基于自注意力改进的逐像素匹配方法

摘要

目前基于 Siamese Network 的跟踪器的一般架构是采用以 ResNet 为代表的现代深度网络获取目标模板与搜索区域的特征，将两张特征图进行相关操作，再通过后处理获取候选框。通常的相关操作是以模板特征作为卷积核与搜索特征做卷积操作，今年来也兴起了一种逐像素匹配的方法，但都无法解决模板中的背景干扰问题。为解决这个问题本文提出在相关操作之前对特征图进行自注意力的特征提取，抑制背景信息的干扰，以提升相关操作的效果，使得模版特征中的目标特征能在结果中得到更高的响应值。

关键词：孪生网络；逐像素匹配；自注意力机制

1 引言

视觉目标跟踪的目的是估计任意一个给定的目标在视频序列的每一帧中的位置。在大多数情况下，仅给跟踪器提供第一帧目标的位置作为初始信息，然后跟踪器需要对下一帧中的目标外观进行建模并计算位置信息。由于目标的不确定性及其特定信息仅在测试时可用，因此无法通过离线网络的训练来获得目标模型。近年来，许多研究者都在探索如何利用深度学习技术的力量来解决跟踪任务。孪生网络是目前最流行的深度学习框架之一。作为先行者，SiamFC [1] 将视觉跟踪问题转化为一个深度目标匹配问题。具体来说，SiamFC 包括两个分支，即目标模板和搜索区域。第一个分支将目标建模为一个固定的样例，后一个分支处理目标可能存在的区域。SiamFC 启发了许多后来的跟踪器 [2–5]，这些跟踪器建立在 Siamese 网络架构之上，可以实现最先进的性能。其中，SiamRPN 引入了区域建议网络 (RPN)，该网络由用于前景背景区分的分类 head 模块和用于锚点细化的回归 head 模块组成。SiamRPN++ [2] 和 SiamFC++ [5] 解决了以往孪生网络采用深度网络性能反而下降的问题，释放了现代深度网络的潜力，如 ResNet [6] 和 GoogleNet [7]，使用这些网络来增强特征表示。受 FCOS [8] 和 CornerNet [9] 等无锚目标检测器的启发，许多无锚跟踪器 [10,11] 采用像素预测方式来执行目标定位。虽然基于 Siamese 的跟踪器已经取得了很好的性能，但仍然存在一个局限性：Siamese 跟踪器容易被背景信息所干扰。在选取模板区域时，由于特征提取网络的需要，选取的区域是以目标位置为中心，截取包括整个目标在内的正方形区域。因此在对模板进行特征提取的特征图中，除了目标的特征之外，也包含了大量背景信息的特征，这导致了无论采用什么方式对模板和搜索区域进行相关操作，都会被模板中的背景信息所干扰，尤其是采用逐像素匹配的方法，这种相关操作更加注重细节的特征信息，缺乏对目标整体结构的搜索，这使得这种方法更容易在与背景特征相关的区域得到更高的响应，从而导致跟踪性能的下降。为了解决这个问题，我们提出在将两个特征图逐像素匹配之前，将模板和搜索特征图输入自注意力

模块（Self-Attention），训练网络让目标在输出结果中获得更高的响应，从而使得将模板和搜索特征图进行逐像素匹配时，使目标区域获得更高的响应值，而抑制背景噪声的干扰。本文主要的工作如下：

- 我们改进了逐像素匹配这种相关操作的流程，使得通过自注意力机制来抑制背景噪声的干扰，防止跟踪器被背景干扰物所影响。
- 我们在 UAV123，OTB100，VOT2016 等数据集上均获得了比改进前的模型更好的跟踪性能。

