arXiv:2301.09063v1 [cs.CV] 22 Jan 2023

DASTSiam：用于提升孪生跟踪性能的时空融合与判别性增强

Yucheng Huang[1*], **Eksan Firkat**[1*], **肖子旺**[1], **Jihong Zhu**[1,2], **Askar Hamdulla**[1 *†]新疆大学信息科学与工程学院，中国乌鲁木齐[2]清华大学精密仪器系，中国北京
{107552101321,eksan, 107552103759}@stu.xju.edu.cn.com,
jhzhu@tsinghua.edu.cn, askar@xju.edu.cn

摘要

随着孪生跟踪器的出现，基于深度神经网络的跟踪任务取得了显著进展。然而，跟踪过程中目标外观常会发生变化，这会降低跟踪器应对宽高比变化、遮挡和尺度变化等挑战时的鲁棒性。此外，杂乱背景可能导致响应图中出现多个高响应点，从而造成目标定位错误。本文提出两个基于Transformer的模块，以改进孪生跟踪，称为**DASTSiam**：时空融合模块和判别性增强模块。时空模块通过交叉注意力累积历史线索，增强模型应对目标外观变化的鲁棒性；判别性增强模块则关联模板与搜索区域之间的语义信息，以提升目标判别能力。此外，修改锚框的标签分配方式也能提高目标定位的可靠性。我们的模块可用于所有孪生跟踪器，并通过对比实验和消融实验，在多个公开数据集上展现出性能提升。代码见 *https://github.com/huangliqwe2020/DASTASiam.*

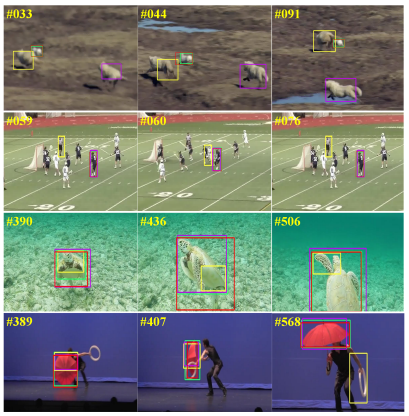


图1：**定性比较**。借助时空模块（ST）和判别性增强（DA），我们的**DASTSiam** 能够应对外观变化、运动模糊和背景杂乱。此外，其性能优于 SiamRPN 和 SiamRPN++。

1 引言

目标跟踪是计算机视觉中的一项基础问题，在无人机、精确制导、移动机器人和视频监控等领域有着广泛的应用。基于孪生网络的跟踪算法是目标跟踪中最广泛使用且前景广阔的方法之一 [Bertinetto 等，2016年；Li 等，2018a；Guo等，2017年；Zhang 等，2018年；Zhang 和 Peng，2019年；Wang等，2019年；Li 等， 2019]。这些算法采用双分支神经网络，将表示目标的模板图像和包含目标的搜索区域图像输入深度卷积神经网络骨干网络，以跟踪目标进行特征提取。随后，对提取的特征应用相关函数，生成指示目标在搜索区域中位置的响应图。近期研究主要关注如何提升基于孪生网络的跟踪算法的性能。例如，SiamFC [Bertinetto 等， 2016] 指的是 DCF 跟踪器 [Henriques 等，2014年；Danelljan 等，2017年；Valmadre 等，2017年；Danelljan 等，2015年；Li 等，2018b] ，并使用孪生网络训练相关滤波器；SiamRPN [Li 等，2018a] 指的是 Fast-RCNN[Gir- shick， 2015] 并提出了基于区域提议网络的锚点设置。这些方法在准确性和实时性能之间实现了良好的平衡。

尽管上述孪生跟踪器已取得出色的跟踪性能，但它们对时空信息的利用仍不充分。在 SiamFC 和 SiamRPN 中，初始模板不会更新；因此，当目标外观在跟踪过程中发生变化时，可能导致跟踪失败。为解决这一问题，近期的孪生跟踪器采用了线性更新技术[Li 等，2018a；Zhu 等，2018] 根据新帧中的外观修改模板。然而，在复杂环境中——

*Yucheng Huang 和 Eksan Firkat 贡献相同。
†Askar Hamdulla 是通讯作者。

存在光照变化和遮挡等因素时，学习能够处理此类变化的自适应目标描述符十分重要。杂乱背景也是孪生跟踪器面临的常见挑战，因为它们很容易在最终响应图中产生多个高响应点，导致跟踪器由于缺乏能够将目标与干扰区分开的深层语义信息而无法准确定位目标。近期的研究方法，如 DaSi-amRPN [Zhu 等， 2018] 以及 RASNet [Wang 等， 2018]分别通过在训练样本中构建语义干扰的负样本集，以及持续学习以提高描述目标语义信息的通道权重，尝试解决这一问题。然而，这些方法面临额外构建样本的困难，且未能充分利用深层特征空间中目标与干扰之间的语义差异。

