中国科学院大学
University of Chinese Academy of Sciences

硕士学位论文

基于无人机监控系统的关键弱小目标感知

作者姓名：蒋楠

指导教师：焦建彬　教授

中国科学院大学微电子学院

学位类别：工程硕士

学科专业：计算机技术

研究所：中国科学院大学微电子学院

2021 年 6 月

Key Tiny Object Perception based on UAV Monitoring System

A Thesis Submitted to

University of Chinese Academy of Sciences

in partial fulfillment of the requirement

for the degree of

Master of Engineering

in Computer Technology

By

Jiang Nan

Supervisor: Professor Jiao Jianbin

School of Microelectronics

University of Chinese Academy of Sciences

June, 2021

中国科学院大学

研究生学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明或致谢。

作者签名：蒋楠

日　　期：2021.5.27

中国科学院大学

学位论文授权使用声明

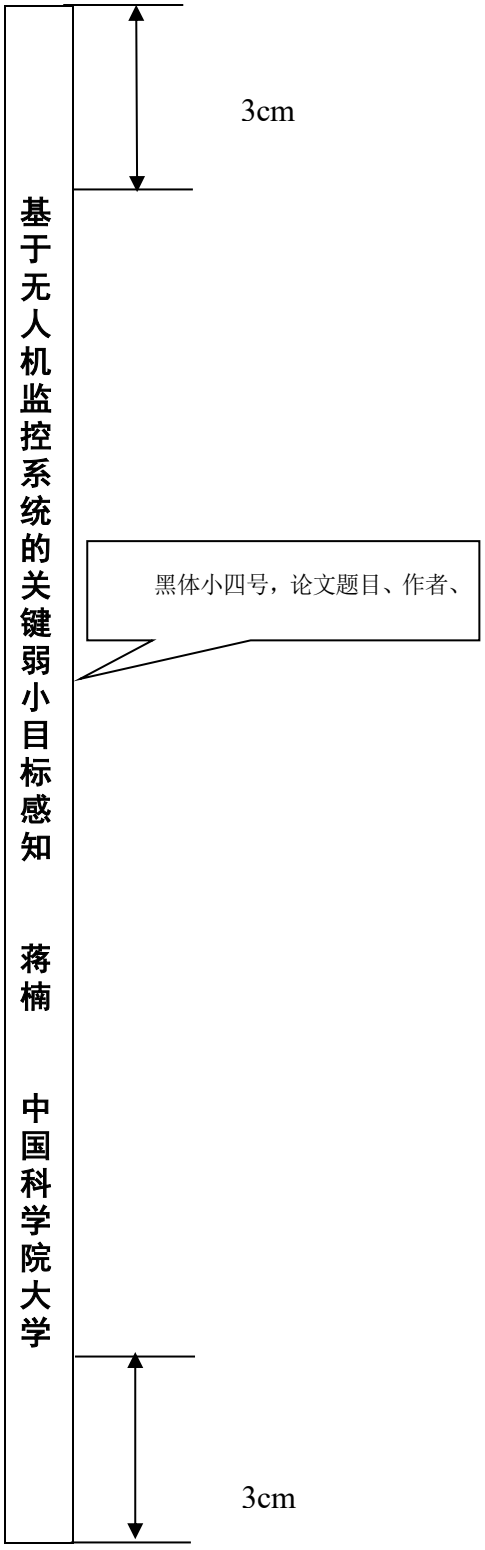
本人完全了解并同意遵守中国科学院有关保存和使用学位论文的规定，即中国科学院有权保留送交学位论文的副本，允许该论文被查阅，可以公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存、汇编本学位论文。

涉密的学位论文在解密后适用本声明。

作者签名：蒋楠　　导师签名：

日　　期：2021.5.27　　日　　期：2021.5.27

书脊（此页仅用于制作书脊，不用单独打印放入论文）



摘 要

弱小目标感知（Tiny Object Perception，TOP）作为计算机视觉领域中一个新兴的重要任务，涉及机器学习、图像处理、多模态传感器、深度学习等多个领域的相关技术。科技的迅速发展使得弱小目标感知在自动驾驶、安防监控、医学图像分析等多个领域均有重要的应用背景。近年来的蓬勃发展，无人机（Unmanned Aerial Vehicle，UAV）技术取得了突破性的进展，商业和民用无人机的普及率和流行度显著上升。无人机的应用正在不断深入，同时带动了弱小目标感知技术日益更新。但是作为一柄双刃剑，无人机技术尽管可以便利人们日常生活，对其的滥用却会造成经济损失、对公众安全构成威胁。在此基础上，本文针对无人机监控和管制两个方面进行了深入探索，形成了一套基于无人机监控系统的关键弱小目标感知技术，研究内容和成果包括：

