时间显著注意力孪生跟踪网络

毛 琳，苏宏阳，杨大伟

（大连民族大学 机电工程学院，辽宁 大连 116600）

摘 要：针对现有孪生神经网络仅利用空间信息，面对目标遮挡、消失、表观剧烈形变等挑战造成跟踪准确度下降问题，提出一种时间显著注意力孪生跟踪网络。该网络通过信息交换“桥梁”，一方面为当前帧添加时间显著注意力，引导网络重点学习目标特征；另一方面对内存网络中历史目标特征进行筛选，将其作为附加模板，提供目标额外表观信息，同时遵循学习目标表观信息与空间位置的变化规律，指导后续检测、分类过程。为提高时间显著注意力能力，提出多尺度特征提取单元，解决骨干网络特征提取不充分的问题。在Got-10k数据集上进行模型测试，与目标跟踪算法时空记忆网络（STMTrack）相比，*AO*值提高2.4%。可视化结果显示，注意力孪生跟踪网络在目标遮挡、消失等挑战中，具有更高准确性。

关键词：计算机视觉；目标跟踪；目标遮挡；多尺度；特征融合；时间显著注意力

中图分类号：TP391 **文献标志码**：A **文章编号**：1671-5497(2024)11-3327-11

DOI：10.13229/j.cnki.jdxbgxb.20230003

Temporal salient attention siamese tracking network

MAO Lin，SU Hong-yang，YANG Da-wei

（*School of Electromechanical Engineering*，*Dalian Minzu University*，*Dalian* 116600，*China*）

Abstract：Aiming at the problem that the existing siamese network only use spatial information，and face the challenges of object obstruction，disappearance，apparent severe deformation and so on，which leads to the decrease of tracking accuracy，a temporal salient attention siamese tracking network is proposed. Through the information exchange “bridge”，the network on the one hand adds salient attention to the current frame，and guides the network to focus on learning the object characteristics；on the other hand，the features of historical object in the memory network are screened，and they are used as additional templates to provide the external appearance information of object，at the same time，the changing rules of the external information and spatial position of object are studied to guide the subsequent detection and classification process. In order to further improve the ability of temporal attention，a multi-scale feature extraction unit is proposed to make up for the insufficient feature extraction of backbone network. The

收稿日期：2023-01-04.

基金项目：国家自然科学基金项目（61673084）；辽宁省自然科学基金项目（20170540192，20180550866，2020-MZLH-24）.

作者简介：毛琳（1977-），女，副教授，博士.研究方向：目标跟踪与多传感器信息融合，目标跟踪与轨迹预测.
E-mail：maolin@dlnu.edu.cn

通信作者：杨大伟（1978-），男，副教授，博士.研究方向：计算机视觉处理技术，目标跟踪与轨迹预测.
E-mail：yangdawei@dlnu.edu.cn

model is tested on Got-10k data set, and compared with the object tracking algorithm STMTrack, the AO value is improved by 2.4%. According to the visualization results, this network has higher accuracy in the challenges of object obstruction and disappearance.

Key words: computer vision; object tracking; object obstruction; multi-scale; feature fusion; temporal salient attention

0 引 言

现有以SiamFC[1]为代表的孪生神经网络,以其简单、快速的优秀点在目标跟踪领域迅速崛起。但大多数孪生神经网络只考虑目标空间信息,忽略时间信息的重要性,而时间信息利用一直是目标跟踪领域核心问题,将其引入目标跟踪算法,可以很好地解决目标遮挡、目标快速运动等导致的跟踪不准确问题。

在目标跟踪中,可将跟踪器按信息利用分为空间跟踪器和时空跟踪器,孪生神经网络在空间跟踪器中占据重要地位。近几年,以SiamRPN[2]为基础,衍生出许多基于区域建议网络(Region proposal network,RPN)的跟踪算法。文献[3, 4]针对SiamRPN在目标表观相似和表观剧烈变化等场景下,对跟踪效果不佳问题进行改进,二者都采用将RPN进行级联的方式,提高网络对目标和背景辨别能力,与SiamRPN相比,RPN有较大提升,但多个RPN进行级联不仅引入了更多计算,在目标遮挡等挑战下仍不具备更强鲁棒性。文献[5]则借鉴FCOS(Fully convolutional one-stage)思想,在孪生跟踪器预测框回归阶段,加入质量评估分支,引入FCOS中心度公式,强化预测框回归效果,成为主流预测回归方法。但其更专注前景背景区分,与SiamRPN相比,RPN虽然取得更好的性能,但依旧无法适应目标遮挡等挑战。文献[6]则针对目标旋转问题,通过将跟踪目标旋转不同角度,构成一组实例集合,统一学习不同目标实例之间的共同特性。因此,它针对目标旋转问题取得良好的成果,但在其他目标跟踪挑战场景中仍显不足。

如今,时间信息利用已然成为目标跟踪的一个研究热点。在文献[7]中,提出一种动态记忆网络(Dynamic memory network),将目标历史表观信息存储在网络外部内存中,将长短期记忆(Long short-term memory,LSTM)用作内存控制器,输入为搜索特征映射,输出为内存块读写过程的控制信号,应用注意力机制使潜在目标特征作为LSTM输入,来适应跟踪过程中目标外观变化,取得良好效果。文献[8]则以编码-解码转换器作为关键组件,编码器包含用于目标定位的目标表观信息,对目标和搜索区域之间的全局时空特征依赖关系进行建模,而解码转换器包含跨帧目标的状态变化,通过学习目标时空运动规律预测目标空间位置,在多个数据集基准上取得先进性能。文献[9]提出一种基于时空记忆网络(Space-time memory network)的跟踪框架,通过存储目标历史信息,并计算当前帧与历史信息的像素级相似度,引导跟踪器聚焦于相似度最高的区域,针对目标外观变化获得优秀适应能力。基于时间信息的跟踪器[8,9]相对于大多仅空间信息的跟踪器[4,5,10,11],无需复杂的后处理来选择最佳预测框作为跟踪结果,在目标遮挡、相似外观等场景下更具鲁棒性。

