基于 DiMP 算法的全局实例跟踪

作者：陈若愚 202118018629015
导师：黄凯奇，陈晓棠，赵鑫

2022 年 6 月 24 日

Abstract

目标跟踪是人类视觉系统的基本能力，计算机视觉任务模拟了目标跟踪。人类可以在任意场景中持续定位任意目标，在复杂甚至对抗环境下依旧保持鲁棒跟踪能力。而目前的单目标跟踪算法大多在简单连续场景上能取得不错的结果，与人类差距仍然比较大，与人类差距仍然比较大。为了更好的探索计算机视觉的类人跟踪能力，本文从全局实例跟踪 [1] 的角度出发，实现类人的目标跟踪算法。本文基于 DiMP 算法 [2] 实现，并在 VideoCube[1] 公开数据集进行算法的训练与测试。

Key Words 目标跟踪，全局实例跟踪，DiMP 算法，VideoCube 数据集

1 全局实例跟踪

全局实例跟踪，英文称为 **global instance tracking (GIT)**，是由 Hu 等人 [1] 提出的一个概念。该任务是在没有任何关于摄像头或运动一致性假设的情况下，搜索视频中任意用户指定的实例，以模拟人类的视觉跟踪能力。如图1所示，主流的目标跟踪算法分为 4 种，即视频实例检测 (VID)，多目标跟踪 (MOT)，单目标跟踪 (SOT)，全局实例跟踪 (GIT)。通常 VID 算法核心是检测，他只能限制于数据集的类别，通常无法针对任意实例，只能针对部分目标。MOT 通常需要检测与跟踪，通常无法面对场景的变换，针对的目标也非常有限。SOT 算法针对任意的目标，对于运动具有一定的假设，但是对于场景通常是固定不变的。与 SOT 不同，GIT 除了针对任意物体外，也可以针对任意场景。从算法上而言全局实例跟踪算法更符合类人跟踪能力。

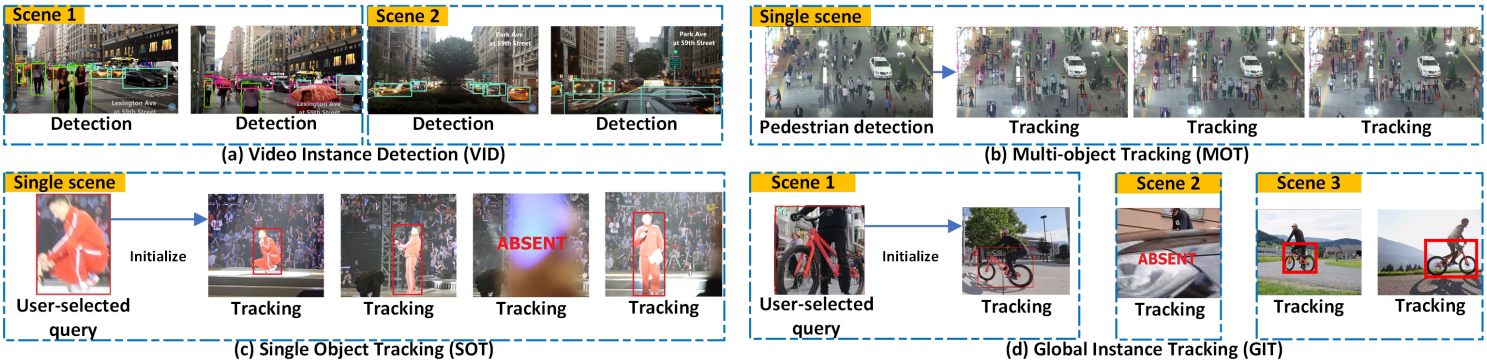

Figure 1: 视觉跟踪的四个类别，视频实例检测 (VID)，多目标跟踪 (MOT)，单目标跟踪 (SOT)，全局实例跟踪 (GIT)。

2 单目标跟踪算法

由于 GIT 的概念在最近提出，很少有关于 GIT 相关的目标跟踪算法。考虑到全局实例跟踪算法的特性，在选择复现的模型时我们更倾向于基于搜索与匹配的算法。传统上，任意对象跟踪的问题是通过专门在线学习对象外观模型来解决的，使用视频本身作为唯一的训练数据。尽管这些方法取得了成功，但他们仅在线的方法本质上限制了他们可以学习的模型的丰富性 [3]。Bertinetto