1）针对空对地的人员安全监控需求，提出了基于无人机航拍的弱小人体目标检测任务。面向海上快速救援的应用背景，作为参与者，发布了弱小人体目标检测数据集 TinyPerson。迎合数据驱动的特性，提出了基于精细尺度匹配的弱小人体目标检测预训练策略。该策略除了引入额外的目标检测数据集外，还从图像级别和实例级别两种尺度匹配的角度出发，将额外预训练数据集的尺度分布向下游目标任务数据集的尺度分布迁移，进一步减少了尺度分布间的差异性，具有直接的指导意义。

2）针对地对空的潜在威胁管制需求，提出了基于多模态互感的无人机目标跟踪任务。发布了高质量多模态无人机目标跟踪数据集 Anti-UAV。充分利用不同视频序列间信息交互的优势，提出了基于双流语义一致性的无人机跟踪训练方法，该方法的引入并没有增加推理阶段的计算量。将跟踪器的训练过程解耦为两个阶段：类别级别语义调制和实例级别语义调制，通过两个级别的模板特征调制分别提升跟踪器的鲁棒性和判别能力。

关键词：弱小目标感知，无人机，深度学习，目标检测，目标跟踪

Abstract

As an emerging and vital branch in the computer vision community, tiny object perception (TOP) is related to many fields such as machine learning, image processing, multimodal sensor, deep learning, and many other aspects. As technology rapidly increases, TOP has been dramatically advanced and widely applied in automatic driving, security surveillance, medical image analysis, and other fields. With the vigorous development of modern science, there is a breakthrough in unmanned aerial vehicle (UAV) technology. With this, the accessibility and popularity of UAVs for commercial and recreational use have significantly surged. UAV applications are deepening at the same time give TOP technology based on deep learning a boost. However, UAV technology is double-edged, for although it can facilitate people's daily lives, UAV technology abuse spawns economic losses and threatens public security. On this basis, this paper explores in great depth of UAV surveillance and regulation, forms a set of key TOP technology based on the UAV monitoring system. The contributions of this dissertation are summarized as follows.

1) Focusing on the air-to-ground requirements of personnel safety monitoring, a tiny person object detection task based on UAV aerial images is proposed. For the application of quick maritime rescue, TinyPerson, a tiny person object detection dataset, is jointly released. To meet the needs of data-driven characteristics, a pre-training strategy based on fine scale matching is proposed. This strategy transfers the scale distribution of the additional pre-training dataset to that of downstream task dataset in two levels, *i.e.*, image level and instance level. This further reduces the difference between the scale distributions, enhancing the guiding significance.

2) Aiming at the ground-to-air requirements of potential UAV threat regulation, a UAV object tracking task based on optical information is proposed. A high-quality multimodal UAV tracking dataset, termed Anti-UAV, is published. Taking full advantage of information interaction between different video sequences, a UAV tracking training method based on dual-stream semantic consistency is proposed with nearly no additional

computation in both training and testing stages. The training process of the tracker is decoupled into two stages, *i.e.*, class-level semantic modulation and instance-level semantic modulation, which improve the robustness and discrimination ability of the tracker, respectively.

Keywords: Tiny Object Perception, UAV, Deep Learning, Object Detection, Object Tracking

目 录

图目录

表目录

第1章　引言

1.1　研究背景和动机

视觉作为人类获取信息的主要来源，承担了绝大部分和外界环境进行交互的任务。计算机视觉作为当前计算机智能感知的重要研究领域，旨在促使计算机具备和人类一样的识别和分析视觉信息的能力。得益于软件和硬件领域欣欣向荣的发展，因而将计算机视觉应用在社会公共安全等各个方面已经成为目前流行的热点研究问题。在软件算法层面，人工智能技术的发展极大地推动了监控安防领域的自动化处理能力，提高了检测的精度。同时作为硬件的载体之一，无人机技术的发展为该领域提供了更高自由度的新型监控手段，也带来一定程度上的隐患。在此基础上，两者的有效结合更是催生了新兴领域的发展。