2 相关工作

2.1 孪生网络视觉追踪

最近，SiamFC [1] 将视觉跟踪任务转换为目标模板与搜索区域之间的一般相似度匹配问题，通过大规模离线训练学习一个通用判别器。后续提出的跟踪器通过引入注意力机制 [4]、设计新的网络架构 [12]、使用增强损失 [13] 或利用强化学习 [14] 来进一步增强 Siamese 框架。在这些后续研究中，值得一提的是 SiamRPN [3] 引入了 RPN 模块来预测长宽比变化的目标的边界框，而不是 SiamFC 的暴力破解离散尺度搜索策略。因此，SiamRPN 将 SiamFC 升级为高级框架。基于 SiamRPN，提出了多种跟踪器。其中，DaSiamRPN [15] 收集了更多样化的训练数据，以增强辨别能力。C-RPN [16] 构建了多阶段 RPN 来更准确地进行状态估计。SiamRPN++ [2] 采用更深层次的 ResNet-50 [6] 网络来增强特征表示。受无锚点目标检测和实例分割的启发，一些跟踪器将原始 RPN 架构修改为逐像素跟踪 [5, 10]。虽然上述 Siamese 跟踪器取得了令人满意的性能，但它们容易受到干扰，即跟踪器的鲁棒性较弱。为了解决这一鲁棒性问题，许多研究人员引入了在线深度学习技术来提高跟踪器的泛化能力。例如，ATOM [9] 和 DiMP [17] 在每次在线跟踪时构建目标专用分类器，并收集历史难负样本来增强分类器。UpdateNet [18] 更新目标模板以合并时态信息。MAML [19] 通过元学习将目标检测器 (如 FCOS [8]) 传递给跟踪器，并增量更新网络权重以适应目标特殊序列。然而，这些跟踪器需要仔细设计在线跟踪协议，以避免错误和冗余的更新。本文提出的是对现有的逐像素匹配方式进行优化，仍然是一种离线训练的模式并不修改网络架构，不会增加模型在推理阶段的计算量。

2.2 模板特征提取中的背景干扰问题

大多数高级跟踪器在选取目标模板时都会截取以目标为中心的正方形区域，而在目标的长宽比不等于 1:1 时这种获取模板的方式会附带大量的背景信息，导致在模板特征图中只有一小部分是目标的特征，其中有相当多的背景特征。在早期的一些跟踪器中 [2, 3]，对于目标模板与搜索区域的匹配采取的是卷积求相关的方式，这样将目标模板视作一个整体，作为卷积核对搜索区域进行匹配的时候是附带着大量背景信息对搜索区域进行相关操作，这会导致在目标背景发生改变或者目标本身形变时难以取得良好的跟踪效果。后来一些跟踪器采取了逐像素匹配的方式进行相关操作，然而这种操作过于注重细节信息，对于背景信息和目标特征一视同仁的进行匹配，依然没有解决大量背景特征对最终响应图的影响。我们提出方法不

同的地方在于，我们采用自注意力机制来对学习区分目标模板中的前景背景信息，对其赋予不同的权重再进行匹配，以此来抑制背景信息的干扰。

3 本文方法

3.1 对 Siamese 跟踪器的优化

标准的 Siamese 跟踪器将模板图像 z 和搜索图像 x 作为输入。图像 z 在第一帧中指出了所关注的目标，在随后的视频帧中要求跟踪器在搜索区域 x 中定位目标。模型结构如图 1所示：

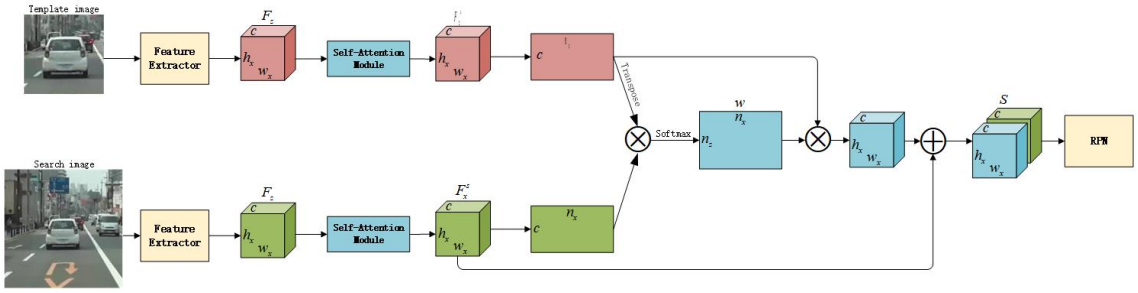

图 1. 基于 Siamese 网络的跟踪器管道包括分类和定位两个子任务。本文提出在逐像素匹配之前先将特征图输入自注意力模块进行处理，有利于在相关操作中提高目标区域的响应值。

