DOI:10.19651/j.cnki.emt.2313046

联合高阶目标感知与相似匹配的目标跟踪算法*

张念超[1] 张宝华[1] 李永翔[2] 谷 宇[1]
(1.内蒙古科技大学信息工程学院 包头 014010;2.内蒙古农业大学能源与交通工程学院 呼和浩特 010018)

摘 要:视觉目标跟踪算法利用自注意力机制增强上下文联系,但面对复杂场景时,自注意力机制中的相关性易发生失配,为此提出一种联合高阶目标感知与相似匹配的目标跟踪算法。构建高阶目标感知模型,针对自注意力机制中的一阶自注意图,利用坍塌的极化过滤方式进行空间和通道维度的正交化建模,优化内部相关性;同时组合非线性拟合函数避免坍塌引起的信息损失,进而获得高阶自注意图,捕获具有高阶上下文信息的感知特征。通过不同维度分解目标的感知特征来细化匹配区域,抑制背景噪声并约束当前帧的响应图,提高网络的判别力。在 OTB100 和 UAV123 基准的实验结果表明,所提算法有更好的跟踪性能,可以有效应对相似干扰等问题。

关键词:计算机视觉;目标跟踪;自注意力机制;高阶目标感知;极化过滤;相似匹配
中图分类号:TP391.41 **文献标识码**:A **国家标准学科分类代码**:520.604

Object tracking algorithm with jointing high order target aware and similarity matching

Zhang Nianchao[1] Zhang Baohua[1] Li Yongxiang[2] Gu Yu[1]
(1. School of Information Engineering, Inner Mongolia University of Science and Technology, Baotou 014010, China;
2. College of Energy and Transportation Engineering, Inner Mongolia Agricultural University, Hohhot 010018, China)

Abstract: The self-attention mechanism is used to enhance context in the visual object tracking algorithm, but in the face of complex scenes, the correlation in the self-attention mechanism is prone to mismatch. Therefore, a high-order target aware and similarity matching object tracking algorithm was proposed. A high-order target aware model was Constructed for the first-order self-attention map in the self-attention mechanism, the collapsed polarization filtering method was used to perform orthogonal modeling of space and channel dimensions, and optimize internal correlation. At the same time, a nonlinear fitting function was combined to avoid information loss caused by collapse, and then a high-order self-attention map is obtained to capture perceptual features with high-order context information. The perceptual features of the target were decomposed in different dimensions to refine the matching area, so the background noise was suppressed and the response map of the current frame was constrained, and improve the discriminative power of the network. The experimental results on OTB100 and UAV123 benchmarks show that the proposed algorithm has better tracking performance, and can effectively deal with problems such as similar interference.

Keywords: computer vision; target tracking; self-attention mechanism; high order target-aware; polarization filtering; similarity matching

0 引 言

目标跟踪在视频监控、智能交通等领域有着广泛应用[1-2]。然而,由于跟踪场景的复杂多样性,目标跟踪任务易遭受目标遮挡、相似干扰等影响[3-4]。早期相关滤波算法(correlation filter,CF)的手动特征对目标多样性表现出较差的鲁棒性,因此传统跟踪算法仍有较大提升空间。近年来,以深度学习方式驱动的跟踪算法快速发展,如

收稿日期:2023-03-09
*基金项目:国家自然科学基金(61962046, 62262048, 62001255, 62066036, 61841204)、内蒙古科技计划项目(2020GG0315, 2021GG0082)、中央引导地方科技发展资金项目(2021ZY0004)、内蒙古草原英才、内蒙古自治区自然科学基金(2022MS06017, 2019MS06003, 2018MS06018)、教育部"春晖计划"合作科研项目(教外司留 1383 号)、内蒙古自治区高等学校科学技术研究项目(NJZY145)资助

SiamRPN[5]将跟踪视为样本检测问题，利用区域建议网络(region proposal network，RPN)区分前景和背景，避免多尺度测试，同时利用互相关聚合模板特征和搜索特征，提高跟踪器性能。但先前算法中平移不变性易受 Padding 影响，为此 SiamRPN＋＋[6]加深网络深度，并提出一种采样策略打破平移不变性的限制，从而提高算法性能。虽然以上算法更加有效，但易引入无目标背景框，表现出较差的跟踪鲁棒性。

视觉注意机制被证明可以有效捕获目标的关键信息，以提升目标表征能力，进而提高跟踪鲁棒性。杨梅等[7]构建通道联合空间的注意力模型提高目标关键特征的关注度。SiamDA[8]构建双重孪生网络，每重网络中嵌入非局部注意模块和通道注意模块，以突出目标区域并抑制背景；Zhang 等[9]使用不同风格的注意力获得更强的语义特征表达，并利用自适应决策融合策略实现稳定跟踪；Wang 等[10]提出目标感知注意力机制，通过联合局部和全局搜索搜素，确保预测目标感知注意力图的空间和时间一致性。此外，部分学者通过设计不同策略来利用语义信息，达到提高跟踪准确性的目的。付谱平等[11]通过增加语义特征分支与原有外观特则会那个分支形成互补，以充分利用两分支特征的异质性来提高算法的判别能力。DaSiamRPN[12]丰富训练过程中的类别信息，并构造有语义的负样本来提高感知目标信息的判别力；SiamCAR[13]以逐像素的方式解决视觉跟踪问题，并嵌入 Centerness 中心度量，避免预测时出现过大位移。然而，现有跟踪网络仍存在一定问题：1)视觉注意力机制内部存在相关性失配问题，即面对复杂背景时易造成语义模糊性，阻碍跟踪性能提升；2)现有孪生跟踪方法的互相关操作易造成较大的匹配区域，进而产生干扰响应，模糊空间信息。