1.1.1　人工智能发展现状

自 20 世纪中旬在达特茅斯会议上科学家们讨论如何用机器模仿人类后，人工智能（Artificial Intelligence）的概念首次诞生：研究使计算机这类机器能够实现类人智能，诞生出像人类一样的思维能力。人工智能未来可行的应用场景就一直萦绕在人类的脑海之中。通过大量的时间和漫长的探索，科研工作者从未放弃心中这一个耀眼璀璨的梦，期间对于人工智能的研究难免遭遇过挫折，但是一直处于方兴未艾的阶段，尤其是近十年来，人工智能在科技领域的发展欣欣向荣，在计算机视觉（Computer Vision）、推荐系统（Recommender Systems）等细分领域中更是有着天翻地覆的变化。

作为人工智能各个领域当前主流的技术实现方法，深度学习（Deep Learning）通过使用多层非线性变换累积而组成了多个处理单元，或者更加复杂的非线性结构来搭建深层的神经网络（Neural Network）对数据进行抽象，即组合浅层特征重组为愈加抽象的深层表示特征，从而学习数据在样本空间中的表示层次和潜在规律。区分于较为传统的浅层学习，深度学习更加验证了表征学习（Representation Learning）的重要性，通过利用大规模数据来学习特定任务需求的特征，从而建立起输入

到输出的函数关系式，尽可能地拟合现实的关系形式。谈及深度学习的发展历程，Hinton 等人[1]在 20 世纪初提出深度学习的概念以及改进相应网络模型的训练方法，一定程度上解决了之前反向传播算法（Backpropagation algorithm，BP）[2]的问题；后来，Alex 等人[3]提出的 AlexNet 在 ImageNet LSVRC2012 的分类任务中获得了卓越的效果。深度学习的成功归结于以下几点：

1）强大的硬件计算能力。GPU 硬件技术的发展，大大提升了计算机的算力，进而可以从海量的数据中学习各种数据特征。

2）充足的训练数据。Li 等人[4]意识到数据的重要性，进而提出规模庞大的 ImageNet 数据集，同时也符合神经网络依赖大量数据驱动的特性。

从此，开启了深度学习的时代。各种应用场景中争相出现：Google 提出的 AlphaGo[5]和 AlphaGo Zero[6]，不断突破围棋界的上限，相继打败人类顶尖高手；Tesla 和 Waymo 等科技公司采集大量数据作为驱动自动驾驶技术发展的燃料，结合深度学习赋能于自动驾驶研发；抖音和快手等短视频社交软件利用深度学习处理内容分析和用户标签关系，从而提供更好的个性化推荐等等。

值得一提的是，人工智能虽然迄今为止蓬勃发展，但是距离达到强人工智能仍有较为遥远的路途要走。强人工智能即为计算机可以独立思考问题，设计解决难题的最优化策略。然而，当下人工智能相关的研究工作都处于在弱人工智能阶段，仅能够让计算机具备一定的感知和推断的能力来决策一些已经在数据场景中出现的问题，暂时尚不具备很强的泛化能力来完美处理训练数据中出现过的问题，以及暂且不能自适应地推测训练数据中不曾出现的状态。计算机视觉领域同样需要更强的人工智能技术，例如，现实世界中对于弱小目标的感知具有较大的研究价值。弱小目标具备绝对尺度小、相对尺度小、信息量弱等特点，相比于通用目标检测会对目标的定位产生巨大的挑战，尤其当其处在广阔且复杂的背景中会更加难以被捕捉。对于如上所述的研究领域迫切需要越来越多的科研工作者不断前赴后继地去完善当前弱人工智能的方案、去探索强人工智能的可行性。

1.1.2 无人机技术发展历程

无人机（Unmanned Aerial Vehicle，UAV）特指可重复使用、无机载人类飞行员的航空器。在此基础上，无人机可以通过人类操作员远程遥控，或者具备自主导航

能力。早在一战、二战时期，无人机在关键战役中大放光彩，具备伤亡率极低甚至零伤亡率、高机动性、低可探测性等等特性，在军用领域受到了广泛认可。因此，无人机在收集场景情报，大范围、多类型监测等各种任务提供了强效有力的支持。