在其他计算机视觉领域中,时间信息利用已经愈发普遍。在目标检测中,文献[12]通过提出一种时空存储网络(Spatial and temporal memory networks,STMN),将历史帧中干扰因素特征储存在内存网络中,在后续检测中,通过与内存网络特征匹配,排除干扰因素特征,有效解决人物视频空间干扰因素,在多个数据集基准下取得先进性能。在目标分割中,文献[13]提出一种半监督视频对象分割方法,通过将历史帧中目标添加掩膜,存储为一个外部内存,当前帧使用内存中掩膜信息进行分割,以此应对外观变化和闭塞等挑战,实现优秀的性能。文献[14]同样提出一种半监督视频目标分割方法——区域记忆网络(Regional memory network,RMNet),在RMNet中,通过记忆目标在历史帧的局部区域,构建区域记忆内存,使当前帧与历史帧进行精确的局部到局部匹配,有效缓解了相似对象在跟踪过程中的模糊性问题,使信息能够高效地从区域记忆内存传递到查询区域。文献[15,16]也通过构建内存网络的方

式，在视频对象分割领域获得了先进性能。

本文将时间信息引入孪生神经网络，提出时间显著注意力网络（Temporal salient attention network，TESANet），与大多数仅使用空间信息的孪生跟踪算法[1,10]不同，TESANet从整个视频序列中动态采样目标特征，存储在内存网络中作为附加模板，提供额外目标表观信息。在Got-10k测试集上，针对目标遮挡、消失等问题，取得了良好性能，相比于基线算法，AO提高2.4%，$SR_{0.5}$提高3.1%，$SR_{0.75}$提高2.3%。

1　时间显著注意力网络

在目标跟踪中，可将视频序列信息分为空间信息和时间信息，空间信息可以有效反映空间特征分布、目标位置、形状大小，进而对目标进行准确定位。但是，目标特征在整个视频序列中并不是一成不变的，会出现遮挡、表观形变等不利于目标定位的情况，此时空间信息无法为算法提供有效信息，甚至提供错误信息。而时间信息反映物体连续的运动轨迹和空间位置，以及表观信息变化规律，算法通过时间信息的利用可以更好地适应目标表观变化，以及在遮挡、表观形变等场景下对目标位置进行预测。

本文引入时间信息，提出目标显著模块（Object Salient Module，OSM）和内存模块（Memory Module，MEM）。在OSM中，通过历史帧对当前帧特征进行处理，为当前帧增加时间显著注意力。具体来说，在内存网络中，包含按时间顺序存储的k个历史帧，通过信息交换“桥梁”，计算特征相似矩阵，并对当前帧进行特征检索，若当前帧某一特征在历史帧中找到相似特征，则赋予其相应权重，通过多帧对比进行权重叠加，为当前帧赋予时间显著注意力，便可对目标特征实现重点学习。而MEM则为跟踪提供额外的表观信息和空间位置信息。

如图1所示，当前帧特征中三角形为目标特征，其他为背景干扰特征。在OSM中，通过特征相似度计算，与全部历史帧进行特征相似计算；将多组相似特征映射融合，获得显著特征；最终，与原始当前帧特征相加，使目标特征突显，帮助网络更好地区分目标特征与背景特征，并学习到正确特征。在MEM中，则通过特征相似度计算，筛选历史帧中的目标特征，将筛选的目标特征输出，作为附加模板，提供额外表观信息。并且，通过对历史帧中目标特征进行学习，获得其运动状态与表观信息变化规律，提高模板匹配准确度，实现更准确的目标位置预测。

1.1　目标显著模块（OSM）

为在多种跟踪挑战场景下，获得更高鲁棒性，如图2所示，在OSM中，主要对当前帧进行3个操作：①多尺度特征融合。为获得更具判别力的特征，使用多尺度特征提取（Multi-scale Extraction，MSE）单元对当前帧特征进行处理。②显著性特征计算。为获取目标可能存在的位置，通过信息交换“桥梁”，与历史帧进行特征相似度计算，获得相似度矩阵，并将相似度矩阵对MSE单元处理过的当前帧进行特征筛选，获得相似特征。③时间显著注意力。通过特征融合将多个相似特征进行相加，形成显著性特征，与原始当前帧特征叠加在

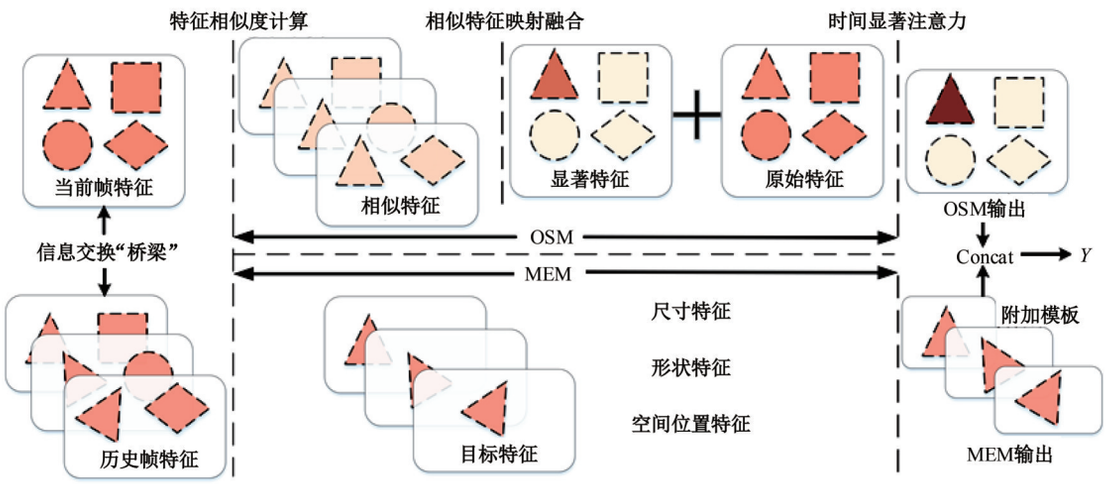

图1　TESANet逻辑示意图

Fig. 1　TESANet schematic diagrams

一起,构成注意力机制,突显目标特征,促使网络学习正确特征。

1.1.1 多尺度特征提取(MSE)

人眼对物体识别时,通常会先注意到轮廓,然后才观察物体纹理细节。神经网络也需要从高到低、从大到小、从粗到细的多层次、多尺度、多分辨率地获取目标信息[17]。如图3(a)所示,受GoogLeNet的Inception结构的启发,设计了MSE单元。为获得目标多尺度特征,MSE单元分别将3×3、5×5、7×7、9×9尺寸的卷积核并联,以此适应跟踪目标尺寸变化。如图3(b)所示,MSE单元不再对多通道特征进行拼接,而是采用直接融合方式,使多个特征复合为更具有判别能力的特征,实现优势互补。