为了解决这些问题，受 UpdateNet[Zhang 等人以及一些基于 Transformer 的视觉任务[Parmar等人，*2018*年；*Carion* 等人，*2020*年；*Zeng* 等人 2020] 的启发，我们提出了一种利用 Transformer 自注意力机制[Vaswani等人 2017] 来增强孪生跟踪的新方法，称为**DASTSiam**。我们的方法解决了现有孪生跟踪器在利用时空信息以及从杂乱背景中区分目标方面的局限性。为提高跟踪器应对目标外观变化的鲁棒性，我们引入了一个轻量级时空（ST）模块。该模块连接彼此孤立的视频帧，并在帧间传递丰富的时间信息。为增强跟踪器从杂乱背景中区分目标的能力，我们提出了一个判别性增强（DA）模块。该模块强化搜索区域特征内部语义相似性特征与交叉注意力机制之间的关联，以利用目标模板与搜索区域之间的语义相似性。我们还在所提方法中加入了两个自适应滤波器，以提高模板更新和改进的解码器输出的可靠性。

总体而言，我们的方法提升了目标模板特征的泛化能力，充分利用了跟踪过程中的时间线索，并增强了在搜索区域中区分目标与背景的能力（如图 1 所示）。我们的贡献总结如下。

- 我们提出了一种轻量时空模块（ST），用于连接彼此孤立的视频帧，并充分利用跟踪过程中的时间线索。该模块增强了跟踪器应对目标形变、遮挡及其他问题的鲁棒性。

- 我们提出了一种判别式增强模块（DA），以增强跟踪过程中区分目标与背景的能力。

- 在公开基准测试上的实验验证了，所提出的 **DASTSiam** 在多个公开数据集上实现了新的最先进性能。

2 方法

本节详细介绍所提出的 **DASTSiam** 框架。该框架包含两个关键组件：时空融合模块和判别式增强模块（DA）。时空融合模块通过修改编码器，提高模板更新对目标外观变化的适应能力。判别式增强模块通过修改解码器，在深层特征空间中建立匹配模板与搜索区域之间的语义关联，从而增强跟踪器的判别能力。图2展示了**DASTSiam**框架的示意图。

2.1 时空（ST）融合模块。

在孪生网络中，模板图像 T_z 和搜索图像 T_x 的特征提取过程由骨干网络完成。提取的特征分别表示为 f_z 和 f_x。对这些特征进行互相关运算，以生成响应图。为了使跟踪系统能够利用时间线索和帧间关系，我们提出在进行互相关运算之前，对 f_z 特征图应用时空融合模块（ST）。该方法的公式如下：

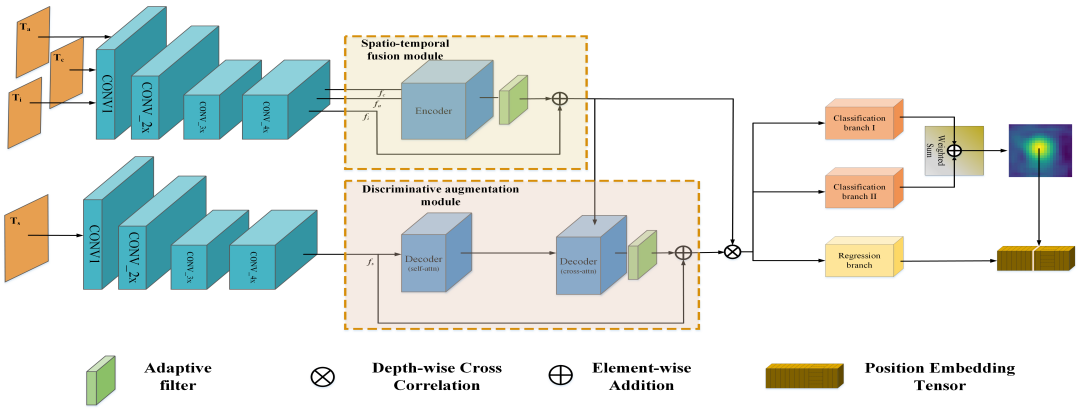
$$ST(f_i, f_a, f_c) = \Phi_{ST}(Encoder(f_a, f_c)) + f_i \tag{1}$$

我们提出的编码器采用注意力机制来扩大感受野并获取更多上下文信息，从而提升时空融合模块（ST）对空间信息的利用率。为高效利用时间线索并提高推理速度，我们采用交叉注意力机制进行多帧融合，而非按通道拼接。我们从同一视频序列中选取三帧，T_i、T_a和 T_c，并提取其特征，分别得到 f_i、f_a和 f_c。随后，我们将 f_a和 f_c 作为编码器的输入，以 f_c 作为查询和值，以 f_a 作为键。特征融合通过交叉注意力机制完成，其数学表示如下：

$$CrossAttn(Q, K, V) = Softmax(\frac{Q_c K_a^T}{\sqrt{d}})V_c \tag{2}$$

交叉注意力机制通过三个不同的全连接层（FC），将置换后的特征图 f_a 和 f_c 映射为 Q_c、V_c和 K_a，其维度为$\mathbb{R}^{\frac{N_t}{2stride} \times \frac{N_t}{2stride} \times C}$。$(\frac{N_t}{2stride})^2$是特征图中的特征点数量，每个点包含 C不同的特征信息，从而增强每个特征点的语义信息。FC 层的权重维度为（N_c， N_c），其中N_c 是通道数。如公式（3）所示， Q_c、V_c和 K_a 中的每个特征点 N_i 都会整合其所有按通道划分的语义信息，极大地丰富特征的空间上下文。

$$\begin{aligned} \hat{N}_{i,j} &= \frac{\sum_{j=0}^{N_c} w_j N_{i,j}}{N_c} \\ N_i &= (N_{i,0}, N_{i,1}, ..., N_{i,j})_{N_c} \\ \hat{N}_i &= (\hat{N}_{i,0}, \hat{N}_{i,1}, ..., \hat{N}_{i,j})_{N_c} \end{aligned} \tag{3}$$