近年以来，无人机产业发展迅速、急速扩张，渐渐从军用领域拓展、延伸到了民用和商用领域，并且民用无人机市场开始占据主流。伴随着新型轻质复合材料和通信技术的发展，再加上飞行控制系统（Flight Control System）逐渐成熟，民用无人机的发展可谓是锦上添花。在此基础上，消费级无人机成本得到很好的把控，配备摄像头的无人机在旅游航拍等情形被广泛使用。除此之外，Amazon、京东和美团等互联网企业也在积极研发无人机配送技术，旨在降低不必要的人力成本，一定程度上提高偏远地区物流的配送效率。

然而，无人机技术给人们的日常生产生活带来了诸多便捷，却也给不怀好意的不法分子提供了犯罪的新工具，他们利用无人机携带危险物品、武器对政府重要人物进行袭击。例如，2019 年 9 月 14 日，歹徒使用无人机袭击沙特国家石油公司设施，给沙特造成惨重的经济损失。此外，该事件还影响到了世界各国。此前若具备反无人机精确的侦察手段，在受到恐怖攻击前可对无人机进行全面的反制，尽可能地减少破坏。不仅如此，在民用领域，无人机也不幸沦为走私、贩毒的常用工具，机场“黑飞”事件以及隐私泄露案件频繁爆出，对社会的公共安全产生了严重的威胁。因此，除国家相关部门积极完善无人机管控的条例外，愈来愈多科研工作者也在进行复杂场景下无人机感知的研究，从而对无人机进行有效的监管，有效遏制无人机带来的内在威胁，提升反无人机的探测系统的能力。

1.2　本文的研究内容

人工智能在基于机器视觉的感知领域中有着举足轻重的地位，其中关键弱小目标感知是一项极其重要的难点问题，给众多科研工作者带来了巨大的挑战。关键弱小目标感知在许多任务中扮演着重要并且不可忽略的角色，有着广泛的应用场景。比如，在炙手可热的自动驾驶领域中，车载摄像头拍摄到的高分辨率场景数据中行人目标或者交通标志太小，以及激光雷达采集到的远距离点云目标的特征稀疏，这些弱小目标的精确感知是实现高级别安全自动驾驶的重要前提。尤其当汽车处于高

速移动的过程中，对于视觉中弱小目标（也即远处的目标）的感知则更为重要。对于医学图像处理领域，精密仪器拍摄的医学图像中各种微小病变区域的成功检测有助于早期诊断病症、防止病情恶化到无以复加的地步。除此之外，军方需要在卫星遥感图像中对于舰船、飞机等武装力量进行有效地检测。许多领域都对弱小目标感知有特定的适用场景，因此，关键弱小目标的感知技术具有很高的研究价值。

依托于飞行器技术的急速发展，基于无人机、直升机等飞行器的关键弱小目标感知渐渐走入大众视野，通过飞行器高空、远距离俯拍地面或者海面进行关键目标的探测和监控的技术是当前应用研究的热点之一。在不便设置监控摄像头的场景，诸如海滩等，可以通过无人机技术进行远距离监控、掠海检测，发现遇难人员即可进行海上快速救援。作为一个岛国，日本已经成立了一支专业精良的海上航空搜救队伍——航空自卫队航空救难团。可以说，日本具备东亚最富有经验的海上航空救援力量。对此，面向海面快速救援的空对地感知需求的重任就落在了关键弱小目标感知上，但是目前公开的目标检测数据集内人体目标的尺度分布尚不足以满足弱小目标感知的研究需求，需要尺度更小的弱小人体目标检测数据集。

除此之外，同时也要谨慎预防无人机技术的滥用，无人机在远端空中快速移动和它的低可探测性也对反无人机探测技术提出了更高的要求。综合利用各种物理传感器（如光电摄像机、红外成像仪等）完成地对空感知需求，来寻找威胁目标，通过物理特性的测量来感知关键弱小目标——无人机，达到实时监控无人机空间位置和运动轨迹等信息，也是亟待解决的关键问题。然而，学术界在无人机目标跟踪领域的研究暂不完善，还没有相关工作提出公开可用的高质量基准数据集、基线方法和评价指标。因此，相关数据集的搭建将极大推动未来反无人机感知前沿技术的发展，成熟的反无人机感知技术无异于给国家和商业安全打上一剂强效镇静剂。

