 LaSOT 数据集上测试

LaSOT 数据集拥有 1400 个大规模的视频序列，并且提供了高质量的密集注释，由于其视频序列平均长度为 2512 帧，主要侧重于评估跟踪器的长期表现能力，采用与 OTB100 相似方法来评测不同算法的性能。将所提算法与 ATOM[6]、SiamFC[2]、

VITAL[22]等9种算法在测试数据集进行实验，得到不同跟踪器的成功率和归一化精度对比结果，如图9所示。结果表明，在两种评估标准下，所提跟踪算法性能优于其他所有算法，在成功率和精度上实现了0.524和0.597的得分，相比ATOM分别提高了0.9个百分点和2.1个百分点，与VITAL的0.390和0.453相比，分别有13.4个百分点和14.4个百分点提升。

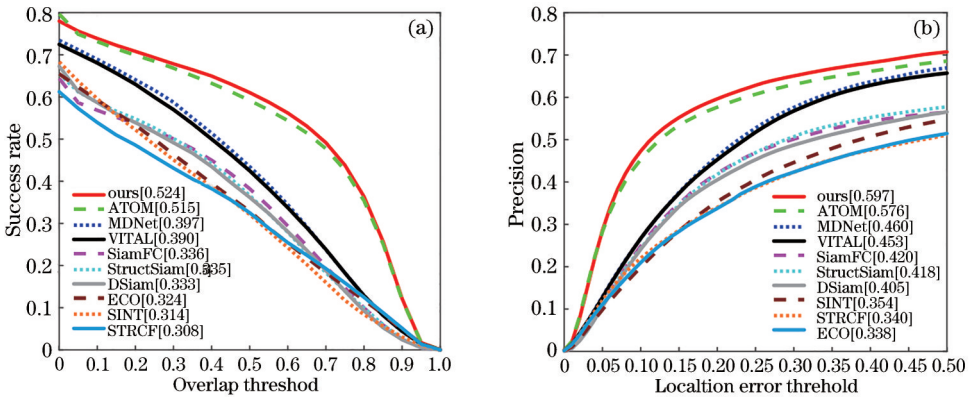

图9　LaSOT数据集上成功率与精度对比。(a)成功率；(b)归一化精度

Fig. 9　Comparison of success rate and accuracy on LaSOT dataset. (a) Success rate; (b) normalized precision

3.2.4　在TrackingNet数据集上测试

TrackingNet是针对目标跟踪任务设计的首个大规模的数据集，训练集由超过30×10^3个视频序列组成，包含大量野外场景，更贴近于真实世界中的跟踪场景。跟踪器根据成功率曲线下的面积(AUC)、精度(precision，Prec.)和归一化精度(normalized precision，Prec._N)进行评估。根据文献[15]对所提算法与表现最好的跟踪算法进行比较，包括ECO[23]、SiamFC[2]、SPM[24]、MDNet[25]、SiamMask[26]、ATOM[6]和D3S[27]。表2显示了对比结果，所提算法在所有性能指标上都取得了最好的成绩，与SiamFC相比，所提算法的成功率提升了大约16个百分点，与基准算法ATOM相比提高了2.5个百分点。

表2　在TrackingNet数据集上的实验结果对比

Table 2　Comparison of experimental results on the TrackingNet dataset

Parameter	ECO	SiamFC	SPM	MDNet	SiamMask	ATOM	D3S	Ours
AUC /%	55.4	57.1	71.2	60.6	72.5	70.3	72.8	72.8
Prec. /%	49.2	53.3	66.1	56.5	66.4	64.8	66.4	67.2
Prec._N /%	61.8	66.6	44.8	70.5	77.8	77.1	76.8	78.9

3.2.5　在GOT10k数据集上测试

GOT10k是最近一个大型高分辨率跟踪数据集，由10×10^3个视频序列组成，主要特点是测试集与训练集在对象类别上没有重叠，注重于评估视觉跟踪器的泛化程度。按照平均重叠度(AO)进行评估，同时在两个重叠阈值0.5和0.75处分别评估成功率(SR)。在测试序列上对所提算法进行评估，并与7个跟踪算法进行对比实验，结果如表3所示。所提算法在阈值为0.5和0.75上分别取得了66.7%和45.4%的成功率。所提算法的平均重叠率达到了58.2%，与SPM相比获得了6.9个百分比的性能提升，相对于基准算法ATOM获得2.6个百分比的性能提升。