[1] VideoCube 数据集: http://videocube.aitestunion.com/

等人 [3] 提出了 SiamFC，旨在通过匹配方式在图像中匹配最相似的模型，SiamFC 的结构如图3所示。然而，仅仅靠孪生网络难以面对场景变换，目标尺度不变性的问题。后期提出的一些算法，如 SiamRPN[4]，SiamRPN++[5]，SiamDW[6]，SiamFC++[7] 等等。

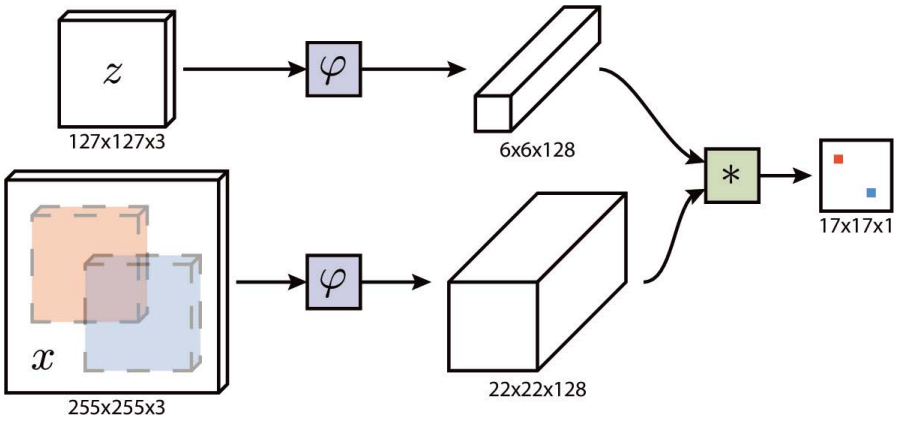

Figure 2: 全卷积孪生网络 SiamFC 结构。[3]

考虑到全局实例跟踪的特性是在没有任何关于摄像头或运动一致性假设的情况下，搜索视频中任意用户指定的实例，以模拟人类的视觉跟踪能力。我们理解为全局搜索并匹配，因此我们认为应当基于搜索或者遍历的方法，并且判断需要通过 Zero Shot 的方式进行判断，因此选择基线我们需要选择基于孪生网络的架构。这里我们预选的框架为 DiMP 网络 [2]。

3 DiMP 目标跟踪算法

DiMP 算法[2]是 Bhat[2] 在 2019 年提出的方法。这篇算法之前的端到端可训练计算机视觉系统的努力对视觉跟踪任务提出了重大挑战。与大多数其他视觉问题相比，跟踪需要在推理阶段在线学习稳健的目标特定外观模型。为了实现端到端的可训练，目标模型的在线学习因此需要嵌入到跟踪架构本身中。由于强加的挑战，流行的连体范式只是简单地预测目标特征模板，而在推理过程中忽略背景外观信息。因此，预测模型具有有限的目标-背景可辨别性。DiMP 算法一个端到端的跟踪架构，能够充分利用目标和背景外观信息进行目标模型预测。