针对上述问题，本文提出联合高阶目标感知与相似匹配的目标跟踪算法，工作如下：1)本文构建一种新的感知网络，称为高阶目标感知模型(High Order Target Aware Model，HTA)。高阶目标感知模型直接对提取的特征进行语义信息优化，该模型利用极化过滤方式，对一阶自注意图求取高阶注意图以降低错误相关性的可能，更好的捕获全局上下文信息关联性。2)受文献 PG-Net[14]启发，在特征融合时，通过构建相似匹配网络(Similar Matching Network，SM)分解模板分支输出特征来缩小匹配区域，避免产生过多的干扰响应点，抑制背景干扰，提高算法判别力。3)在通用数据集上做消融实验和对比实验，证明本文算法的有效性。通过对现有问题的探索与改进，算法的跟踪鲁棒性得到有效提高，且面对相似干扰、遮挡等复杂场景时拥有更好的跟踪性能。

1　联合高阶目标感知与相似匹配的目标跟踪

1.1　基本原理

本文算法框架如图 1 所示。在模板分支中，将第 $t-1$ 帧作为初始帧，利用 ResNet50 网络提取特征，随后引入高阶目标感知模型优化高阶注意权重间的相关性，在空间和通道维度上获取高阶目标感知特征；在搜索分支中，将第 $t\sim(t+T)$ 帧作为待跟踪序列帧来提取特征。在特征融合阶段构建相似匹配网络，以更小的区域实现精确的匹配，得到最终响应特征。最后送入分类回归网络，得到分类和回归响应图，并利用中心度得分帮助跟踪目标。

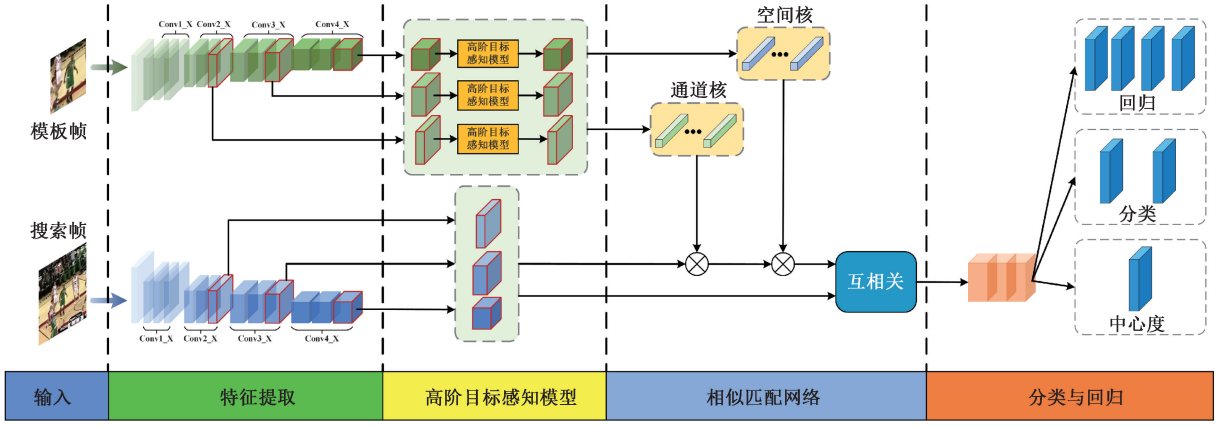

图 1　联合高阶目标感知与相似匹配的跟踪网络框架

1.2　高阶目标感知模型

特征提取部分的 CNN(convolutional neural network)存在感受野形状固定且范围有限，难以捕获全局上下文信息的问题，进而导致难以判别相似目标。为此 DANet[15]引入自注意力机制捕获全局信息，然而通过键值对来获取自注意力权值是独立的，这可能导致注意力分散，使键值对发生计算偏差。为纠正这种负面影响，本文在特征提取网络中嵌入高阶目标感知模型(HTA)，结构如图 2 所示。本文首先计算初始注意力权值 $\boldsymbol{A}$，对原始特征内部像素间进行建模；随后借鉴极化过滤思想[16]，采用空间-通道正交方式对注意力权值 $\boldsymbol{A}$ 进行建模，获取仅关注空间和仅关注通道的高阶注意力权值 $\boldsymbol{A}_S^2$ 和 $\boldsymbol{A}_C^2$，用于获取高阶上下文关系。初始一阶注意力权值可以动态调整输入特征中像素间的关系，而本文高阶注意力权值又可以调整一阶注意力

权值的相互作用,避免因模糊性导致目标与非目标间的错误相关。

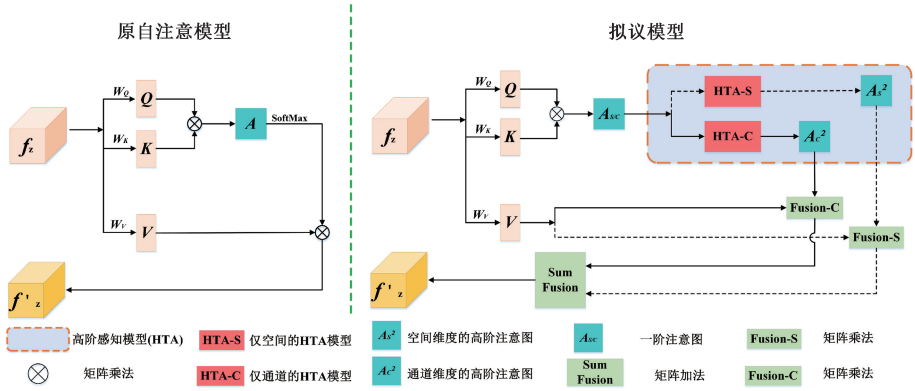

图 2 高阶目标感知模型

如图 2 所示,高阶目标感知模型从通道和空间两个维度对一阶自注意权值进行建模。假定输入高阶目标感知模型的初始特征为 $f_z \in R^{C\times H\times W}$,首先经过 3 组 1×1 卷积(即 W_Q,W_K,W_V)的线性变换,将原始输入特征变换为 $\boldsymbol{Q}$,$\boldsymbol{K}$,$\boldsymbol{V} \in \mathbb{R}^{HW\times C}$;随后计算 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 之间的矩阵相似度得到一阶注意力权值 $\boldsymbol{A}_{S/C}$(其中 $\boldsymbol{A}_S \in \mathbb{R}^{HW\times HW}$,$\boldsymbol{A}_C \in \mathbb{R}^{C\times C}$),其公式如下:

$$\begin{cases}\boldsymbol{A}_S = \alpha \boldsymbol{Q}\boldsymbol{K}^T \\ \boldsymbol{A}_C = \beta \boldsymbol{Q}^T\boldsymbol{K}\end{cases} \tag{1}$$