Fig图2：**基于孪生网络的跟踪框架概览，包含我们提出的模块**

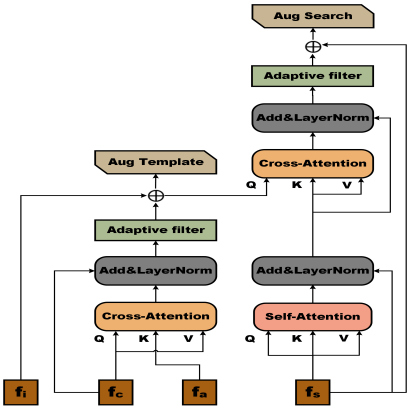


图3：所提出的基于 Transformer 的时空（ST）模块和判别性增强模块（DA）概览。

N_i 指 f_i、f_a、f_c中的特征点。为方便起见，分别将 Q_c、K_a 中的每个特征点标记为 Nc_i 和 Na_i 。为使 ST 能够利用时间信息，我们直接使用矩阵乘法计算 Q_c 和 K_a的注意力矩阵（$AttnMat$）。虽然我们对 Q_c 和 K_a 使用了权重不同的 FC，但其映射规则相同，即可通过 $AttnMat$衡量相同空间位置上 Q_c 与 K_a 之间的相似度。尤为重要的是，f_a 包含丰富且可靠的历史信息，它是 ST 在上一阶段的输出。这样，公式（4）中的 $AttnMat$ 就能使 f_c 更加关注时空相似度较高的位置。

$$\begin{aligned} AttnMat_{i,j} &= P_{i,j}(T_c|T_i, T_{i+1}, ..., T_{c-1}) \\ &= \sum_{j=0}^{N_c} \hat{Nc}_{i,j} \hat{Na}_{i,j}^{\top} \end{aligned} \tag{4}$$

其中，$AttnMat_{i,\ j}$是 $AttnMat$中的一个元素，包含丰富的历史信息。此外，注意力矩阵与 V_c相乘，并通过融合时空信息来调整 f_c 中每个点的值。

时间信息可以增强模板特征，因为其中包含丰富的历史信息和先验知识。

遗憾的是，可能会出现先验误导。具体而言，目标在两个连续帧中的状态变化过大。为了增强最终模板特征的可靠性，我们使用卷积层作为自适应滤波器Φ_{MF} 来纠正先验误导。值得注意的是，自适应滤波的纠正能力有限。如果持续出现极端情况，ST 的输出将失去可靠的特征信息。综合考虑梯度反向传播和特征增强这两个方面，我们继续将初始真实标注帧的模板特征 f_i 添加到 Φ_{MF} 纠正后的输出中，以获得增强的模板特征 f_z^*。这确保了在融合时空信息的同时不会丢失真实目标信息，使我们的跟踪器能够更好地利用时间信息来增强自身的鲁棒性。

2.2 判别性增强模块（DA）

通过改进的编码器获得 f_z^*后，为了提升区分搜索区域 $f_s \in \mathbb{R}^{C\times N_s\times N_s}$ 中目标的能力，我们使用 DA 模块增强搜索区域特征 $f_s \in \mathbb{R}^{C\times \frac{N_s}{stride}\times \frac{N_s}{stride}}$。DA 可表示如下：

$$DA(f_z^*, f_s) = \Phi_{DA}(Decoder(f_z^*, f_s)) + f_s \tag{5}$$

提出改进的解码器，用于在执行互相关运算之前增强 f_s。首先，使用自注意力机制，使 T_s 中的特征点更加关注其他相似的特征点。然后使用交叉注意力机制融合 f_s和 f_z^*。最后生成判别掩码。该掩码的尺寸与 f_s一致。与模板一样，我们使用卷积层Φ_{DA} 来抑制掩码中的干扰信息。最后，将掩码与 f_s 逐元素相加，得到增强后的搜索区域特征f_s^*，并将其用于后续与f_z^* 进行互相关计算，从而获得最终可靠的响应图。图3展示了所提出的两个模块。

2.3 训练与推理

1. **离线训练**。首先，从一个视频序列中随机选取3帧作为不同的模板，即 T_i、T_a、 T_c。经过骨干网络的特征提取后，我们得到f_i、f_a、f_c。在视频中，选取范围为连续50帧。对于 T_i，我们直接从连续50帧中随机选取帧。在预处理的图像增强阶段，为了提高跟踪系统对受损帧的鲁棒性，分别向 T_a和 T_c 中随机添加噪声信息。选择帧时，将随机选取两个连续帧作为 T_a 和 T_c，以确保模块具备融合时间信息的能力。尽可能使推理阶段的执行流程与之相符。