表3　在GOT10k数据集上的实验结果对比

Table 3　Comparison of experimental results on the GOT10k dataset

Parameter	ECO	SiamFC	SPM	MDNet	SiamMask	ATOM	D3S	Ours
AO /%	31.5	34.8	51.3	29.9	51.4	55.6	59.7	58.2
$\text{SR}_{0.75}$ /%	11.1	9.8	35.9	9.9	36.6	40.2	46.2	45.4
$\text{SR}_{0.5}$ /%	30.9	35.3	59.3	30.3	58.7	63.5	67.6	66.7

3.2.6　跟踪速度与性能的权衡分析

为了进一步分析所提算法跟踪的实时性，基于LaSOT长时跟踪数据集对所提算法与其他主流跟踪算法进行跟踪帧率与成功率的对比实验，结果如图10所示。本文采用计算量适中的ResNet-18作为主干网络，并引入轻量级的SK注意力模块和MSI模块，相较于基准算法ATOM，在提升跟踪性能的同时跟踪速度上几乎没受到影响。由于采用的是两阶段的跟踪方法，即先利用相关滤波算法进行粗定位，再利用目标估计网络进行精细定位，所以在跟踪速度上所提算法与Siamese系列算法相比存在一定的差距，但在长时跟踪过程中所提算法具有较高的跟踪性能，并且速度达到27 frame/s，在基本满足实时性要求的同时具有良好的性能。

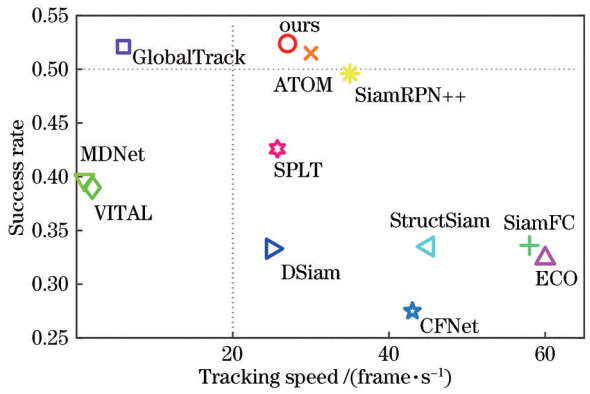

图10　在LaSOT数据集上不同算法的速度与性能对比

Fig. 10　Comparison of speed and performance of different algorithms on LaSOT dataset

3.3　消融实验

为了更深入地分析每个单独组件对性能增益的贡献，对单个模块进行了额外的消融实验，结果如表4所示。通过添加或删除提出的模块，在VOT2018和GOT10k数据集上利用EAO与AO来比较不同变体的性能增益。

通过依次添加SK、MSI模块之后，对比①②⑤行，VOT2018的EAO提高了1.1个百分点和2.0个百分点，GOT10k的AO值提高了0.8个百分点和1.9个百分点，这一显著的改进表明，SK和MSI模块是性能提升的主要贡献者。为了验证PMB的有效性，通过构造一个变体，即在所提算法中移除PMB，对比⑤⑥行，如果去掉特征金字塔调制模块，EAO与AO分数分别下降0.5个百分点和0.7个百分点，这表明提出的特征金字塔调制模块能有效提升跟踪性。

表4　加入各项组件后，所提算法的跟踪结果

Table 4　Tracking results of the proposed algorithm after adding various components

No.	SK	MSI			PMB	VOT2018	GOT10k
		SFB	MFB-MCAM	MFB		EAO	AO
①						0.401	0.556
②	√					0.412	0.564
③	√	√				0.416	0.571
④	√	√	√			0.413	0569
⑤	√	√		√		0.421	0.575
⑥	√	√		√	√	0.426	0.582