这两幅图像被输入到一个共享权重的骨干网络中，分别生成特征图 $F_z \in {}^{H_z \times W_z \times C}$ 和 $F_x \in {}^{H_z \times W_z \times C}$ ，然后利用匹配网络对 F_z 和 F_x 进行处理，得到相似特征图 F：

$$F = \varphi(F_z, F_x) \tag{1}$$

许多流行的 Siamese 跟踪器将其定义为深度交叉相关 (DW-Corr) [20]。近年来，受视频目标分割 [21] 的启发，许多研究者采用非局部关注 [22] 的变体——逐像素相关方法 (PW-Corr) [23, 24] 作为跟踪任务的匹配网络。我所改进的模型采用的是 PW-Corr [23] 的简化版本：

$$w_{ij} = \frac{\exp[(F_z^i \odot F_x^i)/\sqrt{C}]}{\sum_{\forall k} \exp[(F_z^k \odot F_x^i)/\sqrt{C}]} \tag{2}$$

其中，F_z 和 F_x 分别被重塑为 $H_z W_z \times C$ 和 $C \times H_x W_x$ 的矩阵，i 和 j 分别是 F_z 和 F_x 上每个像素的指数。符号 $\odot$ 表示点积运算。然后得到相似矩阵 $w \in {}^{H_z W_z \times H_x W_x}$，相似特征映射 S 的运算为：

$$S = concat(F_x, (F_z)^T \otimes w) \tag{3}$$

其中 $concat()$ 表示矩阵串联，符号 $\otimes$ 表示矩阵乘法。然后将相似特征映射 S 输入到由分类模块 θ_{cls} 和定位模块 θ_{loc} 组成的 RPN head 模块中。本模型采用 Anchor-free 的 RPN head 模块。这样就可以得到分类图 A_{cls} 和回归图 A_{loc}：

$$A_{cls} = \theta_{cls}(S), A_{loc} = \theta_{loc}(S) \tag{4}$$

A_{cls} 是从搜索区域中识别出可能是前景的置信度，A_{loc} 是用于对目标的边界框进行回归。Siamese 跟踪器使用的标准损失函数定义为：

$$loss_{rpn} = \frac{1}{N_{pos}} \sum_{i \in A_{pos}} (loss_{cls}(A^i_{cls}, Y^i_{cls}) + loss_{loc}(A^i_{loc}, Y^i_{loc})) + \frac{1}{N_{neg}} \sum_{i \in A_{neg}} loss_{cls}(A^i_{cls}, Y^i_{cls}) \quad (5)$$

其中 N_{pos} 和 N_{reg} 分别为正样本集 A_{pos} 和负样本集 A_{neg} 的个数，Y_{cls} 和 Y_{loc} 分别表示分类和回归的标签。$loss_{cls}$ 通常使用的是交叉熵损失函数，$loss_{loc}$ 通常使用的是 smooth L1 或 IoU 损失函数。在公式 (2) 中可以看到一个潜在的问题，即模板特征图中实际上有相当多的区域表示的是背景特征。而这个逐像素匹配的相关操作对前景背景是不加区分的，这就导致了网络对前景背景的关注是相同的，这将导致网络会学习到错误的背景信息从而降低对目标的跟踪性能。

3.2 自注意力模块

我们的改进方法基于 Tang 等人的工作 [25]，如上所述，大多数基于 Siamese 的跟踪器通过 ResNet 等深度网络提取特征图后，就直接将这些特征图进行相关操作，忽略了模板特征图中携带的大量背景噪声，这会严重影响跟踪器的鲁棒性。有的跟踪器会在 RPN 网络中进行更加细化的操作来提出背景噪声的干扰 [16]，但如果能在最初输入的特征图中直接抑制背景噪声，则不需要过多的后处理也能提高跟踪器的性能。

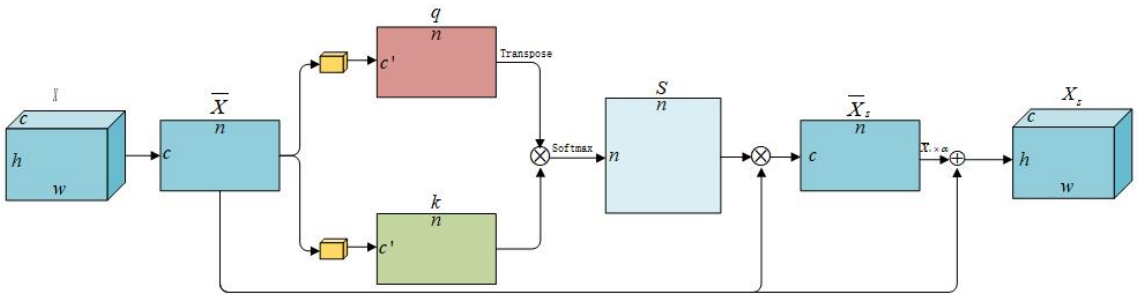

图 2. 自注意力模块，其中符号 $\otimes$ 表示矩阵相乘操作，符号 $\oplus$ 表示矩阵相加操作，黄色立方体表示 1×1 的卷积，$n = h + w$，c' 是将输入特征图降低的通道数，α 为调整自注意力权重的超参。