2. **损失函数**。 对于回归，我们采用与 SiamRPN 一致的 smooth-L_1 loss，以预测锚框中心到真实标注中心的归一化距离。对于分类，我们使用两个分支。其中一个与 SiamRPN 相同，采用交叉熵损失来预测响应图中的正负样本。我们还添加了另一个分类分支，该分支采用与 FCOS [Tian 等人,2019] 相关的二元交叉熵（BCE），以提高自适应模板更新的置信度。在 SiamRPN 中，分类分支采用基于交并比（IoU）的正负样本划分方法。对于每个锚框，计算其交并比值。如果交并比值大于预设阈值，则将这些锚框设为正样本，否则设为负样本。根据 [Zhang 等人，2020], 的分析，基于中心距离的标签分配方法比基于交并比的方法能获得更高的平均精度均值，而回归方法对平均精度均值的结果影响很小。在我们的方法中，分类分支采用基于锚框中心点与真实标注中心点之间距离的方案，其形式如下:

$$
\begin{aligned}
c_{pos} &= \phi(C_{pos}) = \frac{C_{pos} - ori}{stride} \\
distance_{x_i,y_i} &= (\frac{c_{y_{lt}} + c_{y_{rb}}}{h} - row_i)^2 \\
&\quad + (\frac{c_{x_{lt}} + c_{x_{rb}}}{w} - col_i)^2
\end{aligned} \tag{6}
$$

C_{pos} 是边界框的角点坐标，其中pos 包含左上角（lt）坐标和右下角（rb）坐标。通过函数 ϕ，我们可以将边界框的原始坐标值转换到分类分支输出的响应特征空间中，而ori 是锚点的预设原点值。得到 c_{pos} 后，我们可以计算映射后的边界框中心点与特征图上各锚框中心点之间的距离，即 $distance_{x_i,\ y_i}$。row_i 表示特征图的第 i 行， col_i 表示特征图的第 i 列。进一步将$distance_{x_i,\ y_i}$ 小于阈值的锚框设为正样本，否则设为负样本。

最终总损失 L_{total} 函数为:

$$
\begin{aligned}
L_{cls} &= \lambda L_{cls1} + L_{cls2} \\
L_{total} &= \lambda_1 L_{cls} + L_{reg}
\end{aligned} \tag{7}
$$

其中， λ 和 λ_1 是需要进行平衡的超参数。

3. **推理**。 f_i 表示初始帧，这是最可靠的参考模板，在整个跟踪过程中始终不会改变； f_c 是跟踪器处理每一帧后更新的当前模板； f_a 是 ST 的输出结果。首先，使用初始帧 T_i 的特征图作为 f_i、 f_a、 f_c，其中 f_i 在初始真实标注帧中被框选。在每个后续帧中，我们根据中心距离对分类分支的预测结果进行加权和，选择置信度最高的目标，并将其发送至 ST，以便在下一步增强我们的模板。

3 实验

3.1 实现细节

1. **方法细节**。 我们使用 SiamFC 和 SiamRPN 作为基础跟踪器。我们未对 SiamFC 的原始版本进行任何修改。对于 SiamRPN，我们将其骨干网络更改为修改后的 ResNet50[He 等， 2016]。同时，我们通过将基于 IoU 的锚点设置更改为基于中心距离的设置，修改了 SiamRPN 的标签分配方式。这简化了处理流程和超参数的设置，并减少了参数数量。对于所提出的模块嵌入方案，我们根据 SiamFC 和 SiamRPN 之间的差异采用不同的嵌入方法。由于 SiamFC 中存在特征金字塔，判别式增强模块不涉及多层特征的处理，因此只能嵌入时空融合编码器。在 SiamRPN 中，两个模块均可嵌入。在 SiamRPN 中，通过大量实验，我们决定将更新阈值设为 1.18，这将最大限度地降低模板退化的可能性。在 SiamFC 中，与修改后的 SiamRPN 类似，我们直接针对响应图的最大值设置阈值。为方便起见，我们使用 DASTSiam 来指代经我们方法修改后的 SiamRPN。

2. **训练细节**。 训练数据由 LaSOT[Fan 等人、 2019], VID[Russakovsky等人和 2015] 的训练集组成。DASTSiam 共训练 50 个训练轮次。前 10 个训练轮次中，骨干网络的参数保持固定。在其余训练过程中，学习率以对数空间从 0.005 逐渐降至 0.0005。采用动量为 0.9 的随机梯度下降（SGD）作为优化器，小批量大小为 12 对。模板和搜索图像的尺寸分别为 287*287 像素和 127*127 像素。

3. **平台**。 我们的跟踪器使用 Python 3.6 和 PyTorch 1.1.0 实现。实验在配备 NVIDIA 12G 3080Ti GPU 的服务器上进行。

4.**基准测试**。我们在多个公开跟踪基准数据集上评估结果: VOT2018 [Kristan 等， 2018],

LaSOT 测试集[Fan 等， 2019], OTB100 [Wu 等, 2013], GOT-10k [Huang 等， 2019]。

3.2 比较

为全面衡量我们的方法，采用上述四个基准数据集来评估该方法的性能，并将其与其他先进的跟踪算法进行比较。

首先，我们采用所提出的方法以及几种先进的跟踪算法，在 OTB100 上进行测试和评估。DASTSiam 基于 SiamRPN，融合了时空信息和判别式增强模块。实验结果表明，将这两个模块嵌入基线后，跟踪性能均得到提升。SiamFC 的成功率从 58.32 提升至 59.30。SiamRPN 的成功率从 64.10 提升至 67.09。相比之下，DASTSiam 实现了完整的模块嵌入，增强了跟踪器的时间信息整合能力和判别能力。同时，采用了基于目标中心距离的标签分配方法。最后，通过多分支加权计算置信度分数，从而提高了自适应模板更新的置信度。因此，DASTSiam 在基线基础上得到了显著提升。如图 4 所示，DASTSiam 在成功率和精确率指标上均有所提高，其性能超过了 CFNet、GradNet、SiamDW 和 DaSiamRPN 等先进算法。