式中:α 和 β 是用于对抗数值爆炸的尺度因子。然后,将注意力权值 $\boldsymbol{A}_S$ 和 $\boldsymbol{A}_C$ 作为输入,送入 HTA 模型中用于计算高阶注意力权值,以引导一阶注意力权值更好的调整原始输入特征中像素间的关系,如图 3 所示。

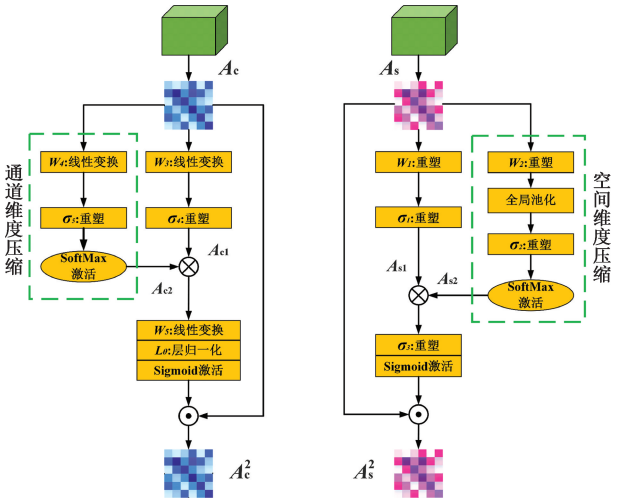

图 3 高阶注意力图获取

在高阶感知模型(HTA)中,采用正交的方式,将不同的一阶注意力权值送入两个不同的分支中:$\boldsymbol{A}_S$ 被送入仅空间感知模型(high order target aware model in spatial dimension,HTA-S)的分支,在空间维度帮助其向良好的方向调整;而 $\boldsymbol{A}_C$ 被送入仅通道感知模型(high order target aware model in channel dimension,HTA-C)的分支,用于更好的优化通道关联性,这个过程被表示为非线性映射 φ:input → output。在 HTA-S 分支中,首先通过线性变换 W_1 和重塑 σ_1 将 $\boldsymbol{A}_S$ 变为 $\boldsymbol{A}_{S1} \in \mathbb{R}^{(D/2)\times D}$(其中 $D = HW$),目的是将特征的空间维度保持在高分辨率水平,并适当缓解计算效率的增长趋势;同时通过线性变换 W_2、全局池化 $F_{gp}(\cdot)$ 以及重塑 σ_2 将 $\boldsymbol{A}_S$ 转换为 $\boldsymbol{A}_{S2} \in \mathbb{R}^{1\times(D/2)}$;最后,对 $\boldsymbol{A}_{S1}$ 和 $\boldsymbol{A}_{S2}$ 执行矩阵乘法,通过 σ_3 重塑和 Sigmoid 函数将参数维持在 0~1 之间,并和 $\boldsymbol{A}_S$ 拼接得到最终的 $\boldsymbol{A}_S^2 \in \mathbb{R}^{HW\times HW}$。基于空间维度的高阶注意力计算中,通过保持空间维度高分辨率并折叠压缩通道维度,实现了极化过滤;组合非线性拟合函数,用 SoftMax 函数 $F_{sm}(\cdot)$ 增加动态关注范围,随后用 Sigmoid 函数 F_{sig} 动态映射,对压缩造成的损失强度范围进行信息增强,使此模型具备更强的拟合能力,实现特定空间的加权来感知相同语义的像素。完整计算过程如下:

$$\begin{cases}\boldsymbol{A}_S^2 = F_{sig}\{\sigma_3[F_{sm}(\sigma_2(F_{gp}(W_2\cdot\boldsymbol{A}_S)))\times\sigma_1(W_1\cdot\boldsymbol{A}_S)]\} \\ \boldsymbol{A}_C^2 = F_{sig}\{L_\theta[W_5(\sigma_4(W_3\cdot\boldsymbol{A}_C))\times F_{sm}(\sigma_5(W_4\cdot\boldsymbol{A}_C))]\}\end{cases} \tag{2}$$

不同于 HTA-S 分支,HTA-C 分支仅关注通道维度,对特定的通道进行加权以输出最佳分数。如式(2)所示,首先利用线性变换 W_3 和重塑操作 σ_4 将 $\boldsymbol{A}_C$ 转换为 $\boldsymbol{A}_{C1} \in \mathbb{R}^{(C/2)\times C}$,同时利用线性变换 W_4 和重塑 σ_5 将转换为 $\boldsymbol{A}_{C2} \in \mathbb{R}^{C\times 1}$;随后信息压缩的 $\boldsymbol{A}_{C2}$ 利用 SoftMax 进行增强;最后,通过线性变换 W_5 和 L_θ 将通道数恢复并做层归一化以捕获权值矩阵中的关系,并利用 Sigmoid 进行非线性拟合得到 $\boldsymbol{A}_C^2$。因此,高阶目标感知模型的总体可以表达为:

$$\mathrm{HTA}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = [\mathrm{SoftMax}(\boldsymbol{A}_S{}^2) + \mathrm{SoftMax}(\boldsymbol{A}_C{}^2)]V \tag{3}$$