为了解决这个问题，我们提出将自注意力机制与逐像素匹配方法相结合。如图 1中所示，原模型中并没有添加 Self-Attention Module，而是直接将模版和搜索的特征图进行逐像素相关操作，这样会导致在训练和推理阶段都受到大量背景噪声的干扰。我们的修改在 ResNet 后添加了自注意力模块，自注意力模块的结构如图 2所示。首先将输入的特征图经过卷积再 reshape 为二维矩阵得到 $q, k \in {}^{c \times n}$，再进行如下操作得到权重图：

$$S = Softmax(q^T \otimes k) \quad (6)$$

将输入特征图与权重图 S 做矩阵乘法，并以一定的权重与原始输入 $\overline{X}$ 做残差得到：

$$X_s = \alpha \times (\overline{X} \otimes S) + \overline{X} \quad (7)$$

自注意力模块会随着网络的训练，对特征图中的前景和背景区域赋予不同的权重，而不是像原方法一样一视同仁的对所有区域都进行相关操作。这样会在相关之前就在一定程度上抑制背景噪声，从而使目标区域的响应值更高，而背景区域则受到抑制。

4 复现细节

4.1 与已有开源代码对比

此部分为必填内容。如果没有参考任何相关源代码，请在此明确申明。如果复现过程中引用参考了任何其他人发布的代码，请列出所有引用代码并详细描述使用情况。同时应在此部分突出你自己的工作，包括创新增量、显著改进或者新功能等，应该有足够差异和优势来证明你的工作量与技术贡献。

本文使用了 Tang 等人提出的基于排名的改进方法的代码 [25]，基于此代码进行了改进。在保留了该网络大体骨架的同时，在经过 ResNet 提取的特征图后加入自注意力模块，自注意力机制来增强目标区域在特征图中的响应权重，以此来达到抑制背景干扰的效果。同时自注意力机制配合逐像素匹配的方法，使得网络对目标细节处的特征识别能力更强，在应对形变或遮挡的情况时拥有更强的稳定性。

4.2 实验环境搭建

我们的跟踪器使用 Pytorch 跟踪平台 PySOT 实现，并在一张 3090 显卡上进行训练。

4.3 界面分析与使用说明

如图 3跟踪器会选择视频第一帧的目标区域作为模板图像，在后续帧中对目标进行持续跟踪

4.4 创新点

我们在原模型的基础上，为解决跟踪器在长时跟踪下易被背景干扰和变形遮挡等问题，在逐像素匹配方法的基础上引入了自注意力机制，试图通过自注意力模块加强网络对目标细节特征的感知能力，使得跟踪器在应对目标变形和遮挡时有更好的鲁棒性。

5 实验结果分析

5.1 原模型的效果

在展示我们的效果之前，首先展示一下原模型与其他模型的对比效果。

VOT2016 效果如表 1：原模型是在 SiamBAN 的基础上做优化得到的，相比起 SiamBAN，SiamPW-RBO 在鲁棒性和 EAO 上均有显著的提升。