此外还对MSI中各个子模块进行了单独分析，设计SFB的动机是在层内挖掘目标的多尺度特征信息，MFB则是利用MCAM自适应地将低层特征的上下文信息和高层特征的语义信息相结合。对比②③⑤行，在依次添加SFB和MFB模块后EAO值提高了0.4个百分点和0.9个百分点，AO值提高了0.7个百分点和1.1个百分点。MCAM利用全局和局部注意力捕获不同特征通道之间的依赖关系，对比④⑤行，MFB-MCAM仅采用常规拼接操作对特征进行融合，可以发现在添加MCAM后，EAO值与AO值分别提高了0.8个百分点和0.6个百分点，表明适当的通道权重分配有助于增强最有效的通道，提升跟踪性能。

3.4　定性分析

为了直观理解所提算法与各对比跟踪算法在应对复杂跟踪环境时的实际跟踪性能，选取OTB100中具有代表性的5组视频序列（skating、bird1、soccer、board、diving）与其他经典算法（ATOM[6]、SiamRPN[3]、RT-MDNet[28]、SRDCF[21]）进行了定性对比实验，如图11所示。所提跟踪算法能够很好地应对复杂场景，尤其是处理尺度变化、目标变形和旋转等问题。在昏暗场景下跟踪滑冰者时，场地相似目标较多，背景也发生了剧烈的变化，ATOM和RT-MDNet等跟踪算法发生漂移，所

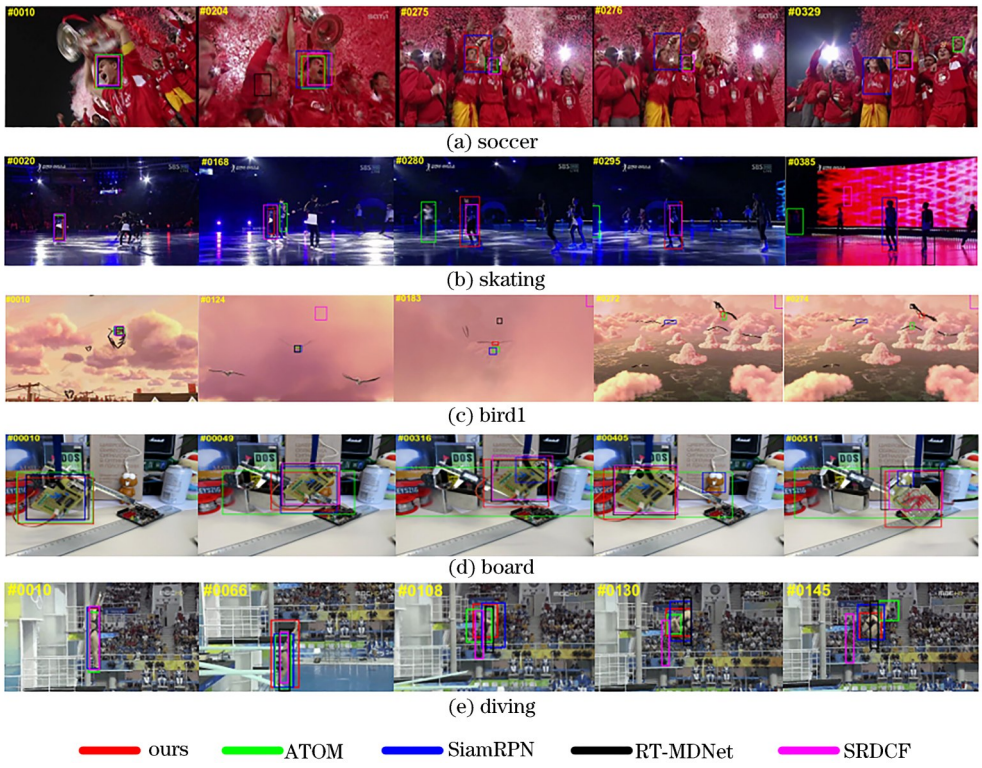

图 11　5种算法的跟踪效果

Fig. 11　Tracking results of five algorithms

提跟踪算法仍然成功地跟踪了目标。大多数跟踪算法都不足以应对足球运动员中人脸遮挡和背景干扰的问题，虽然所提算法在跟踪时同样发生了短暂漂移现象但能够感知上下文信息的变化进行快速纠正。在跳水运动员视频序列中，由于人体姿态发生了巨大的形变与旋转，各种跟踪算法不能生成精确的边界框，而所提算法的工作表现良好。