1.3 相似匹配网络建模

传统孪生跟踪器采用相似性度量的方法,将模板特征视为卷积核,对搜索特征卷积实现匹配;但真实匹配区域

大于理想区域，因而运算时易引入背景噪声，掩盖目标特征信息，增大区分干扰的难度。因此，本文利用相似匹配网络缩小匹配的有效区域，优化互相关过程，通过迭代搜索缩小匹配区域实现相似计算。

1)模板特征分解

如图 4 所示，对模板分支的感知特征进行分解以减小匹配时的核尺寸，利用局部特征突出目标的细节信息。感知特征 f_z 在空间维度划分为 n_s 个 $1\times1\times c$ 的空间核，所有空间核集合用 $k_{zs}=\{k_{zs}^1,k_{zs}^2,\cdots,k_{zs}^{n_s}\},k_{zs}\in Z^{n_s\times(1\times1\times c)}$ 表示。同时将模板特征在通道维度划分为 c 个 $1\times1\times n_s$ 的通道核，通道核集合表示为 $k_{zc}=\{k_{zc}^1,k_{zc}^2,\cdots,k_{zc}^c\},k_{zc}\in Z^{c\times(1\times1\times n_s)}$。

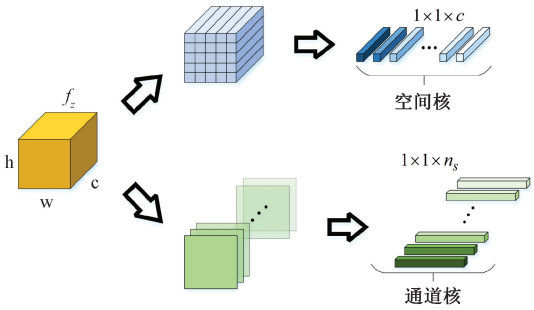

图 4　特征分解模块

2)建立相似匹配模型

相似匹配部分将分解后的感知特征与搜索特征计算相似度，结构如图 5 所示。

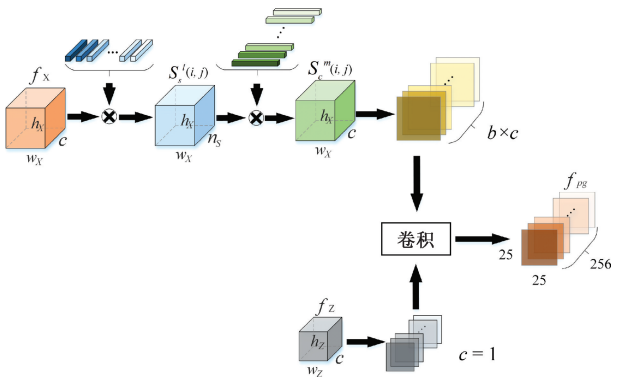

图 5　相似匹配模块

首先将空间核 k_{zs} 与搜索特征 $f_X(i,j)$ 逐像素计算，获得空间相似度：

$$S_s^l(i,j)=\psi(f_X(i,j),k_{zs}^l),\ l=1,2,\cdots,n_s \tag{4}$$

其中，$f_X(i,j)$ 为搜索特征的第 i 行第 j 列位置的像素特征；k_{zs}^l 为感知特征的第 l 个空间核；$\psi(,)$ 表示高维张量乘法；$S_s^l(i,j)$ 表示第 (i,j) 位置的搜索特征与第 k_{zc} 个空间核 k_{zs} 的空间相似性。

由于分解空间核 k_{zs} 仅关注模板每个小区域上的有效信息而忽略两分支的整体关联性，因此进一步利用通道核 k_{zc}，获得与感知特征之间的通道相似性，并保证有效信息的完整性。用 $S_s^l(i,j)$ 逐像素与通道核 k_{zc} 计算相似性度量 S_c，公式如下：

$$S_c^m(i,j)=\psi(S_s^l(i,j),k_{zc}^m),m=1,2,\cdots,c \tag{5}$$

其中，k_{zc}^m 表示感知特征第 m 个通道核；$\psi(,)$ 表示张量乘法；$S_s(i,j)$ 为通道相似信息，表示与第 m 个通道核的相似度。

对于式(4)、(5)中搜索区域的每个像素特征，计算其与感知特征的相似度，使复杂背景下的目标具有很好的判别性，且尽可能多的利用有效目标前景区域来帮助训练，对目标的边界信息足够友好。通过特征分解和相似匹配过程，得到更理想的响应特征。整个计算过程用式(6)表示：

$$f_{pg}=\left(\sum_{m=1}^{c}\sum_{l=1}^{n_s}f_X(i,j)\otimes k_{zs}^l\otimes k_{zc}^m\right)*f_Z \tag{6}$$

式中：$*$ 表示卷积运算，f_z 表示模板感知特征，f_{pg} 表示经过特征分解和匹配之后的输出特征。

1.4　训练损失

本文分类分支预测类别信息，回归分支计算对应位置边界框。首先，对于经过感知与匹配的输出特征 $F\in\mathbb{R}^{C\times H\times W}$，其特征图上的每个点 $A(i,j)$ 均可映射回搜索区域。由于分类和回归任务不同，将特征 F 经过两种不同卷积得到分类特征图 $F_{cls}\in\mathbb{R}^{2\times H\times W}$ 和回归特征图 $F_{reg}\in\mathbb{R}^{4\times H\times W}$，分类特征图包含前景和背景的类别信息，而回归特征图计算映射点到搜索区域边界框的边距信息（l、r、t、b）。公式如下：

$$\begin{cases}l=x_0-x_a,t=y_0-y_a\\r=x_b-x_0,b=y_b-y_0\end{cases} \tag{7}$$

式中：(x_a,y_a)，(x_b,y_b) 分别表示真实目标边界框的左上、右下角点坐标，(x_0,y_0) 表示搜索区域上的预测位置。此外，对于样本的判定，有以下定义：

$$S_{(i,j)}=\begin{cases}1, & l,r,t,b>0\\0, & \text{其他}\end{cases} \tag{8}$$

可以看到，当 l、r、t、b 均大于零时，预测点落入边界框内，此时判定为正样本，否则为负样本。因此，对于回归损失如下：

$$Loss_{reg}=\frac{1}{S_{(i,j)}}\sum S_{(i,j)}L_{IOU}(F_{reg}(l,r,t,b)) \tag{9}$$