1.3 本文的主要贡献

随着人工智能领域的兴起，依托于深度学习的目标感知相比于传统方法精确度和效率都得到了较大提升[7][8]。同时无人机技术也日新月异，价格也变得更加亲民。无人机技术和目标感知算法的充分结合，顺势而生了许多涉及到社会公众安全的应用场景。

基于无人机的航拍图像进行区域的安全监控，以面向海上快速救援的应用背景为例，通过无人机、直升机等飞行器在海滩周围进行航拍人体目标感知，拯救落水遇难人员、监控临海区域安全；同时利用多模态光学传感器对于存在安全隐患的无人机进行反无人机探测感知，有效地遏制潜在威胁。因此，我们将空对地的弱小人体目标检测和地对空的光学无人机弱小目标跟踪技术整合形成一项基于可控的无人机监控系统的关键弱小目标感知流程，如图 1.1 所示。通过图 1.1 中的天地协同的弱小目标感知流程，一方面完成基于航拍的人员安全监测，另一方面完成反无人机的无人机目标跟踪技术以加强管控。

然而，基于深度学习的算法依赖数据驱动。数据量级越大，场景多样性越丰富，以这类数据当作神经网络的训练集，有利于提升弱小目标感知算法的性能和鲁棒性。当前缺少相关场景的数据限制了人工智能算法在弱小目标感知领域的发展，亟需收集、筛选、标注并整理出相应的数据集。以高质量数据集作为弱小目标感知研究的燃料，能够更好地研究如何高效地进行弱小目标感知，不仅在相应的数据集上取得较好的性能，还要符合实际的场景探测要求，在实际的应用中也大放异彩才是产学研结合的最高的追求。

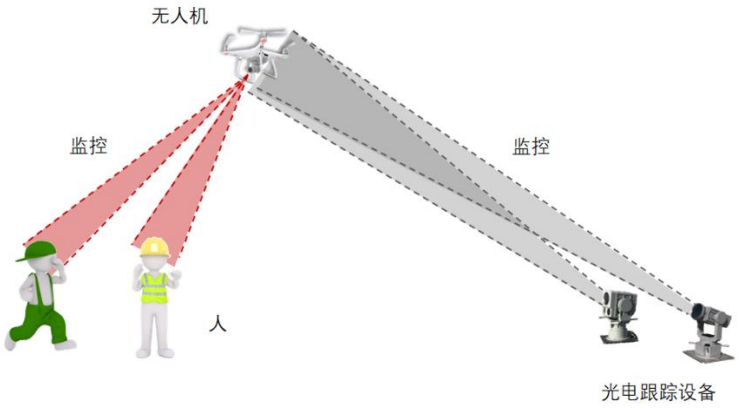


图 1.1　基于无人机监控系统的关键弱小目标感知

Figure 1.1　Key tiny object perception based on UAV monitoring system

针对以上所述的光学弱小目标感知的研究问题，本文的工作是通过相应无人机监控装置完成研究数据的原始收集，将其整理成为高质量的数据集可驱动关键弱小目标感知算法的研究。在此基础上提出公开可用的高质量基准数据集、基线方法和评价指标，同时，基于相应数据集的特点，对研究内容进行深入地剖析，进行关键

弱小目标感知算法的迭代、优化。工作内容可以分为几下几点：

1）制备了关键弱小目标感知相关研究所需的数据集。作为参与者，收集、标注并整理发布了基于无人机航拍的弱小人体目标检测数据集 Tinyperson。同时面向反无人机的应用场景，收集、标注并整理发布了基于多模态互感的无人机跟踪数据集 Anti-UAV。

2）开展了基于无人机航拍的弱小人体目标检测的研究。深入剖析了 Tinyperson 数据集和现有相近领域的公开数据集的异同，体现了弱小人体目标检测任务的独立性和互补性。并且，总结和分析了基于深度学习的目标检测及小目标检测领域内相关检测算法，结合了当前针对数据特性的预训练算法的研究内容，提出了基于精准尺度匹配的弱小人体目标感知的预训练算法，其中共同提出了基于图像级别的尺度匹配算法，并将图像级别推进到实例级别，提出了基于实例级别的尺度匹配算法。