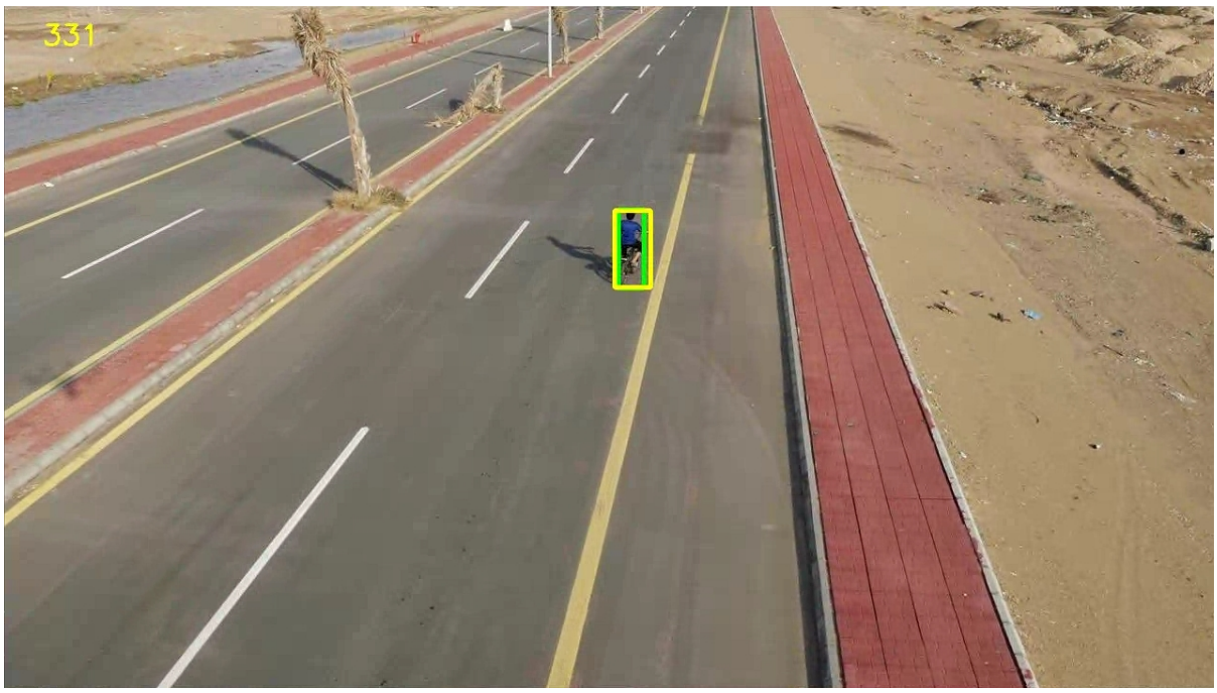

图 3. 跟踪器效果示意

Tracker	A(↑)	R(↓)	EAO(↑)
SiamRPN	0.578	0.312	0.337
SiamRPN++	0.642	0.196	0.463
Ocean	0.625	0.158	0.486
SiamBAN	0.632	0.149	0.502
SiamDW	0.580	0.240	0.371
SiamAttn	0.680	0.140	0.537
SiamPW-RBO	0.617	0.098	0.531

表 1. 原模型 SiamPW-RBO 在 VOT2016 数据集上与各大跟踪器的对比。

OTB100，TC128，UAV123，NFS30 效果如表 2：SiamPW-RBO 相较于 SiamBAN 在性能上有显著提升。

Tracker	OTB100	TC123	UAV123	NFS30
SiamRPN++	69.6	57.3	61.3	50.2
SiamBAN	69.6	58.4	61.4	59.0
SiamCAR	65.7	57.8	61.4	53.3
Ocean	67.2	55.1	59.2	51.8
SiamRPN++-ACM	71.2	-	63.4	-
SiamBAN-ACM	72.0	-	64.6	-
SiamPW-RBO	69.8	59.3	64.5	60.1

表 2. 原模型 SiamPW-RBO 在 OTB100，TC123，UAV123，NFS30 四个数据集上与各大跟踪器的对比。

5.2 我们的改进与原模型效果对比

接下来是我们的改进与原模型的性能对比。

OTB100：如表 3，我们在 OTB100 数据集上验证了我们提出的跟踪器，该数据集由 100 个完全注释的序列组成。如表 1 所示，我们进行改进的跟踪其在该数据集上的效果略弱于原模型，可能是由于自注意力机制需要更长的时间来进行收敛导致的。但是除 OTB100 之外的其他数据集，我们的模型所展现出的性能相较于原模型有显著的提升。

数据集及其指标		SiamPW-RBO	SiamSA-RBO(ours)
OTB100	Success(↑)	65.6	65.1
	Precision(↑)	87.3	87.1

表 3. 我们的模型与原模型在 OTB100 上的性能对比。

VOT2016：如表 4，我们的改进效果再 VOT2016 上效果显著，在保证了鲁棒性的前提下，对目标跟踪的 Accuracy 从 60.0% 提升至 61.6%，提升 1.6%，EAO 从 0.442 提升至 0.452。