对于分类分支采用交叉熵损失作为分类损失。考虑预测点到目标中心的距离会影响边界框预测质量，因此使用了中心损失来抑制过大的位移。中心度特征图 $F_{cen}\in\mathbb{R}^{1\times H\times W}$ 同样使用一组卷积得到，相应的中心得分可以表示为：

$$C_{(i,j)}=S_{(i,j)}\times\sqrt{\frac{\min(l,r)\times\min(t,b)}{\max(l,r)\times\max(t,b)}} \tag{10}$$

由公式可以看出，当预测位置落入背景时，中心度得分为零。因此，中心损失如下：

$$L_{cen}=\frac{-1}{S_{(i,j)}}\sum_{S_{(i,j)}=1}C_{(i,j)}\times\log F_{cen}+(1-C_{(i,j)})\times\log(1-F_{cen}) \tag{11}$$

最后,整体损失为:

$$L_{total}=\lambda_1\times L_{cls}+\lambda_2\times L_{reg}+\lambda_3\times L_{cen} \tag{12}$$

其中,L_{cls} 为交叉熵损失,且超参数 λ_1、λ_2、λ_3 分别为 1、3、1。最后,本文算法在训练阶段的 3 种损失函数以及总损失函数曲线如图 6 所示。

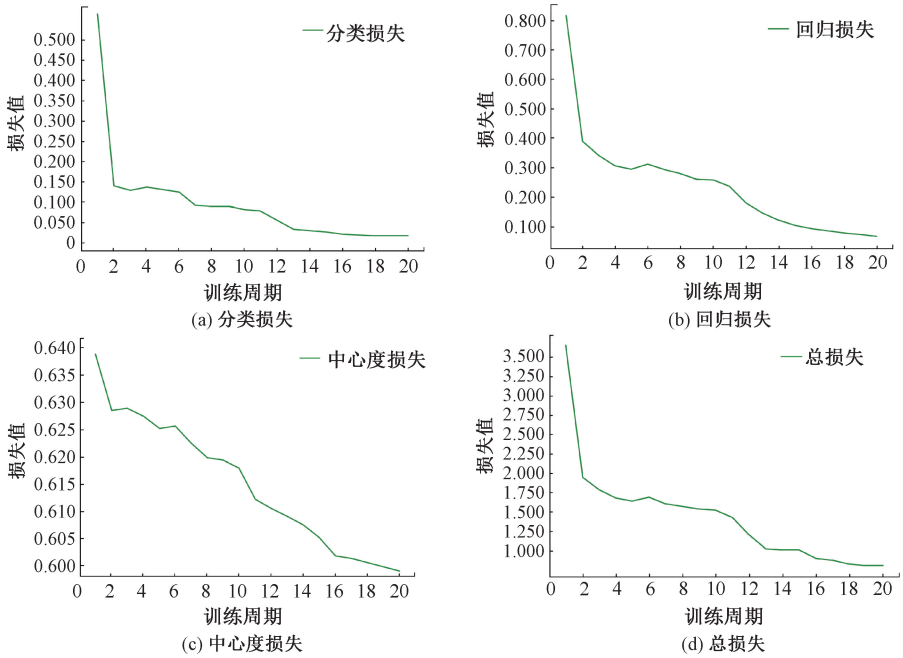

图 6 训练损失函数曲线图

2 实验结果与分析

2.1 实验平台与参数设定

本实验的硬件配置采用 E5-1650 V4 处理器,内存 32 G;GPU 为 11 G 显存 GTX 1080Ti;系统为 Ubuntu18.04,PyTorch1.2.0 框架,python 3.6.9 版本。本文在 GOT-10K[17] 和 ILSVRC2015-VID[18] 数据集上离线训练,训练批大小为 16,使用随机梯度下降 SGD。对前 5 个 epoch 使用预热训练,初始化学习速率为 2×10^{-4},最终学习率升为 1×10^{-3};从第 6 个 epoch 开始,初始学习率为 1×10^{-3},最终学习率下降为 1×10^{-4},动量设置 0.9;总迭代 20 个周期。

2.2 OTB100 基准评估算法性能

文中算法与几种主流跟踪算法 CFNet[19]、SiamFC[20]、SiamDWfc[21]、SiamRPN[5]、DaSiamRPN[12]、GradNet[22] 等在 OTB100[23] 跟踪基准上进行实验对比并评估其性能。

1)定量分析

(1)跟踪成功率和准确率是目标跟踪中两种评估指标,代表覆盖率和中心位置误差。图 7 展示各算法的成功率和准确率。由图可知,本文算法准确率取得 87.0%的次优成绩;而跟踪成功率取得 66.0%的最优成绩,比 DaSiamRPN 算法高出 0.2%。较基准算法在准确率上高出 3.6%,而成功率高出基准算法 1.8%,跟踪更加稳定。