3）开展了基于多模态互感的无人机跟踪的研究。首先分析了 Anti-UAV 数据集和目标跟踪领域其他公开数据集各自的侧重，并且总结了不同模态下基于深度学习的跟踪器的发展现状。结合基于图像信息交互的训练策略和 Anti-UAV 数据集的特性，提出了基于双流语义一致性的无人机训练策略，该方法仅仅作用于跟踪器的训练阶段，在推理阶段没有增加任何计算量。

本文对于上述的每个研究内容，都进行了大量实验以验证和分析，并且与其他优秀的方法做了对比分析。最后在搭建的数据集中，实验结果表明本文提出的关键弱小目标感知算法的性能优于同类型的其他优秀算法。

1.4 本文的组织结构

第一章，引言，起初对人工智能领域及其当前主流方法深度学习的概念和发展历程作了陈述，从个人角度简要分析了深度学习兴起的必要条件。同时介绍了无人机技术的发展，分析了无人机技术带来的优势和随之而来的安全问题。通过当前的弱小目标感知的背景呈现，结合无人机技术的依托，阐明了目前对于特定场景下的关键弱小目标感知技术有很高的研究价值，从而引申出本文的主要研究内容。首先亟需高质量的数据集作为支撑，在其基础上开展基于无人机监控系统的关键弱小目标感知技术的两方面研究——基于无人机航拍的弱小人体目标检测算法和基于多

模态互感的无人机目标跟踪算法。最后介绍了本文的主要贡献和组织结构。

第二章，相关研究，分为基于无人机航拍的弱小人体目标检测和基于多模态互感的无人机跟踪两个部分，第一个部分介绍了当前相关的公开数据集，并从多个角度分析了不同数据集间的异同，简要归纳了基于深度学习的目标检测模型和小目标检测算法的研究现状。第二个部分首先介绍了目标跟踪领域现有的不同模态的公开数据集，通过比较相同点、不同点阐明了反无人机感知的研究价值，并分析了当前在不同模态数据下基于深度学习的跟踪器类型。

第三章，基于无人机航拍的弱小人体目标检测，依托于海上快速救援的应用背景，作为参与者，提出了基于无人机航拍的弱小人体目标检测数据集 TinyPerson，并搭建了相关的基准评测平台。通过分析现有预训练策略并结合弱小目标检测数据集的特点，提出了基于精细尺度匹配的弱小目标检测预训练策略。最后，通过大量实验验证了从图像级别和实例级别进行单向尺度迁移的有效性，以及深入剖析了算法提升性能的缘由。

第四章，基于多模态互感的无人机跟踪，为了填补当前缺少反无人机相关数据集的空缺，提出了面向无人机管控的多模态无人机跟踪数据集 Anti-UAV，并搭建了高质量的相关基准算法评测平台，对 40 多个跟踪器进行评测及分析。从 Anti-UAV 仅有无人机这一通用类别的特性出发，首先分析了当前基于信息交互的训练策略，提出了基于双流语义一致性的训练策略，进而在 Anti-UAV 上进行大量实验验证和分析。最后，从监督任务和权重因子两个方面对提出的训练策略进行了深度剖析。

第五章，总结与展望，归纳了全文的研究工作。其中主要从数据集形式、当前算法改进和其他思路出发，改善基于无人机监控系统的关键弱小目标感知，提出了未来研究的工作方向。

第2章　相关研究

鉴于深度学习依赖于海量数据的驱动，因此，使用深度学习的方法来研究某一个具体领域需要特定条件下数据的支持，而直接使用已经发布的不太相关的公开数据集会严重影响神经网络的性能。目前，学术界在特定领域的关键弱小目标感知领域的研究逐渐开始，要夺取这王冠上的明珠，需要首先建立合理的高质量基准数据集、基线方法和评价指标，其中数据集具体细节分别可以查阅第三章、第四章对应部分。在此基础上，进行弱小目标感知算法的研究，深入分析相近领域算法，特定问题采用特定分析的方式，通过对症下药的方法论来迎接关键弱小目标感知这一巨大的挑战。