4　结　　论

提出了一种结合注意力机制与特征融合网络调制的目标跟踪算法。引入SK注意力模块为模型提供动态感受野，同时采用层内多尺度特征与层间不同特征结合的方式进一步提升网络对目标的识别能力；通过金字塔调制模块充分利用目标的空间信息，构建目标在不同尺度下的外观模型，提高了跟踪器对目标变化的适应性。实验结果表明，所提算法在跟踪精确率与成功率上都有较大的提升，在形变和尺度变化下具备较好的鲁棒性。但当目标在较为复杂的背景下，出现长期遮挡时会影响算法的适应性，因此在未来的工作中，针对目标跟踪的特点对注意力网络进行改进，同时探索更灵活的特征融合策略，进一步提高跟踪性能是关注的重点。

参　考　文　献

[1] 毛宁，杨德东，李勇，等. 基于形变多样相似性的空间正则化相关滤波跟踪[J]. 光学学报，2019，39(4)：0415002.
Mao N，Yang D D，Li Y，et al. Spatial regularization correlation filtering tracking via deformable diversity similarity[J]. Acta Optica Sinica，2019，39(4)：0415002.

[2] 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.

[3] 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.

[4] Ren S Q，He K M，Girshick R，et al. Faster R-CNN：towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2017，39(6)：1137-1149.

[5] 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.

[6] Danelljan M, Bhat G, Khan F S, et al. ATOM: accurate tracking by overlap maximization[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: 4655-4664.

[7] Jiang B, Luo R, Mao J, et al. Acquisition of localization confidence for accurate object detection [M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11218: 816-832.

[8] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[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: 7132-7141.

[9] Li X, Wang W H, Hu X L, et al. Selective kernel networks[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: 510-519.

[10] 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.

[11] Yu Y C, Xiong Y L, Huang W L, et al. Deformable Siamese attention networks for visual object tracking [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 6727-6736.

[12] Cao Y, Xu J R, Lin S, et al. GCNet: non-local networks meet squeeze-excitation networks and beyond[C]//2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), October 27-28, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 1971-1980.

[13] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11211: 3-19.

[14] Fan H, Lin L T, Yang F, et al. LaSOT: a high-quality benchmark for large-scale single object 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: 5369-5378.

[15] Müller M, Bibi A, Giancola S, et al. TrackingNet: a large-scale dataset and benchmark for object tracking in the wild[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11205: 310-327.

[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] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context[M]//Fleet D, Pajdla T, Schiele B, et al. Computer vision-ECCV 2014. Lecture notes in computer science. Cham: Springer, 2014, 8693: 740-755.

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

[19] Kristan M, Leonardis A, Matas J, et al. The sixth visual object tracking VOT2018 challenge results [M]//Leal-Taixé L, Roth S. Computer vision-ECCV 2018 workshops. Lecture notes in computer science. Cham: Springer, 2018, 11129: 3-53.

[20] Danelljan M, Häger G, Khan F S, et al. Convolutional features for correlation filter based visual tracking[C]//2015 IEEE International Conference on Computer Vision Workshop, December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 621-629.

[21] Danelljan M, Häger G, Khan F S, et al. Learning spatially regularized correlation filters for visual tracking[C]//2015 IEEE International Conference on Computer Vision, December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 4310-4318.

[22] Song Y B, Ma C, Wu X H, et al. VITAL: visual tracking via adversarial learning[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: 8990-8999.

[23] 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, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6931-6939.

[24] Wang G T, Luo C, Xiong Z W, et al. SPM-tracker: series-parallel matching for real-time visual object 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: 3638-3647.

[25] Nam H, Han B. Learning multi-domain convolutional neural networks for visual tracking[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 4293-4302.

[26] Wang Q, Zhang L, Bertinetto L, et al. Fast online object tracking and segmentation: a unifying approach [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: 1328-1338.

[27] Lukežič A, Matas J, Kristan M. D3S: a discriminative single shot segmentation tracker[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 7131-7140.

[28] Jung I, Son J, Baek M, et al. Real-time mdnet[M]// Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11208: 89-104.