表 1：VOT2018（红色表示最佳值，绿色表示第二佳值。）

跟踪器	UpdateNet	UPDT	SiamRPN	ECO	DASTSiam
准确性	0.518	0.536	0.576	0.484	0.585
鲁棒性	0.454	0.184	0.323	0.276	0.295
EAO	0.244	0.378	0.324	0.280	0.366

表 2：GOT-10k 上的比较（红色表示最佳值，绿色表示次佳值。）

跟踪器	ECO	SiamRPN++	ATOM	SiamCAR	DASTSiam
AO	0.316	0.517	0.556	0.569	0.567
SR_{50}	0.309	0.615	0.634	0.670	0.656
SR_{75}	0.111	0.329	0.402	0.415	0.452

表 3：用于验证更新策略的消融研究

标签分配	OTB100		GOT-10k		
	成功率	精确率	AO	SR_{50}	SR_{75}
锚框交并比	59.86	0.78	0.416	0.422	0.153
锚框中心距离	**60.70**	**0.80**	**0.423**	**0.426**	**0.158**

为验证 DASTSiam 应对典型跟踪问题的能力，我们在 LaSOT 数据集划分出的测试集上进行了全面的性能测试。如图5所示，与基线相比，DASTSiam 在应对各种跟踪问题方面取得了显著提升。具体而言，为应对形变、长宽比变化、尺度变化等问题，我们嵌入了时空融合模块来整合时间信息，增强匹配模板的特征，从而提升

表 4：时空融合模块（ST）的验证消融研究

跟踪器	ST	OTB100	
		成功率	精确率
SiamFC		58.32	0.77
	✓	**59.30**	**0.78**
修改后的 SiamRPN		60.70	0.80
	✓	**63.34**	**0.83**

表5：用于验证判别性增强的消融研究

跟踪器	ST	DA	OTB100		GOT-10k		
			成功率	精确率	AO	SR_{50}	SR_{75}
修改后的 SiamRPN			60.70	0.80	0.423	0.426	0.158
		✓	**62.40**	**0.81**	**0.458**	**0.543**	**0.251**
	✓		63.34	0.83	0.463	0.548	0.254
	✓	✓	**67.09**	**0.89**	**0.567**	**0.656**	**0.452**

DASTSiam 应对外观变化的鲁棒性。我们引入判别式增强模块来增强搜索区域特征，使 DASTSiam 比基线更能适应背景分离问题。成功率曲线和精确率曲线给出的整体性能结果如图 6 所示，结果表明，我们的 DASTSiam 超越了 SiamRPN++、SiamMask、SiamDW 及其他最先进的跟踪器。

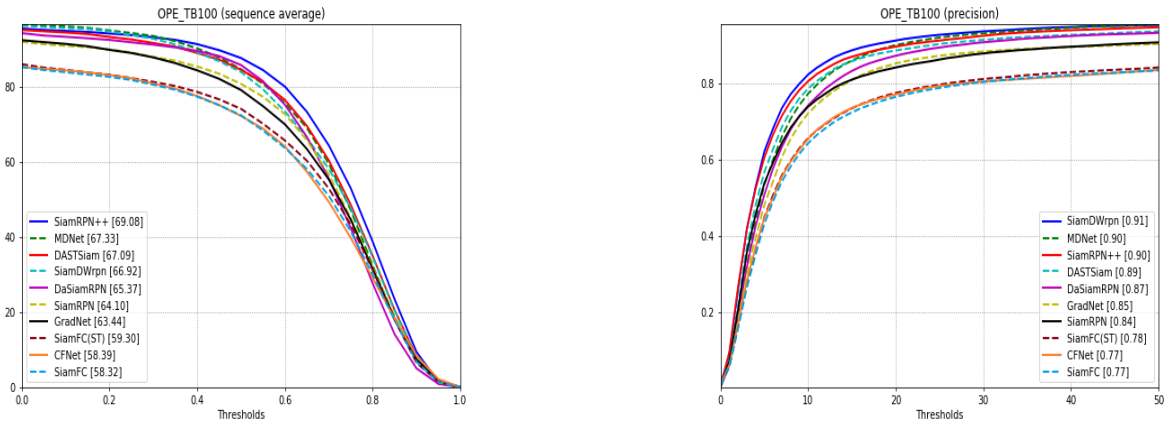
此外，使用 VOT2018 数据集通过不同协议进行评估。DASTSiam 的跟踪性能通过 EAO 进行衡量，EAO 由两个指标计算得出：准确性和鲁棒性。如图 7 所示，在 VOT2018 中，与基线相比，我们的方法在光照变化、相机运动、运动变化、尺寸变化和遮挡方面具有更好的跟踪性能，具体信息请参见表 1。同时，我们还在 GOT-10k 上对 DASTSiam 进行了评估和验证。采用 GOT-10k 提出的协议，即类别平衡指标 AO 和 SR，来衡量我们方法的跟踪性能。如表 2 所示，我们的方法优于 SiamRPN++ 等当前先进算法。这证明了我们方法的有效性。