2.1　基于无人机航拍的弱小人体目标检测

2.1.1　目标检测数据集

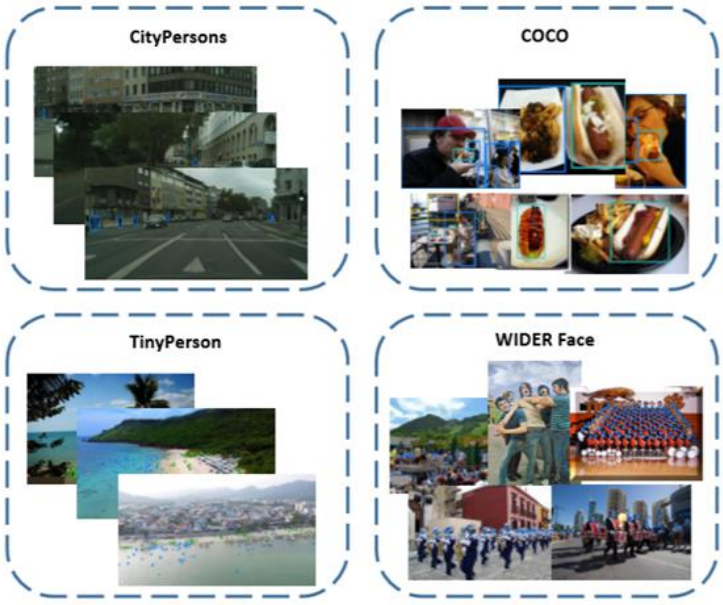


图 2.1　弱小人体目标检测数据集 TinyPerson 及相关数据集的展示

Figure 2.1　The display of tiny person detection dataset TinyPerson and related datasets

作为驱动神经网络的必备燃料，愈来愈多的数据集被公开发布。弱小人体目标检测[9]作为一个处于萌芽阶段的新兴领域，尚未得到很好的研究。弱小人体目标检

测中的弱小人体目标被定义为信息量弱小、绝对尺度和相对尺度都很小的人体。其中，绝对尺度定义为目标所占的像素大小，相对尺度定义为目标所占的像素和整个样本面积的比值的大小。为了更好地阐述这个任务和其他相近任务的区别，我们将从和弱小人体目标检测关联程度较高的任务中选取经典数据集进行分析，来更好地明确这个任务所具有的特点。如图 2.1 所示，我们选取了几个具有代表性的数据集作为展示，横坐标为目标的尺度大小（根号下相应面积大小）。这其中囊括了通用目标检测（General Object Detection）、行人检测（Pedestrian Detection）和人脸检测（Face Detection）等，对于每个任务分别对应选取了 MS COCO[10]、CityPerson[11]、WIDER Face[12]。选取这些数据集的原因有如下几点：

1）MS COCO 作为经典的通用目标检测数据集受到了广泛的研究，其次人这一类在 MS COCO 中占据了大部分，和弱小人体目标检测人这唯一类别相关度高。

2）WIDER Face 是大规模的人脸检测数据集，其中人脸的包围框普遍较小，且和 TinyPerson 数据集的尺度分布相近，如图 2.2 所示。

3）CityPerson 和 TinyPerson 同样为以人为主体的目标检测数据集。

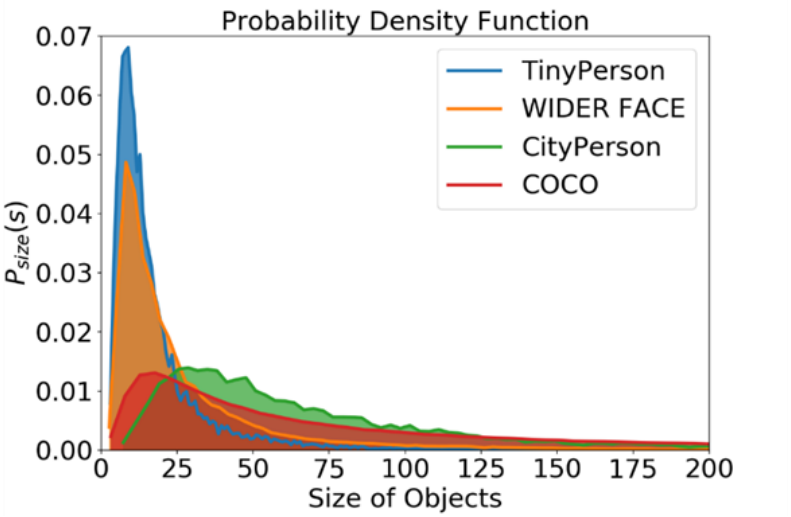


图 2.2　不同目标检测数据集训练集的尺度分布

Figure 2.2　The scale distribution of training sets for different object detection datasets