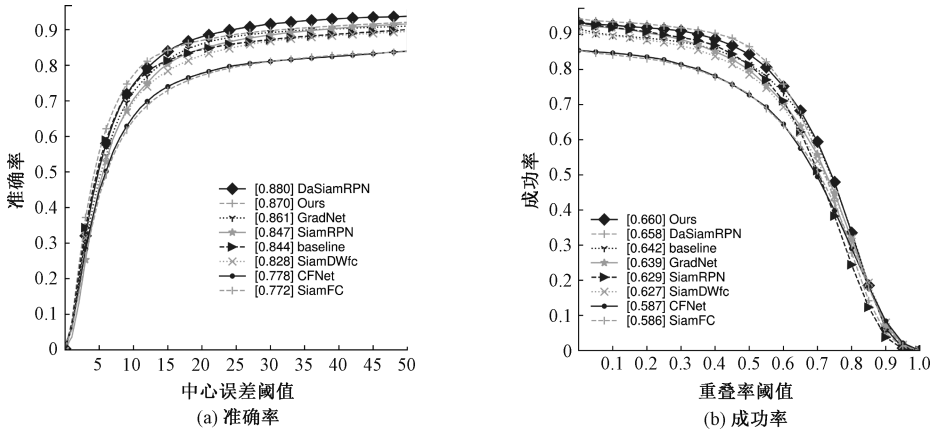

图 7 OTB100 数据集跟踪曲线图

(2)本文利用 OTB 中 11 种属性用于评估算法性能，如表 1、2 所示。本文算法成功率、准确率在多种属性上取得最高得分和次优得分。尤其针对背景干扰、遮挡、快速运动等任务取得优异性能。

表 1　OTB100 视频属性准确率得分

算法	视频属性										
	背景杂波	形变	快速移动	遮挡	超出视野	尺度变化	光照变化	平面内旋转	运动模糊	低分辨率	平面外旋转
GradNet[22]	0.822	0.795	0.838	**0.838**	0.789	0.841	0.844	0.860	0.855	**0.999**	0.872
SiamFC[20]	0.692	0.691	0.744	0.723	0.673	0.736	0.736	0.743	0.707	0.900	0.673
DaSiamRPN[12]	**0.856**	**0.878**	0.818	0.811	0.717	0.852	0.869	0.886	0.819	0.937	0.863
CFNet[19]	0.756	0.714	0.705	0.699	0.601	0.731	0.707	0.786	0.680	0.888	0.759
SiamDWfc[21]	0.762	0.763	0.808	0.798	0.781	0.819	0.794	0.824	0.841	0.901	0.829
SiamRPN[5]	0.799	0.825	0.789	0.780	0.726	0.838	0.859	0.854	0.816	0.978	0.851
SiamCAR[13]	0.768	0.829	0.823	0.788	0.786	0.842	0.828	0.846	0.876	0.829	0.819
Proposed	0.842	0.855	**0.873**	0.822	**0.793**	**0.862**	**0.878**	**0.889**	**0.881**	0.842	**0.877**

注：黑体为最优，下划线为次优。

表 2　OTB100 视频属性成功率得分

算法	视频属性										
	背景杂波	形变	快速移动	遮挡	超出视野	尺度变化	光照变化	平面内旋转	运动模糊	低分辨率	平面外旋转
GradNet[22]	0.611	0.571	0.624	0.615	0.583	0.614	0.643	0.627	0.645	**0.669**	0.628
SiamFC[20]	0.527	0.512	0.571	0.549	0.509	0.556	0.575	0.559	0.554	0.618	0.561
DaSiamRPN[12]	**0.642**	**0.645**	0.621	0.611	0.537	0.637	0.655	0.652	0.625	0.636	0.634
CFNet[19]	0.561	0.526	0.554	0.527	0.454	0.546	0.541	0.567	0.540	0.614	0.533
SiamDWfc[21]	0.574	0.560	0.630	0.601	0.590	0.613	0.622	0.606	0.654	0.596	0.612
SiamRPN[5]	0.591	0.617	0.599	0.585	0.542	0.615	0.649	0.628	0.622	0.639	0.625
SiamCAR[13]	0.578	0.612	0.638	0.593	0.586	0.641	0.647	0.635	0.681	0.603	0.607
Proposed	0.632	0.627	**0.671**	**0.616**	**0.595**	**0.660**	**0.681**	**0.668**	**0.685**	0.596	**0.641**

注：黑体为最优，下划线为次优。

综上所述，本文算法满足实时性要求，在多种属性上的成功率和准确率取得最优和次优得分，说明在 OTB100 评估基准上，本文算法凭借感知模型和良好的匹配策略，更好的应对快速运动、遮挡等场景。

2)定性分析

如图 8 所示，本文选取 4 个代表性跟踪序列来展示跟踪效果，并做如下分析：

(1)背景杂波及相似干扰挑战。如图 8(a)所示，Basketball 序列在前期、中期和后期多次出现相似球员干扰情况；由此可知，高阶目标感知模型能优化目标的深层语义信息和空间位置信息，增强判别能力；同时优化匹配策略，准确区分目标和其他背景信息，从而保持良好跟踪。

(2)目标遮挡挑战。如图 8(b)所示，从 Jogging 序列第 71 帧开始，目标在正常运动时出现遮挡情况。当完全遮挡时，本文算法能利用高阶感知捕获全局上下文信息，对目标特征有较强判别力，并对目标变化拥有较好鲁棒性，进而成功预测目标位置并估计目标大小。

(3)光照变化挑战。如图 8(c)所示，Singer2 序列中展示各算法在光照环境变化下的状态，在视频序列的前期、中期和后期，大部分算法因背景灯光影响开始出现跟踪漂移现象，仅有本文算法和 SiamRPN、DaSiamRPN 算法可以稳定捕获目标信息。当光照突变影响目标时，本文算法能够利用高阶感知模型和匹配网络准确捕获目标像素的变化，进而区分目标与干扰因素，抑制背景光照变化影响。

(4)快速运动挑战。如图 8(d)所示，纵观 Ironman 序列，本文算法跟踪状态更稳定。当发生较快运动时，本文算法能够从快速变化的区域中提取目标有效信息，因而在复杂场景中更好地适应目标快速运动。