3.3 消融实验

通过采用相同的训练数据集、训练技术和平台配置，减少无关干扰，从而公平地评估和验证所提出模块的有效性。用于比较的基准模型为 SiamFC 和 SiamRPN。

首先，使用原始 SiamRPN 和修改后的 SiamRPN 验证模板更新方案的有效性。两者均采用相同的骨干网络 ResNet50[He等人， 2016]。通过比较 OTB100 和 GOT-10k，如表 3 所示，我们发现后者提升了跟踪器的性能，这表明它能够带来更可靠的模板更新。

其次，为了验证时空融合模块，我们采用修改后的 SiamRPN 和原始 SiamFC 作为基线，并保持训练策略一致。如表 4 所示，通过在 OTB100 上进行实验发现，将多帧融合模块分别嵌入基线跟踪器后，其性能均有所提升。多帧融合模块的有效性得到了证



(a) OTB100 成功率曲线图　　(b) OTB100 精度曲线图

图 4：成功率和精　　度曲线图展示了一次通过评估（OPE）中我们的方法与 OTB100 数据集上最先进跟踪器的比较 et.

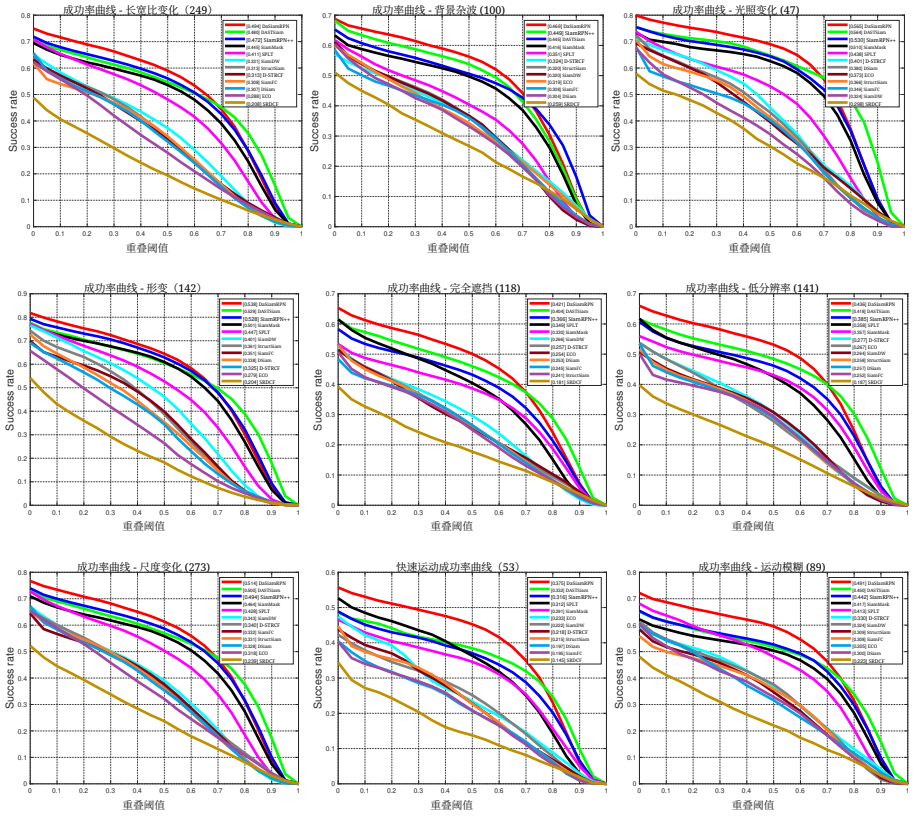
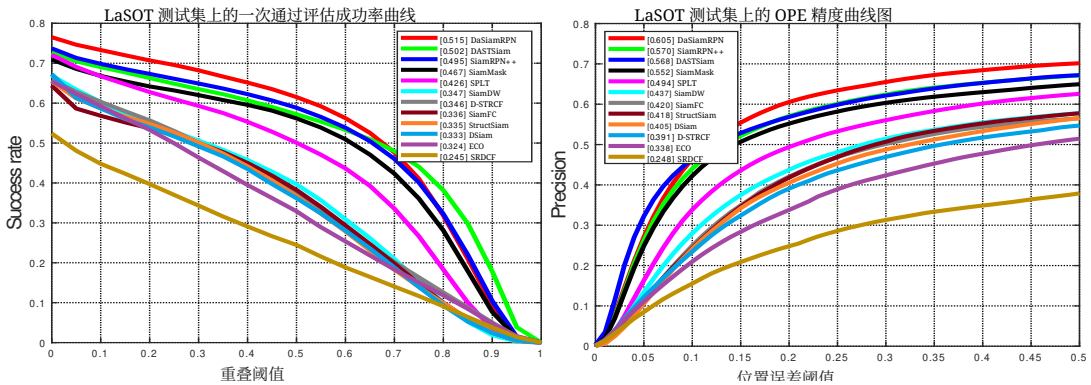


图5：基于属性的评估。实验结果展示了 9 种不同的挑战因素（包括宽高比变化、背景杂波、光照变化、形变、完全遮挡、低分辨率、尺度变化、快速运动和运动模糊）。每种属性挑战因素对应的视频序列数量标注在括号中。