为减少网络参数量,在图3(c)中,使用多个3×3卷积分别等效替换[18]5×5、7×7、9×9卷积,但每个MSE单元中仍有10个3×3卷积,因此在图3(d)中,通过共享卷积方式,进行卷积间残差连接,最终每个MSE单元中仅有4个3×3卷积。为了保证后续卷积能正确学习目标特征,采用分段融合方式,实现细节纹理特征和语义特征准确性的统一。MSE单元最终结构如图3(e)所示。

为更好地学习目标特征,OSM将MSE单元进行级联,不断进行高低层特征融合,实现细节纹理特征和语义特征的统一,在每个OSM中采用4个MSE单元进行级联,以生成目标更具代表性的特征。

1.1.2 显著性特征

为更好地定位目标,获得目标特征,可以使当前帧与内存网络进行信息交互,计算当前帧与历史帧的像素级相似度,对目标实现像素级定位。相似度矩阵即为信息交换“桥梁”,定义如下。

定义1 存在当前帧特征$x\in\mathbb{R}^{C\times HW}$,历史帧特征集合$M=\{z_r\in\mathbb{R}^{C\times HW}, r=1,2,\cdots,k\}$,计算$x$与集合$M$中$k$个历史帧$z$的相似度矩阵$\boldsymbol{P}$,$\boldsymbol{P}$的计算可表示为:

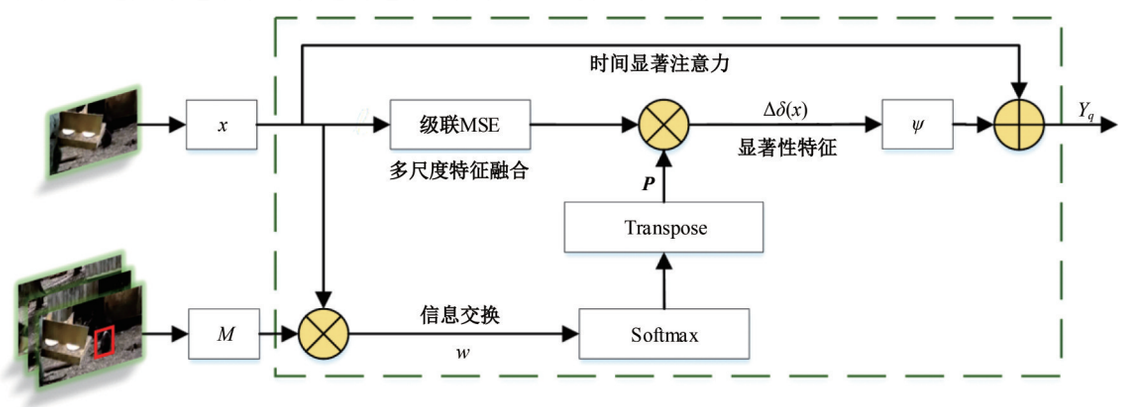

图2 OSM结构图

Fig. 2 OSM structure diagram

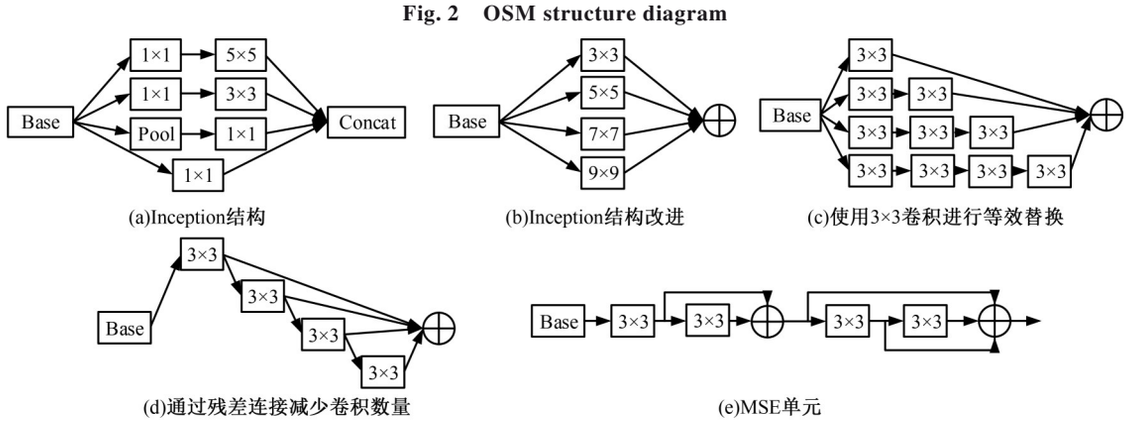

图3 多尺度特征提取

Fig. 3 Multi-scale extraction

$$\boldsymbol{P}=\underset{r=1}{\overset{k}{\mathrm{concat}}}\left\{\boldsymbol{P}_r|(\boldsymbol{P}_r)_{ij}=\frac{\exp((x\otimes M_r)_{ij}/\sqrt{C})}{\sum_{i=0}^{HW}\sum_{j=0}^{HW}\exp((x\otimes M_r)_{ij}/\sqrt{C})}\right\}$$
$$r=1,2,\cdots,k \tag{1}$$

式中：r为集合M中元素索引；M_r为历史帧；x与M_r按照特征通道数C进行矩阵叉乘；为方便起见，将$x\otimes M_r$记作w，w维度为$HW\times HW$；i、j为w宽和高索引，w_{ij}为w每一元素；$\sqrt{C}$为防止指数化过程中造成溢出所除的常数；exp为指数化操作；$\boldsymbol{P}_r$为M_r对应的相似度矩阵；concat为拼接操作，$\boldsymbol{P}$为最终相似度矩阵。

目标在运动过程中，会时刻发生表观和位置变化，初始目标信息可能与当前目标信息有很大差别，尤其是目标发生表观变化和遮挡等情况时，初始信息不足以支撑算法进行目标跟踪。因此，本文通过信息交换“桥梁”，计算任一M_r与x特征相似度，将相似度矩阵$\boldsymbol{P}$作为约束条件，对当前帧特征进行筛选，获得x与目标运动过程中相似的特征映射，将其称为显著性特征$\Delta\delta(x)$。这一过程可表示为：

$$\Delta\delta(x)=\underset{r=1}{\overset{k}{\mathrm{concat}}}(\boldsymbol{P}_r\otimes x),r=1,2,\cdots,k \tag{2}$$