数据集及其指标		SiamPW-RBO	SiamSA-RBO(ours)
VOT2016	Accuracy(↑)	0.600	0.616
	Robustness(↓)	0.144	0.144
	EAO(↑)	0.442	0.452

表 4. 我们的模型与原模型在 VOT2016 上的性能对比。

UAV123：如表 5，在 UAV123 上我们的改进使 Success 从 60.3% 提升至 60.7%，Precision 由 82.5 提升至 82.6，性能在整体上有一定的提升。

数据集及其指标		SiamPW-RBO	SiamSA-RBO(ours)
OTB100	Success(↑)	60.3	60.7
	Precision(↑)	82.5	82.6

表 5. 我们的模型与原模型在 UAV123 上的性能对比。

可以看出在加入了自注意力模块之后，我们的改进方法使模型的跟踪性能在多数情况下取得了一定幅度的提升，在进一步优化之后能够得到比原模型更好的跟踪性能。

6 总结与展望

在本文中，我们提出了在逐像素匹配之前对特征图进行自注意力特征提取的方法，在自注意力模块中抑制背景噪声，减少背景信息对跟踪器性能的影响。在自注意力模块加入之后，跟踪器可以在推理过程中预先排除掉一些背景信息的干扰，使网络能够更加专注于目标所在区域，在该区域获得更高的响应值，从而获得更好的跟踪效果，实验结果表明在一般情况下在加入我们的改进方法之后，能使模型的跟踪性能获得 $0.4\% \sim 1.6\%$ 不等的提升，考虑到模型的鲁棒性并没有降低，该方法带来的提升比较显著。

参考文献

[1] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II 14*, pages 850–865. Springer, 2016.

[2] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4282–4291, 2019.

[3] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8971–8980, 2018.

[4] Qiang Wang, Zhu Teng, Junliang Xing, Jin Gao, Weiming Hu, and Stephen Maybank. Learning attentions: residual attentional siamese network for high performance online visual tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4854–4863, 2018.

[5] Yinda Xu, Zeyu Wang, Zuoxin Li, Ye Yuan, and Gang Yu. Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12549–12556, 2020.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[8] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019.

[9] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018.

[10] Zedu Chen, Bineng Zhong, Guorong Li, Shengping Zhang, and Rongrong Ji. Siamese box adaptive network for visual tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6668–6677, 2020.

[11] Fei Du, Peng Liu, Wei Zhao, and Xianglong Tang. Correlation-guided attention for corner detection based visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6836–6845, 2020.

[12] Guangting Wang, Chong Luo, Zhiwei Xiong, and Wenjun Zeng. Spm-tracker: Series-parallel matching for real-time visual object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3643–3652, 2019.

[13] Xingping Dong and Jianbing Shen. Triplet loss in siamese network for object tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 459–474, 2018.

[14] Ning Wang, Wengang Zhou, Guojun Qi, and Houqiang Li. Post: Policy-based switch tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12184–12191, 2020.

[15] Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, and Weiming Hu. Distractor-aware siamese networks for visual object tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 101–117, 2018.

[16] Heng Fan and Haibin Ling. Siamese cascaded region proposal networks for real-time visual tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7952–7961, 2019.

[17] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6182–6191, 2019.

[18] Lichao Zhang, Abel Gonzalez-Garcia, Joost Van De Weijer, Martin Danelljan, and Fahad Shahbaz Khan. Learning the model update for siamese trackers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4010–4019, 2019.

[19] Guangting Wang, Chong Luo, Xiaoyan Sun, Zhiwei Xiong, and Wenjun Zeng. Tracking by instance detection: A meta-learning approach. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6288–6297, 2020.

[20] Kean Chen, Jianguo Li, Weiyao Lin, John See, Ji Wang, Lingyu Duan, Zhibo Chen, Changwei He, and Junni Zou. Towards accurate one-stage object detection with ap-loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5119–5127, 2019.

[21] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9226–9235, 2019.

[22] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[23] Zhihong Fu, Qingjie Liu, Zehua Fu, and Yunhong Wang. Stmtrack: Template-free visual tracking with space-time memory networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13774–13783, 2021.

[24] Dongyan Guo, Yanyan Shao, Ying Cui, Zhenhua Wang, Liyan Zhang, and Chunhua Shen. Graph attention tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9543–9552, 2021.

[25] Feng Tang and Qiang Ling. Ranking-based siamese visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8741–8750, 2022.