多帧融合模块的有效性得到了证明。

最后，如表5所示，通过修改后的 SiamRPN 验证了判别式增强模块。首先，我们移除时空融合模块，然后添加判别式判别式增强模块。在 OTB100 和 GOT-10k 上的测试表明，跟踪性能得到了提升。



(a) LaSOT 成功率曲线　　(b) LaSOT 精确率曲线

图 6：LaSOT 数据集上的一次通过评估精度曲线和成功率曲线。不同颜色的实线表示 11 种不同跟踪方法的结果。图例分别列出了精度曲线的中心位置误差和成功率曲线的平均重叠得分。(a) LaSOT 成功率曲线。(b) LaSOT 精度曲线图。

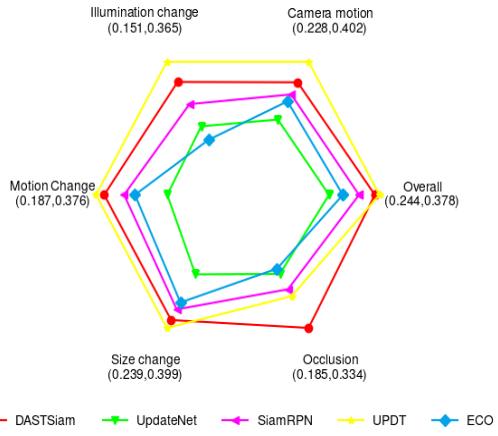


图 7：VOT2018 上以下视觉属性的期望平均重叠率比较：相机运动、光照变化、遮挡、尺寸变化和运动变化。括号中的值表示各属性以及所有跟踪器整体的期望平均重叠率范围。

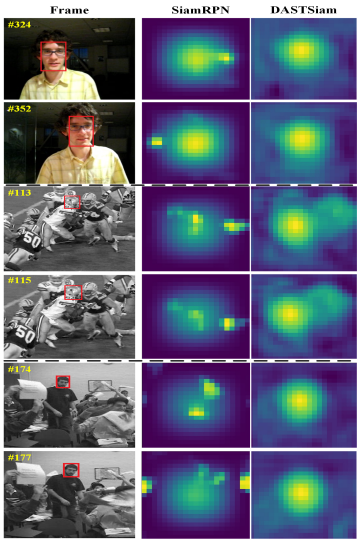


图 8：SiamRPN 和 DASTSiam 的分类响应图可视化。「帧」列展示用于提取模板的搜索区域和真实框。「SiamRPN」展示在 SiamRPN 中将初始模板应用于搜索区域后得到的响应图。对于「DASTSiam」，使用时空融合编码器模块和判别式增强解码器模块增强模板和搜索区域的特征，从而获得更准确的分类响应图。

4 结论

本文提出了一种新颖的目标跟踪方法，该方法利用时空融合模块（ST）和判别式增强模块（DA），以应对跟踪过程中目标形变、遮挡和尺度变化等挑战。在四个基准数据集（OTB100、LaSOT、GOT-10k、VOT-2018）上的实验结果表明，所提出的方法能够有效提升跟踪器的性能。

总之，我们提出的方法通过有效解决现有基于孪生网络的跟踪器在利用时空信息以及从杂乱背景中区分目标方面的局限性，为目标跟踪领域作出了重要贡献。这项工作为特征适配等方向的进一步研究，以及将所提出的方法应用于其他相关任务，开辟了新的机会。

我们相信，我们的方法有望显著提升真实场景下目标跟踪的性能。

致谢

Yucheng Huang 和 Eksan Firkat 做出了同等贡献。感谢 Eksan Firkat、Jihong Zhu 和 Askar Hamdulla 的指导。

参考文献

Luca Bertinetto、Jack Valmadre、Joao F Henriques、An-drea Vedaldi 和 Philip HS Torr。用于目标跟踪的全卷积孪生网络。发表于欧洲计算机视觉会议，第 850–865 页。施普林格，2016年。Nicolas Carion、Francisco Massa、Gabriel Synnaeve、Nico-las Usunier、Alexander Kirillov 和 Sergey Zagoruyko。使用 Transformer 的端到端目标检测。发表于欧洲计算机视觉会议，第 213–229 页。施普林格，2020年。Martin Danelljan、Gustav Hager、Fahad Shahbaz Khan 和 Michael Felsberg。学习用于视觉跟踪的空间正则化相关滤波器。发表于*IEEE*国际计算机视觉会议论文集，第 4310–4318 页，2015年。Martin Danelljan、Goutam Bhat、Fahad Shahbaz Khan 和 Michael Felsberg。ECO：用于跟踪的高效卷积算子。发表于*IEEE*计算机视觉与模式识别会议论文集，第 6638–6646 页，2017年。Heng Fan、Liting Lin、Fan Yang、Peng Chu、Ge Deng、Sijia Yu、Hexin Bai、Yong Xu、Chunyuan Liao 和 Haibin Ling。LaSOT：用于大规模单目标跟踪的高质量基准测试。发表于电气电子工程学会/计算机视觉基金会计算机视觉与模式识别会议论文集，第 5374–5383 页，2019年。Ross Girshick。Fast R-CNN。发表于*IEEE*国际计算机视觉会议论文集，第 1440–1448 页，2015年。Qing Guo、Wei Feng、Ce Zhou、Rui Huang、Liang Wan 和 Song Wang。学习用于视觉目标跟踪的动态孪生网络。发表于*IEEE*国际计算机视觉会议论文集，第 1763–1771 页，2017年。Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。用于图像识别的深度残差学习。发表于*IEEE*计算机视觉与模式识别会议论文集，第 770–778 页，2016年。Jo˜ao F Henriques、Rui Caseiro、Pedro Martins 和 Jorge Batista。使用核化相关滤波器进行高速跟踪。*IEEE*模式分析与机器智能汇刊，37(3):583–596，2014年。Lianghua Huang、Xin Zhao 和 Kaiqi Huang。GOT-10k：用于野外通用目标跟踪的大规模、高多样性基准测试。*IEEE*模式分析与机器智能汇刊，43(5):1562–1577，2019年。Matej Kristan、Ales Leonardis、Jiri Matas、Michael Felsberg、Roman Pflugfelder、Luka ˇCehovin Zajc、Tomas Vojir、Goutam Bhat、Alan Lukezic、Abdelrahman Eldesokey 等人。第六届视觉目标跟踪 VOT2018 挑战赛结果。发表于欧洲计算机视觉会议（*ECCV*）研讨会论文集，第 0–0 页，2018年。Bo Li、Junjie Yan、Wei Wu、Zheng Zhu 和 Xiaolin Hu。使用孪生区域实现高性能视觉跟踪候选框网络。在电气电子工程师学会计算机视觉与模式识别会议论文集，第8971–8980页，2018年。