在OSM中，$\Delta\delta(x)$代表当前帧与历史目标的相似特征映射，相对于整个目标运动过程而言，每个相似特征映射都是片面的，不能确定此相似特征映射是正确的。因此，将获得的显著性特征$\Delta\delta(x)_i$通过函数进行特征融合，形成更能够代表目标的特征。计算公式如下所示：

$$\begin{cases}\Delta\delta(x)\in\mathbb{R}^{C\times THW}\\ \Delta\delta(x)_1\in\mathbb{R}^{C\times T_1HW}\\ \Delta\delta(x)_2\in\mathbb{R}^{C\times T_2HW}\\ \quad\vdots\\ \Delta\delta(x)_i\in\mathbb{R}^{C\times T_iHW}(i=1,2\cdots T)\\ \psi(\Delta\delta(x))=\\ \quad\dfrac{\Delta\delta(x)_1+\Delta\delta(x)_2+\cdots+\Delta\delta(x)_T}{T}\in\mathbb{R}^{C\times HW}\end{cases} \tag{3}$$

式中：$\Delta\delta(x)$为4维矩阵；$\Delta\delta(x)_1$、$\Delta\delta(x)_2$；…；$\Delta\delta(x)_T$为$\Delta\delta(x)$按照维度T进行切片，获得的3维特征矩阵；$\psi()$为特征融合函数。特征融合后的$\varphi(\Delta\delta(x))$为显著性特征映射，是在多个相似特征映射中取交集，更能够代表目标特征。

1.1.3　时间显著注意力

为面对目标遮挡、形变等挑战，使网络更好地学习目标特征，如图2所示，本文通过计算当前帧与历史帧中目标特征相似性，获取显著性特征映射$\psi(\Delta\delta(x))$，将其作为软权重，通过残差连接，为当前帧特征增加时间显著注意力。这一过程表示为：

$$Y_q=\psi(\Delta\delta(x))+x \tag{4}$$

式中：$\psi(\Delta\delta(x))$为显著性特征映射；x为MSE单元处理后的当前帧特征；Y_q为OSM模块输出。

时间显著注意力可以理解为：若当前帧某一特征在越多的历史帧中找到相似特征，则在$\psi(\Delta\delta(x))$中权重就越大，网络就会着重学习此特征；相反，若与历史帧中任何一特征都不相似，则网络就会忽略此特征的学习，以此达到重点学习目标特征的效果。

1.2　内存模块MEM

为更好地解决目标遮挡、目标消失等问题，内存模块以当前帧特征x和历史帧特征集合M作为输入，通过x与M相似特征计算，将相似特征作为附加模板进行输出，为模板匹配提供额外的表观信息。同时，指导后续网络学习目标表观信息和空间位置的变换规律，进而更好地对目标位置进行预测。MEM结构如图4所示。

与OSM模块相同，MEM模块也由信息交换桥梁与当前帧进行信息交互，通过相似性矩阵的计算对历史帧进行特征筛选，获得与当前帧相似的特征，表达式为：

$$Y_m=M\otimes\boldsymbol{P} \tag{5}$$

式中：M为k个历史帧特征集合；Y_m为MEM的输出。

最终，将提出的两个模块进行整合，构成TESANet，并通过级联OSM模块，进一步增强目标重点学习的能力，网络结构如图5所示。

在TESANet中，将MEM的输出Y_m与级联OSM的输出Y_q进行拼接，作为TESANet的最终

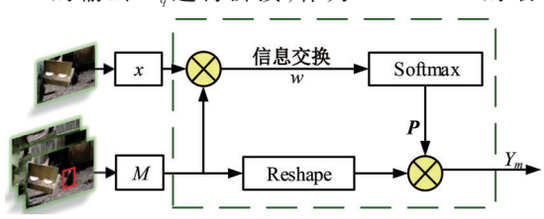

图4　MEM结构图

Fig. 4　MEM structure diagram

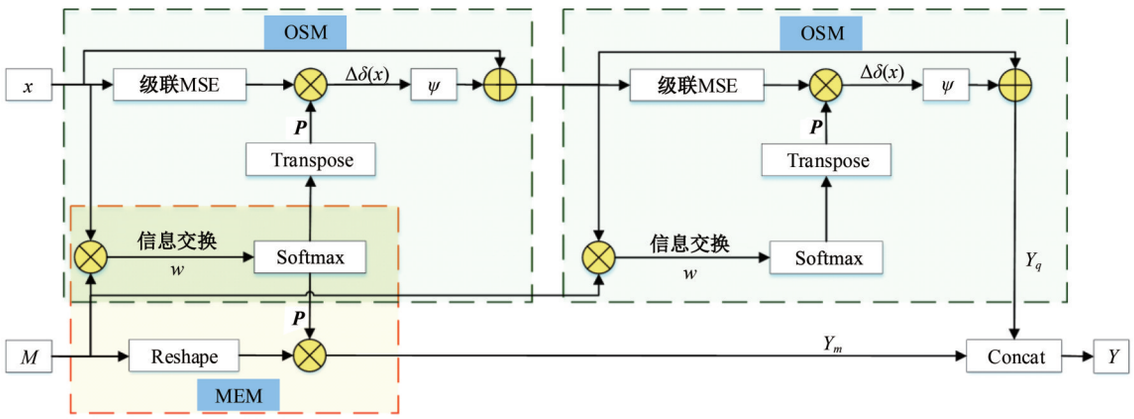

图 5 TESANet

Fig. 5 TESANet

输出 Y,输入到后续分类回归网络进行检测框的标定工作。Y的计算过程可表示为:

$$Y=\mathrm{concat}(M\otimes P,\psi(\Delta\delta(x))+x) \tag{6}$$

2 网络整体框架

将 TESANet 与孪生神经网络结合在一起,构成了本文整体网络,整体网络结构图如图 6 所示。

网络整体框架主要包括 3 个部分:骨干网络、TESANet、分类回归网络。其中,骨干网络采用孪生神经网络框架进行跟踪帧与历史帧的特征提取,并将动态采样的历史帧特征存储在内存网络中;TESANet 对骨干网络提取的特征进行特征交互,分别对当前帧特征和历史帧特征进行处理,生成后续网络更容易利用的特征;分类回归网络进行检测框的标定。