2.3　面向无人机的跟踪性能评估

1)定性分析

针对不同属性的视频序列，用表 3、4 展示 5 种对比算法在 12 种属性上的结果。

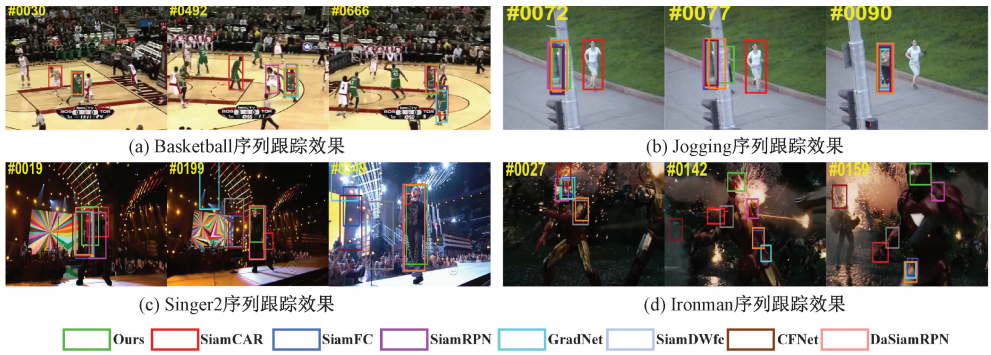

图 8 跟踪效果可视化

表 3 UAV123 视频属性准确率得分

算法	视频属性											
	光照变化	尺度变化	快速运动	超出视野	背景杂波	低分辨率	宽高比率	相机运动	部分遮挡	相似目标	完全遮挡	视点变化
ECO[22]	0.628	0.673	0.537	0.564	0.626	0.657	0.632	0.676	0.638	0.704	0.542	0.631
DaSiamRPN[12]	0.710	0.754	0.737	0.693	0.597	0.663	0.756	0.786	0.701	0.747	0.633	0.753
SiamRPN[5]	0.707	0.744	0.690	0.708	0.590	0.637	0.738	0.787	0.673	0.705	0.559	0.768
SiamCAR[13]	0.762	0.781	0.742	0.717	**0.677**	**0.695**	0.745	0.787	0.705	0.717	0.628	0.784
Proposed	**0.769**	**0.801**	**0.762**	**0.756**	0.670	0.681	**0.763**	**0.818**	**0.729**	**0.778**	**0.691**	**0.824**

注:黑体为最优,下划线为次优。

表 4 UAV123 视频属性成功率得分

算法	视频属性											
	光照变化	尺度变化	快速运动	超出视野	背景杂波	低分辨率	宽高比率	相机运动	部分遮挡	相似目标	完全遮挡	视点变化
ECO[22]	0.407	0.465	0.342	0.406	0.373	0.362	0.425	0.476	0.432	0.490	0.292	0.431
DaSiamRPN[12]	0.500	0.544	0.520	0.509	0.407	0.411	0.537	0.581	0.493	0.517	0.379	0.563
SiamRPN[5]	0.520	0.556	0.502	0.526	0.406	0.419	0.541	0.593	0.483	0.507	0.341	0.587
SiamCAR[13]	0.573	0.596	0.549	0.543	**0.466**	**0.466**	0.568	0.604	0.517	0.533	0.394	0.627
Proposed	**0.577**	**0.610**	**0.569**	**0.582**	0.444	0.450	**0.583**	**0.626**	**0.533**	**0.577**	**0.438**	**0.654**

注:黑体为最优,下划线为次优。

(1)本文算法虽然在完全遮挡属性上的准确率略低于对比算法 DaSiamRPN,但同属性的成功率最高;其次,针对部分遮挡属性的无人机视频,本文算法保持良好的跟踪成功率和准确率。未做模板更新的情况下,本文算法凭借高阶目标感知模型和相似匹配网络的优势,依然可以很好的对无人机目标进行跟踪。

(2)面对相似目标干扰,本文算法的跟踪成功率和准确率均优于对比算法和基线算法。本文算法对提取到的特征进行优化,增强模型的判别能力,并利用小区域的匹配机制准确找到无人机目标的响应位置,从而达到抑制相似物干扰的目的。

2)定量分析

为验证本文算法在无人机领域中的性能,同 ECO[24]、SiamRPN[5]、DaSiamRPN[12] 和 SiamCAR[13] 等算法在 UAV123[25] 做对比试验。实验结果如表 5 所示。

表 5 各算法在 UAV123 标准下的评估得分

策略	跟踪器	UAV123	
		准确率(%↑)	成功率(%↑)
Correlation filter	ECO[24]	0.688	0.525
Anchor based	DaSiamRPN[12]	0.781	0.569
	SiamRPN[5]	0.772	0.581
Anchor free	SiamCAR[11]	0.785	0.583
	Proposed	0.796	0.607

表 5 为各对比算法在 UAV123 标准下的评估结果。由表可知，本文算法相较于基线，在成功率指标上提升了 2.4%，在准确率指标上提高 1.1%；同时，本文算法相对于其他算法均有不同程度的优势。相对于其他对比算法，以成功率指标为例：相对于 ECO、SiamRPN、DaSiamRPN 算法，本文算法分别提高 10.8%、2.4%和 1.5%的性能。

2.4　消融实验

为证明本文算法有效性，用 GOT-10k 的 500 个序列用于训练模型，并在 OTB100 上做消融实验。

如表 6 所示，对于高阶目标感知模型，仅空间或仅通道的高阶感知方式都能在一定程度上提升跟踪性能，证明其可行性。然而，当通过不同连接方式对两种感知子模型进行组合时发现，串联方式并不能对算法的跟踪性能产生积极影响，主要原因是极化过滤的方式在挖掘空间(或通道)维度信息时，完全压缩正交的通道(或空间)维度而丢失通道相关信息，导致在之后的串联法中无法捕获通道(空间)维度的联系，因而只能通过并联的方式进行高阶目标感知模型的设计，达到高算法性能的目的。

表 6　不同组合方式在 OTB100 数据集上的实验结果

组合方式	准确率(%↑)	成功率(%↑)
baseline	0.702	0.471
HTA-S	0.743	0.499
HTA-C	0.750	0.495
HTA-S/C	0.705	0.469
HTA-S//C	0.769	0.514