Feng Li、Cheng Tian、Wangmeng Zuo、Lei Zhang 和 Ming- Hsuan Yang。学习用于视觉跟踪的时空正则化相关滤波器。载于电气电子工程师学会计算机视觉与模式识别会议论文集，第4904–4913页，2018年。Bo Li、Wei Wu、Qiang Wang、Fangyi Zhang、Junliang Xing 和 JunjieYan。Siamrpn++：深度网络推动的孪生网络视觉跟踪演进。载于*IEEE/CVF*计算机视觉与模式识别会议，第4282–4291页，2019年。Niki Parmar、Ashish Vaswani、Jakob Uszkoreit、Lukasz Kaiser、Noam Shazeer、Alexander Ku 和 Dustin Tran。《图像变换器》。载于国际机器学习会议，第4055–4064页。机器学习研究论文集，2018年。Olga Russakovsky、Jia Deng、Hao Su、Jonathan Krause、Sanjeev Satheesh、Sean Ma、Zhiheng Huang、Andrej Karpathy、Aditya Khosla、Michael Bernstein等。《ImageNet大规模视觉识别挑战》。《国际计算机视觉杂志》，115(3)：211–252，2015年。Zhi Tian、Chunhua Shen、Hao Chen 和 Tong He。FCOS：全卷积单阶段目标检测。载于*IEEE/CVF*国际计算机视觉会议论文集，第9627–9636页，2019年。Jack Valmadre、Luca Bertinetto、Joao Henriques、Andrea Vedaldi 和 Philip HS Torr。基于相关滤波器的跟踪端到端表征学习。载于*IEEE*计算机视觉与模式识别会议论文集，第2805–2813页，2017年。Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。Attention Is All You Need。《神经信息处理系统进展》，第30卷，2017年。Qiang Wang、Zhu Teng、Junliang Xing、Jin Gao、Weiming Hu 和 Stephen Maybank。学习注意力：用于实现高性能在线视觉跟踪的残差注意力孪生网络。载于*IEEE*计算机视觉与模式识别会议论文集，第4854–4863页，2018年。Qiang Wang、Li Zhang、Luca Bertinetto、Weiming Hu 和 Philip HS Torr。快速在线目标跟踪与分割：一种统一方法。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第1328–1338页，2019年。Yi Wu、Jongwoo Lim 和 Ming-Hsuan Yang。在线目标跟踪：一项基准测试。载于*IEEE*计算机视觉与模式识别会议论文集，第2411–2418页，2013年。Yanhong Zeng、Jianlong Fu 和 Hongyang Chao。学习联合时空变换以实现视频修复。载于欧洲计算机视觉会议，第528–543页。施普林格，2020年。

张志鹏和彭厚文。用于实时视觉跟踪的更深更宽孪生网络。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第4591–4600页，2019年。张云华、王利军、齐金庆、王栋、冯梦阳和卢湖川。用于实时视觉跟踪的结构化孪生网络。载于欧洲计算机视觉会议（欧洲计算机视觉会议）论文集，第351–366页，2018年。张立超、Abel Gonzalez-Garcia、Joost van de Weijer、Martin Danelljan 和 Fahad Shahbaz Khan。学习孪生跟踪器的模型更新。载于*IEEE/CVF*国际计算机视觉会议论文集，第4010–4019页，2019年。张世峰、迟成、姚永强、雷震和李占山。通过自适应训练样本选择弥合基于锚框与无锚框检测之间的差距。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第9759–9768页，2020年。Zheng Zhu、Qiang Wang、Bo Li、Wei Wu、Junjie Yan 和 Weiming Hu。用于视觉目标跟踪的干扰物感知孪生网络。载于欧洲计算机视觉会议（欧洲计算机视觉会议）论文集，第101–117页，2018年。