本文网络实施步骤如下:

步骤 1 将 k 个历史帧和当前帧分别输入骨干网络,进行历史帧和当前帧的特征提取。骨干网络皆采用 Inception V3 网络,但是二者网络参数并不共享。并且,为减少参数量,加快网络运行速度,将 Inception V3 网络最后的 7×7 平均池化和全连接层阶段去除[9],同时通过自适应卷积层对输出维度进行调节。

步骤 2 在历史帧特征提取后,输出为 k 个历史帧特征,以及对应的 k 个前景背景标签映射 L,存储在内存网络中。L 将真值框内的像素标为 1,而真值框外的像素标为 0,以区分目标特征和背景特征,确保后续网络可以正确提取目标特征。当前帧通过骨干网络进行特征提取,以及自适应卷积层调整特征维度,最终与内存网络一起作为 TESANet 的输入。

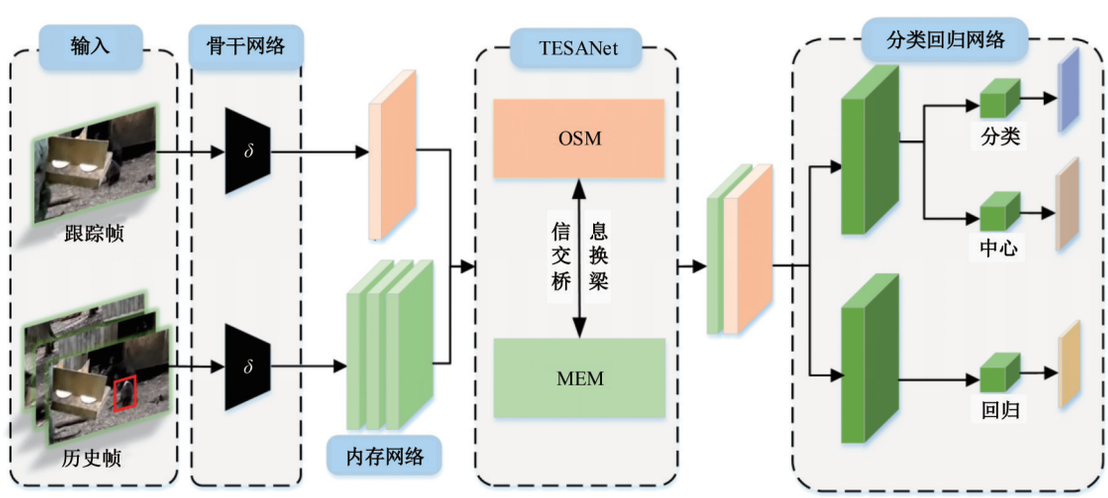

图 6 网络整体框架

Fig. 6 Overall framework of the network

步骤 3　获得当前帧与历史帧特征后，TESANet 通过 OSM 增加目标特征权重，使跟踪目标在当前帧内的特征突显，增强网络对跟踪目标局部特征的学习能力，并且，通过级联 OSM 的方式，不断增强这一效果。同时，MEM 利用多组历史帧特征，通过信息交换“桥梁”，筛选当前帧与历史帧中相似的特征，将多个相似特征作为附加模板，提供额外表观信息，促使网络更好地对当前帧中目标进行定位。最终，将 OSM 与 MEM 输出进行拼接，送入后续网络进行预测框标定。

步骤 4　采用无锚头网络[5]进行检测框的标定和回归。通过文献阅读发现，无锚检测器[5,19]比基于锚的方法[20]具有更好的性能和更少的参数，如图 6 所示，整个无锚头网络有 3 个分支：分类分支、中心度分支、回归分支，每个分支首先使用包含 7 个卷积的轻量级网络对 TESANet 传入的信息进行处理，随后使用单个卷积将其降维，以进行后续分类回归任务。分类分支用于前景背景分类，判断特征属于目标或者背景。中心度分支引入 FCOS 的中心度公式，强化预测框回归效果。回归分支直接估计目标位置，最终进行检测框的标定。

3　实验结果分析

3.1　实验设计

本算法使用 1 张 NVIADIA GeForce 1080Ti 显卡，在 Ubuntu16.04 环境中配置 PyTorch 深度学习框架训练和测试网络模型。整个训练过程为 10 个周期，每个周期进行 38 000 次迭代，批尺寸设为 8。骨干网络使用预训练的 Inception V3，在训练开始前冻结骨干网络反向传播过程，仅训练 TESANet 与预测回归头网络，在第 4 个训练周期对骨干网络解冻，与整体共同训练。

学习率采用分段方式，在第 1 个周期，初始学习率设为 1×10^{-6}，并采用线性学习率(LinearLR)，使其增至 6×10^{-3}，在第 2、3 个周期，采用余弦退火学习率(CosineAnnealingLR)，学习率从 6×10^{-3} 降至 1×10^{-6}，并在后续所有周期中保持 1×10^{-6} 不变。如图 7 所示，为学习率变化曲线。

在训练阶段，内存网络中存储的历史帧为 3 帧，经过骨干网络特征提取，并通过自适应卷积层处理后，历史帧与当前帧特征维度均为 512，送入 TESANet 进行信息交互。本文选用 Got-10k 数据集进行训练与测试，Got-10k 数据集共有 4 个评价指标，分别为真值与检测框的平均重叠率(AO)、重叠阈值为 0.5 的检测成功率($SR0.5$)、重叠阈值为 0.75 的检测成功率($SR0.75$)，以及算法运行速度(FPS)[21]。在 Got-10k 论文中，以 AO 为例，评价指标表达式为：

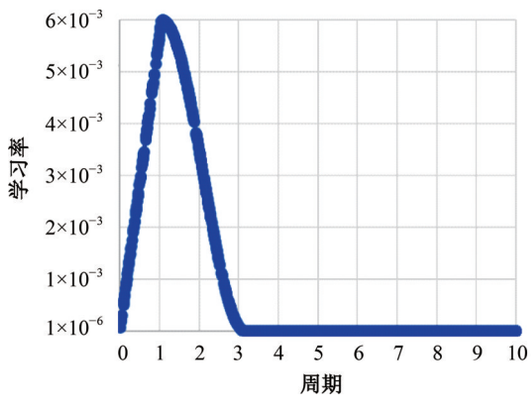

图 7　学习率变化曲线

Fig. 7　Learning rate change curve

$$mAO=\frac{1}{C}\sum_{C=1}^{C}\left(\frac{1}{|S_C|}\sum_{i\in S_C}AO_i\right) \tag{7}$$