其中 HTA-S 为仅对空间维度做处理，HTA-C 仅对通道维度处理；HTA-S/C (high order target aware model in series mode)和 HTA-S//C (high order target aware model in parallel mode)分别表示串联和并联方式的高阶目标感知模型。最终可以通过表 6 看出，在并联法的高阶目标感知模型下，跟踪模型的性能提升 4.3%的跟踪成功率和 6.7%的跟踪准确率。SMN 为使用相似匹配网络进行互相关。

最后，对算法的不同模块在 OTB100 上进行消融实验，如表 7 所示，其中 HTA(即表 5 中的 HTA-S//C)指在原有基础上融入高阶目标感知模型；SMN 指在原有基础上使用相似匹配模型缩小了匹配区域。利用相似匹配网络对搜索特征和优化后的高阶感知特征计算相似性，相比基线分别在成功率和准确率上提高 9.9%和 8.2%，证明本文所提方法的有效性。

表 7　算法不同模块在 OTB100 上的消融实验

方法	准确率(%↑)	成功率(%↑)	帧率
baseline	0.702	0.471	52
Baseline＋HTA	0.769	0.514	44
Baseline＋SMN	0.746	0.490	41
Baseline＋HTA＋SMN	0.801	0.553	37

3　结　　论

本文提出一种基于孪生网络的端到端跟踪算法，通过高阶目标感知模型提高位置信息和深层语义信息的感知能力，并利用相似匹配网络精细化匹配区域，提高算法性能，实现对单目标的跟踪任务。通过在 OTB100 和 UAV123 跟踪基准的实验表明，在满足实时性的前提下，本文算法提高了对通用单目标跟踪的成功率和准确率，同时可以很好的缓解相似目标干扰、遮挡等问题。

参考文献

[1] MARVASTI-ZADEH S M, CHENG L, GHANEI Y H, et al. Deep learning for visual tracking: A comprehensive survey [J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(5): 3943-3968.

[2] 郭秋蕊，李建良，田垚，等. 基于改进 KCF 算法和多特征融合的车辆跟踪研究[J]. 电子测量与仪器学报，2022，36(4)：231-240.

[3] 董美琳，任安虎. 基于深度学习的高速公路交通事件检测研究[J]. 国外电子测量技术，2021，40(10)：108-116.

[4] 孟琭，杨旭. 目标跟踪算法综述[J]. 自动化学报，2019，45(7)：1244-1260.

[5] LI B, YAN 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, 2018: 8971-8980.

[6] 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, 2019: 4282-4291.

[7] 杨梅，贾旭，殷浩东，等. 基于联合注意力孪生网络目标跟踪算法[J]. 仪器仪表学报，2021，42(1)：127-136.

[8] PU L, FENG X, HOU Z, et al. SiamDA: Dual attention Siamese network for real-time visual tracking[J]. Signal Processing: Image Communication, 2021, 95, DOI: 10.1016/j.image.2021.116293.

[9] ZHANG X, MA J, LIU H, et al. Dual attentional siamese network for visual tracking [J]. Displays, 2022, 74, DOI: 10.1016/j.displa.2022.102205.

[10] WANG X, TANG J, LUO B, et al. Tracking by joint local and global search: A target-aware attention-based approach[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(11): 6931-6945.

[11] 付谱平，叶俊. 融合语义特征网络的孪生网络目标跟踪算法[J]. 电子测量技术，2022，45(8)：136-142.

[12] ZHU Z, WANG Q, LI B, et al. Distractor-aware siamese networks for visual object tracking [C].

Proceedings of the European conference on computer vision(ECCV),2018: 101-117.

[13] GUO D Y, WANG J, CUI Y, et al. SiamCAR: Siamese fully convolutional classification and regression for visual tracking[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 6269-6277.

[14] LIAO B, WANG C, WANG Y, et al. Pg-net: Pixel to global matching network for visual tracking[C]. European Conference on Computer Vision. Springer, Cham, 2020: 429-444.

[15] FU J, LIU J, TIAN H, et al. Dual attention network for scene segmentation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 3146-3154.

[16] LIU H, LIU F, FAN X, et al. Polarized self-attention: Towards high-quality pixel-wise mapping [J]. Neurocomputing, 2022, 506(28): 158-167.

[17] HUANG L, ZHAO X, HUANG K. Got-10k: A large high-diversity benchmark for generic object tracking in the wild[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 43(5): 1562-1577.

[18] RUSSAKOVSKY O, DENG J, SS H, et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[19] VALMADRE J, BERTINETTO L, HENRIQUES J, et al. End-to-end representation learning for correlation filter based tracking[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2805-2813.

[20] BERTINETTO L, VALMADRE J, HENRIQUES J F, et al. Fully-convolutional siamese networks for object tracking[C]. Proceedings of European Conference on Computer Vision, 2016: 850-865.

[21] ZHANG Z P, PENG H W. Deeper and wider siamese networks for real-time visual tracking[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 4591-4600.

[22] LI P X, CHEN B Y, OUYANG W L, et al. GradNet: Gradient-guided network for visual object tracking [C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 6162-6171.

[23] WU Y, LIM J, YANG M H. Object tracking benchmark [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 37(9): 1834-1848.

[24] DANELLJAN M, BHAT G, SHAHBAZ K F, et al. Eco: Efficient convolution operators for tracking [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 6638-664.

[25] MUELLER M, SMITH N, GHANEM B. A benchmark and simulator for uav tracking [C]. Proceedings of European Conference on Computer Vision, 2016: 445-461.

作者简介

张念超,硕士研究生,主要研究方向为计算机视觉、目标跟踪。

E-mail: zhangnc_imust@163.com

张宝华(通信作者),教授,硕士生导师,主要研究方向为智能图像处理、目标检测与跟踪、行人重识别等。

E-mail: zbh_wj2004@imust.cn。

李永翔,副教授,主要研究方向为智能交通等。

谷宇,副教授,硕士生导师,主要研究方向为计算机视觉、医学图像处理、计算机辅助检测。