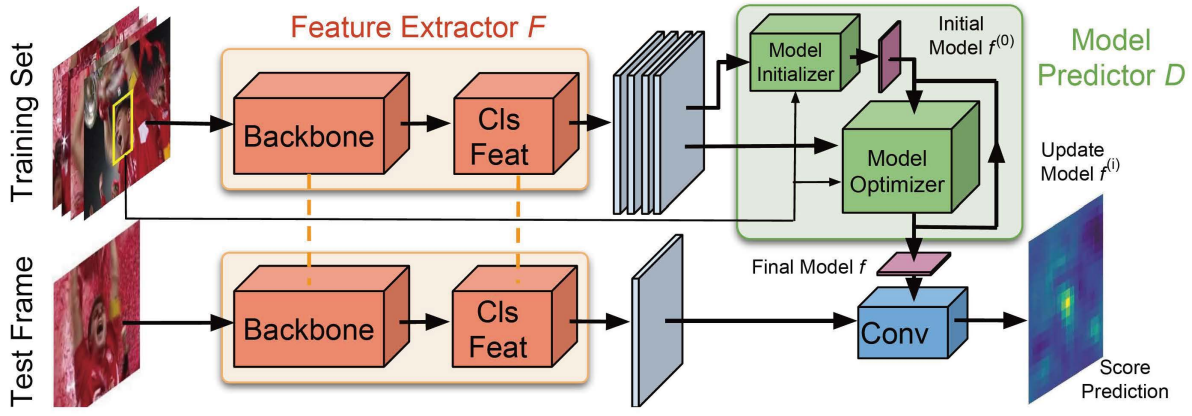

Figure 3: DiMP 结构。[2]

本文主要包括两个原则：

- 判别式学习损失提高了学习目标模型的鲁棒性。
- 确保快速收敛的强大优化策略。

[2]DiMP: https://arxiv.org/abs/1904.07220

下面将详细介绍该模型细节以及核心算法。

3.1 Discriminative Learning Loss

模型的 Model Predictor 包含一个训练集 $S_{\text{train}} = \{(x_j, c_j)\}_{j=1}^{n}$。由特征提取网络 F 提取的深度特征图 $x_j \in \mathcal{X}$。每一个样本将与相应的目标中心坐标配对 $c_j \in \mathbb{R}^2$。给定此数据，本文的目的是为了预测一个目标 $f = D(S_{\text{train}})$。
首先作者定义了残差方程：

$$r(s, c) = v_c \cdot (m_c s + (1 - m_c)\max(0, s) - y_c) \tag{1}$$

则损失函数：

$$L(f) = \frac{1}{|S_{\text{train}}|} \sum_{(x,c)\in S_{\text{train}}} \|r(x * f, c)\|^2 + \|\lambda f\|^2 \tag{2}$$

其中 s 为目标置信度得分 $s = x * f$，目标区域由 m_c 定义，其中 $m_c(t) \in [0, 1]$，$t \in \mathbb{R}^2$ 为空域。物体区域 $m_c \approx 1$，背景区域 $m_c \approx 0$。其中目标掩膜 m_c，空间权重 v_c，正则化因子 λ 甚至回归目标 y_c，都由模型学习得到。

3.2 算法流程

整个算法的流程如图4所示。这便是 DiMP 算法的全部核心，接下来我们的算法也将依赖于该模型复现。

Algorithm 1 Target model predictor D.

Input: Samples $S_{\text{train}} = \{(x_j, c_j)\}_{j=1}^{n}$, iterations N_{iter}

1: $f^{(0)} \leftarrow \texttt{ModelInit}(S_{\text{train}})$ # Initialize filter (sec 3.3)
2: **for** $i = 0, \ldots, N_{\text{iter}} - 1$ **do** # Optimizer module loop
3: $\nabla L(f^{(i)}) \leftarrow \texttt{FiltGrad}(f^{(i)}, S_{\text{train}})$ # Using (1)-(2)
4: $h \leftarrow J^{(i)} \nabla L(f^{(i)})$ # Apply Jacobian of (2)
5: $\alpha \leftarrow \|\nabla L(f^{(i)})\|^2 / \|h\|^2$ # Compute step length (5)
6: $f^{(i+1)} \leftarrow f^{(i)} - \alpha \nabla L(f^{(i)})$ # Update filter
7: **end for**

Figure 4: DiMP 算法流程。[2]