式中：C 为数据集不同视频类别；S_C 为 C 中图片；$|S_C|$ 为 C 中图片数量。

如式(7)所示，首先对某个视频类别中所有图片的平均重叠率 AO 进行求和，并计算平均值，得到单个视频类别的 mAO 值；再对所有的视频类别的 mAO 求和取平均，便得到了最终的 mAO。$SR0.5$、$SR0.75$ 与 mAO 的计算方式相同，$SR0.75$ 相比于 $SR0.5$ 而言，对跟踪器的要求更严格。在 Got-10k 官网上以 mAO 为首要评价指标，mAO 即为 AO。

3.2　消融实验

为验证 OSM 模块的有效性，对 OSM 模块进行级联操作，与原始 STMTrack 进行对比，对比结果如表 1 所示。

由表 1 可知，不同个数的 OSM 级联均提升了算法精度，但在 3 个 OSM 级联出现了精度回降的情况。这是因为骨干网络已经进行了特征提取，但仍有不足，所以需进一步特征提取，以及高低层

表 1　OSM 对 AO 的影响

Table 1　Impact of the OSM on AO

跟踪器	OSM 个数	AO
STMTrack	0	0.642
TESANet	1	0.660
TESANet	2	0.666
TESANet	3	0.657

特征融合,在两个OSM包含的8个MSE单元处理后达到了局部最佳状态,再进一步特征提取便越过了精度极大值点,造成算法精度下降。

为进一步验证OSM与MSE单元不同组合方式对跟踪器造成的影响,TESANet中保持8个MSE单元不变,进行4次算法仿真实验,实验结果如表2所示。

如表2所示,在保持8个MSE单元不变基础上,根据不同组合方式,分别将其命名,各型TESANet均获得良好性能,其中Ⅱ型TESANet获得最高精度,综合第一组消融实验,本文认为,在每个OSM模块中包含4个MSE单元可以起到最好特征提取效果,因此将TESANet-Ⅱ作为本文最终算法。

表2 在Got-10k评估集上,OSM与MSE单元不同组合方式对*AO*的影响

Table 2 Impact of different combinations of OSM and MSE units on *AO* about the Got-10k evaluation set

名称	OSM数量	MSE数量	*AO*
TESANet-Ⅰ	*1	*8	0.658
TESANet-Ⅱ	*2	*4	0.666
TESANet-Ⅳ	*4	*2	0.661
TESANet-Ⅷ	*8	*1	0.662

3.3 算法对比

为了验证本文提出算法的有效性,在Got-10k数据集上进行测试并与基线算法STMTrack和其他部分代表性算法进行对比,对比结果如表3所示,在AO、$SR_{0.5}$、$SR_{0.75}$ 3个性能指标上,均取得了良好的结果,相比于基线算法,STMTrack也有较大提升。

表3 在Got-10k测试集上,TESANet-Ⅱ与其他跟踪器的比较

Table 3 Comparison of TESANet-Ⅱ compares to other trackers about the Got-10k test set

跟踪器	AO ↑	$SR_{0.5}$ ↑	$SR_{0.75}$ ↑
TESANet-Ⅱ	0.666	0.768	0.598
STMTrack[9]	0.642	0.737	0.575
MixFormer-1k[22]	0.712	0.799	0.658
SBT large[23]	0.704	0.808	0.647
STARK[8]	0.688	0.781	0.641
TrDiMP[24]	0.671	0.777	0.583
AutoMatch[25]	0.652	0.766	0.543
Siam R-CNN[11]	0.649	0.728	0.597
FCOT[26]	0.634	0.766	0.521
SBT light[23]	0.602	0.685	0.530
D3S[27]	0.597	0.676	0.462
SiamFC++[5]	0.595	0.695	0.479
SiamRPN++[4]	0.517	0.616	0.325

同时,为了验证模型泛化能力,将TESANet-Ⅱ在OTB-2015数据集上进行测试,并与基线算法和其他算法进行对比,对比结果如表4所示,在OTB-2015数据集上TESANet-Ⅱ的表现依旧优于大部分跟踪器,取得了0.716的成功率。可见,模型有较强泛化能力。

表4 在OTB-2015数据集上,TESANet-Ⅱ与其他跟踪器的比较

Table 4 Comparison of TESANet-Ⅱ compares to other trackers about the OTB-2015 dataset

跟踪器	Success	Precision	跟踪器	Success	Precision
TESANet-Ⅱ	0.716	0.923	ToMP-50[28]	0.701	—
STMTrack[9]	0.719	0.934	MixFormer-1k[22]	0.696	0.911
SBT large[23]	0.719	0.924	SiamRPN++[4]	0.696	0.914
SAOT[29]	0.714	0.926	KYS[30]	0.695	—
SiamAttn[31]	0.712	0.926	Ocean[10]	0.684	0.899
UPDT[32]	0.702	0.919	SiamFC++[5]	0.683	—

为进一步验证模型的有效性,将模型在VOT2018数据集上进行测试,根据VOT2018数据集的评估方法,本文算法在预期平均重叠(EAO),准确性(A)和鲁棒性(R)方面进行了测试,并与其他先进算法进行对比,对比结果如表5所示。

通过对比表5的数据,可以发现本文算法在预期平均重叠、准确性和鲁棒性方面均有所提升。特别是在鲁棒性方面,相较于其他目标跟踪算法表现更优异。本算法能够处理一些复杂的情况,如目标消失、目标遮挡及背景干扰等,并且在这些

表5 在VOT2018测试集上,TESANet-Ⅱ与其他跟踪器的比较

Table 5 Comparison of TESANet-Ⅱ compares to other trackers about the VOT2018 test set

跟踪器	EAO ↑	A ↑	R ↓
TESANet-Ⅱ	0.449	0.591	0.157
STMTrack[9]	0.447	0.590	0.159
D3S[27]	0.489	0.640	0.150
Ocean[10]	0.489	0.592	0.117
SiamAttn[31]	0.470	0.630	0.160
KYS[30]	0.462	0.609	0.143
SiamBAN[33]	0.452	0.597	0.178
PrDiMP-50[24]	0.442	0.618	0.165
DiMP-50[34]	0.440	0.597	0.153
Siam R-CNN[11]	0.408	0.609	0.220
SiamFC++[5]	0.426	0.587	0.183
SiamRPN++[4]	0.414	0.600	0.234

情况下表现依然出色。然而，与其他算法相比，本文算法在预期平均重叠和准确性等方面存在一定差距。这是因为VOT2018评估工具中提供的跟踪框是旋转的，而本文算法回归的跟踪框是平行于x轴和y轴的，因此，在VOT2018数据集中的表现会受到一定影响。尽管如此，这也进一步验证了本文算法在面对目标遮挡、目标消失和背景干扰等情况时，具有更高的鲁棒性，面对复杂场景时的跟踪效果在可视化分析部分进行展示。

3.4　可视化分析

图8为游艇水上航行视频，存在着目标遮挡、相似背景等挑战。在图8(b)、(c)STMTrack跟踪结果中，基线算法在目标受到栏杆遮挡后，进行模板匹配过程中无法找到目标特征，因此将背景船体误识别为目标，造成了跟踪失败，即使在图8(d)中目标重新出现，但由于模板信息受到前一帧跟踪结果干扰，依旧将背景误判为目标，造成了跟踪失败。而TESANet仅通过目标暴露的部分便准确推断出目标位置，如图8(c)第二行所示，良好地解决了目标遮挡问题。

如图9所示，为动物穿过桥体视频，存在目标消失挑战。由图9(c)(d)(e)STMTrack跟踪结果可以看出，基线算法在目标消失时，模板匹配再次失败，选取与目标距离最近、外观最相似的背景当作跟踪结果，造成了跟踪失败。而TESANet通过时间显著注意力，可防止误判的情况发生，并通过目标运动规律的学习，成功地对目标位置进行预测。如图9(c)第二行所示，即使在目标消失情况下，也同样预测出目标位置。

图10为大猩猩运动视频，存在着表观剧烈形变、目标遮挡等挑战。由图10(c)第一行STMTrack跟踪结果可以看出，基线算法无法适应目标表观剧烈形变，对背景进行误判，造成跟踪失败。TESANet则利用时间信息，通过当前帧与历史帧进行信息交互，计算多个显著性特征，将多个特征进行融合形成更能代表目标的特征，再通过MSE单元获得高级语义特征，使目标表观发生剧烈形变时，也成功地进行追踪。由此可见，本文算法在多种目标跟踪挑战场景下具有更强的鲁棒性。