4 数据集描述与分析

VideoCube 数据集是 Hu 等人 [1] 发布并针对全局实例跟踪的数据集。由于其数据集非常大，且不利于训练，因此我们在训练集中下载了部分数据，即前 20 条图像数据作为训练集，测试集也从中选择了 5 条作为测试。所使用的的数据集信息如表1所示。

Table 1: 本实验中所使用的数据集信息。

	使用数据编号
训练集	002, 003, 004, 010, 011, 013, 014, 016, 017, 018 019, 020, 021, 023, 024, 025, 026, 028, 030, 031
测试集	077, 079, 083, 087, 089

5 实验分析

5.1 实现细节

我们基于 pytracking[3]框架进行代码复现。我们使用 DiMP-50 作为本次实验的 backbone。由于机器性能不足，我们的模型在 DiMP-50 预训练模型[4]上进行微调。其中我们的输入批次为 4 段，每段 16 帧。学习率为 0.01。其余参数与原始项目推荐参数保持一致，然后在表格1列出的训练集上进行训练。

5.2 数据划分

尽管在表1中我们指定了训练所需的数据，但是每一段视频的帧数并不相同。我们对每一段视频进行随机的采样 16 帧，每个视频采样 5 次，即总共有 50 个数据作为输入。测试阶段对每个视频全部进行测试，不再单独划分帧数。

5.3 训练结果

模型的训练过程如图5，这个是 pytracking 中提供的可视化结果。

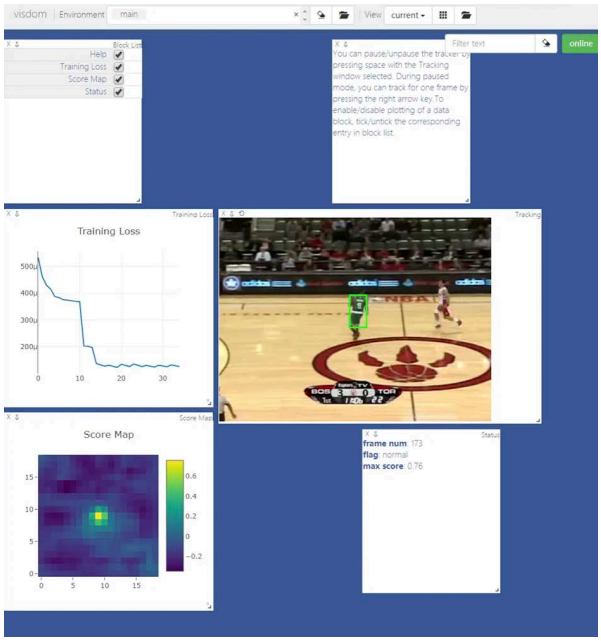

Figure 5: 模型训练结果。

由于 VideoCube 数据集不方便测试，因此部分的测试结果如图6所示，第一帧为人工框选，具有一定跟踪功能。

[3]pytracking: https://github.com/visionml/pytracking

[4]DiMP-50 预训练模型: https://drive.google.com/file/d/1qgachgqks2UGjKx-GdO1qylBDdB1f9KN/view

Figure 6: 模型测试结果。

6 结论

本文从全局实例跟踪的任务角度出发，重新回顾了全局实例跟踪任务。考虑到类人类跟踪的技术，我们认为基于孪生网络的搜索与匹配方法更适合我们的模型，我们选择了 DiMP 算法做为我们的实现算法。我们基于 pytracking 框架进行了模型的训练与可视化跟踪结果，具有一定性能。尽管模型的精度仍然非常欠缺，但是已经取得较好的效果。未来我们将针对全局实例跟踪模型设计更具有鲁棒性的算法。

References

[1] Shiyu Hu, Xin Zhao, Lianghua Huang, and Kaiqi Huang. Global instance tracking: Locating target more like humans. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (01):1–1, 2022.

[2] 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.

[3] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pages 850–865. Springer, 2016.

[4] 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.

[5] 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.

[6] Zhipeng Zhang and Houwen Peng. Deeper and wider siamese networks for real-time visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4591–4600, 2019.

[7] 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.