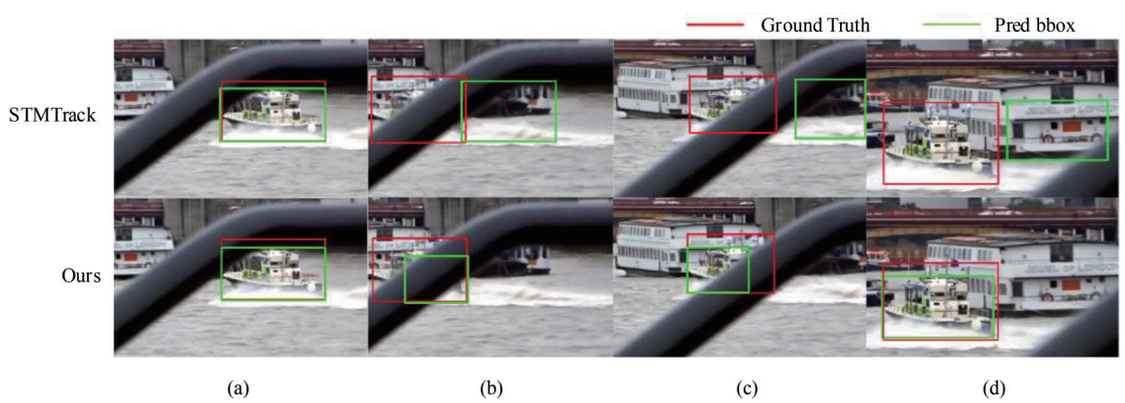

图8　目标遮挡的可视化对比

Fig. 8　Visual contrast of object obstruction

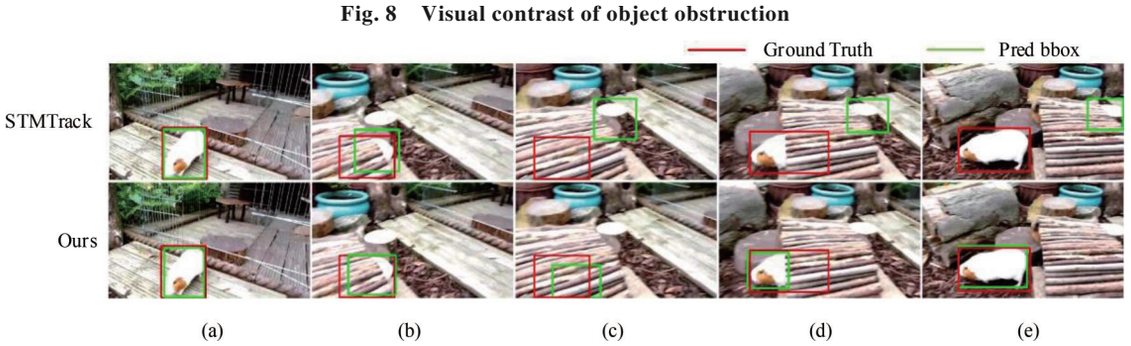

图9　目标消失的可视化对比

Fig. 9　Visual contrast of object disappear

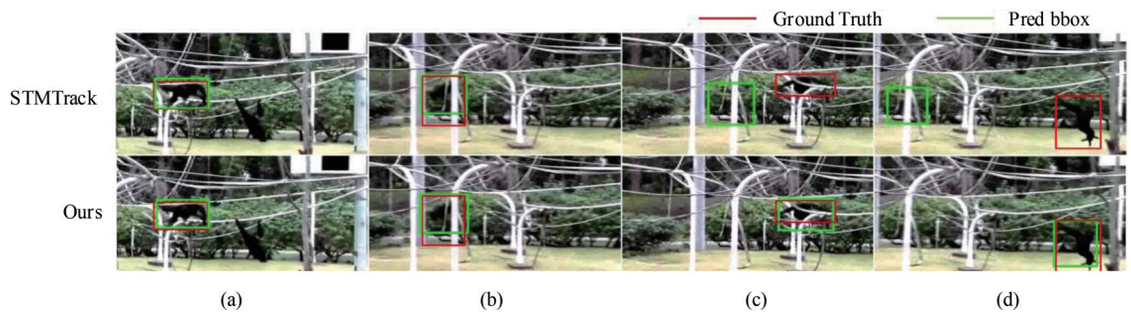

图10 表观剧烈形变的可视化对比

Fig. 10 Visual contrast of apparent violent deformations

4 结束语

为解决孪生神经网络因缺少时间信息利用而造成的目标遮挡等场景下跟踪不准确问题,本文将时间与空间信息结合在一起,提出多尺度时间显著注意力孪生跟踪网络问题,通过信息交换"桥梁",分别对当前帧特征和历史帧特征进行处理。OSM通过时间显著注意力,强化网络对目标特征学习能力,并且,MSE单元通过进行高低层特征融合,实现纹理细节特征与语义特征的统一,解决了骨干网络特征提取不充分问题。MEM则通过筛选多个历史帧中目标特征,作为附加模板,为目标跟踪提供额外表观信息,并挖掘潜在运动状态变化规律,提高了目标空间位置预测准确度。与现有网络相比,TESANet改善了在目标遮挡、目标消失、表观剧烈形变等挑战下跟踪不准确的问题。后续工作中将进一步增强算法预测能力,提高在目标遮挡、消失场景下的准确度。

参考文献:

[1] Bertinetto L, Valmadre J, Henriques J F, et al. Fully-convolutional siamese networks for object tracking [C]// European Conference on Computer Vision, Berlin, Germany, 2016: 850-865.

[2] Li B, Yan J J, Wu W, et al. High performance visual tracking with siamese region proposal network[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 8971-8980.

[3] Fan H, Ling H B. Siamese cascaded region proposal networks for real-time visual tracking[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 7952-7961.

[4] Li B, Wu W, Wang Q, et al. Siamrpn++: evolution of siamese visual tracking with very deep networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 4282-4291.

[5] Xu Y D, Wang Z Y, Li Z X, et al. Siamfc++: towards robust and accurate visual tracking with target estimation guidelines[C]// Proceedings of the AAAI Conference on Artificial Intelligence, New York, USA, 2020: 12549-12556.

[6] Gupta D K, Arya D, Gavves E. Rotation equivariant siamese networks for tracking[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021: 12362-12371.

[7] Yang T Y, Chan A B. Learning dynamic memory networks for object tracking[C]// Proceedings of the European Conference on Computer Vision (ECCV), Munichi, Germany, 2018: 152-167.

[8] Yan B, Peng H W, Fu J L, et al. Learning spatio-temporal transformer for visual tracking[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2021: 10448-10457.

[9] Fu Z H, Liu Q J, Fu Z H, et al. Stmtrack: template-free visual tracking with space-time memory networks [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021: 13774-13783.

[10] Zhang Z P, Peng H W, Fu J L, et al. Ocean: object-aware anchor-free tracking[C]//European Conference on Computer Vision, Berlin, Germany, 2020: 771-787.

[11] Voigtlaender P, Luiten J, Torr P H, et al. Siam R-CNN: visual tracking by re-detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 6578-6588.

[12] Eom C, Lee G, Lee J, et al. Video-based person re-identification with spatial and temporal memory networks[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2021: 12036-12045.

[13] Oh S W, Lee J Y, Xu N, et al. Video object segmentation using space-time memory networks[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, South Korea, 2019: 9226-9235.

[14] Xie H Z, Yao H X, Zhou S C, et al. Efficient regional memory network for video object segmentation [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021: 1286-1295.

[15] Paul M, Danelljan M, Van G L, et al. Local memory attention for fast video semantic segmentation[C]// 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 2021: 1102-1109.

[16] Wang H, Wang W N, Liu J. Temporal memory attention for video semantic segmentation[C]// 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, USA, 2021: 2254-2258.

[17] Yu F, Wang D Q, Shelhamer E, et al. Deep layer aggregation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 2403-2412.

[18] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 2818-2826.

[19] Tian Z, Shen C H, Chen H, et al. Fully convolutional one-stage object detection[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, South Korea, 2019: 9626-9635.

[20] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]// Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 2017: 2980-2988.

[21] 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, Intelligence Machine, 2019, 43(5): 1562-1577.

[22] Cui Y T, Jiang C, Wang L M, et al. Mixformer: end-to-end tracking with iterative mixed attention [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 13608-13618.

[23] Xie F, Wang C Y, Wang G T, et al. Correlation-aware deep tracking[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 8751-8760.

[24] Wang N, Zhou W G, Wang J, et al. Transformer meets tracker: exploiting temporal context for robust visual tracking[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021: 1571-1580.

[25] Zhang Z P, Liu Y H, Wang X, et al. Learn to match: automatic matching network design for visual tracking[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2021: 13339-13348.

[26] Cui Y T, Jiang C, Wang L M, et al. Fully convolutional online tracking[J]. Computer Vision and Image Understanding, 2022, 224: 103547.

[27] Lukezic A, Matas J, Kristan M. D3S-a discriminative single shot segmentation tracker[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 7133-7142.

[28] Mayer C, Danelljan M, Bhat G, et al. Transforming model prediction for tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 8731-8740.

[29] Zhou Z K, Pei W J, Li X, et al. Saliency-associated object tracking[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2021: 9866-9875.

[30] Bhat G, Danelljan M, Gool L V, et al. Know your surroundings: exploiting scene information for object tracking[C]// European Conference on Computer Vision, Berlin, Germany, 2020: 205-221.

[31] Yu Y C, Xiong Y L, Huang W L, et al. Deformable siamese attention networks for visual object tracking [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 6728-6737.

[32] Bhat G, Johnander J, Danelljan M, et al. Unveiling the power of deep tracking[C]// Proceedings of the European Conference on Computer Vision (ECCV), Munichi, Germany, 2018: 483-498.

[33] Chen Z D, Zhong B E, Li G R, et al. SiamBAN: target-aware tracking with siamese box adaptive network [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(4): 5158-5173.

[34] Bhat G, Danelljan M, Gool L V, et al. Learning discriminative model prediction for tracking[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, South Korea, 2019: 6182-6191.