接下来将从数据集的具体细节上分别深入分析它们之间的区别：

■　通用目标检测

通用目标检测通常为多尺度、多类别的目标检测数据集。MS COCO 将物体分为大、中、小三个尺度，其中小目标被定义为尺度范围在小于 32*32 个像素点的物体，中目标被定义为像素点在 32*32 到 96*96 范围内的物体，而大于 96*96 个像素点的物体则为大目标。然而，专注于弱小人体目标检测数据集 TinyPerson 可以视为单一类别、单一尺度的数据集，其尺度分布如图 2.2 所示，尺度分布较为尖锐且靠近 y 轴，表明整个数据集中的人体目标基本都集中在小尺度这一范围内。

除此之外，MS COCO 中小目标所占的比例和尺度都达不到弱小目标检测的要求。如表 2.1[13]所示，根据 MS COCO 中所有标注的包围框的统计，小目标的分布极度不均匀，即意味着所有包围框中有 41.4 %是小目标，占据多尺度目标的大部分，但只有近一半的样本中包含了小目标。除小目标外的其余目标分布都相对均匀，在总样本数量中占比 70%以上。简而言之，小目标的包围框数量很多但分布非常不均匀，有接近 50%的样本中没有小目标的存在。模型在训练中会更倾向于拟合出现频次高的物体，因此在这种情况下不利于网络模型对于小目标的学习。

表 2.1　MS COCO 训练集中不同尺度目标的分布

Table 2.1　Distribution of objects with different scales in MS COCO training set

统计信息	小目标	中目标	大目标
在总样本框数量中所包含的占比（%）	41.4	34.3	24.3
在总样本图数量中所包含的占比（%）	52.3	70.7	83.0

面向海上救援的弱小人体目标检测中人体目标大多尺度范围都在小于 20*20 个像素点，更关注于相对尺度和绝对尺度均比通用目标检测中小目标更小的目标，并且只关注于这些目标，因此可以认为是是通用目标检测在尺度上的补充，而并不简单是通用目标检测其中作为小目标的一部分。

■　行人检测

行人检测关注的目标主体和弱小人体目标检测相同，都是只关注于人这一个主体，但是其中有所侧重。以 CityPerson 为例，一方面该数据集的收集主要通过车载摄像头进行正面拍摄，以汽车前视的正面视角呈现出来。在这种情况下，行人被拍摄到的视角相对单一。其次，由于是平视的视角，行人之间的遮挡现象也是行人检

测任务中一个巨大的挑战，但是这样的遮挡情况在航拍视角下很少存在。

另一方面，在标注框的长宽比分布上，CityPerson 标注框中行人的包围框多为行走中或站立中的人，因而包围框的长宽较为单一，基本聚集在 2.44 左右。在目标的尺度分布上，CityPerson 要比 MS COCO 更贴近弱小目标的定义范畴，但是仍有一定的差距。

■ 人脸检测

如图 2.2 所示，为了更鲜明地展示尺度分布，实际上截断了分布中在 x 轴上大于 200*200 个像素点的部分，在这其中人脸检测数据集 WIDER Face 在目标绝对尺度分布上是最接近弱小人体目标检测数据集 TinyPerson 的。即使 WIDER Face 中小尺度的人脸占据了很大的比重，但是仍然有部分大尺度的人脸，这是不可忽略的一部分。WIDER Face 中人脸包围框的长宽比相对单一的，绝大部分性能表现优秀的人脸检测器均使用 1 或者 1.5 作为锚点框的预设长宽比，在这方面上类似于行人检测数据集 CityPerson。

需要注意的是，虽然 WIDER Face 在绝对尺度和相对尺度上都是在这些数据集中最接近弱小人体目标检测的，但是往往人脸作为目标是和人体一起出现的，这两部分是不可分割的。一定程度上可以将人体作为背景信息进行探索、利用，两者相辅相成，所以可以认为人脸检测这个任务并不是一个纯粹的弱小目标检测。同时，由于人脸检测任务发起较早，目前已经出现了很多研究人脸检测的工作，出现了很多专为人脸检测数据集所设计的网络结构及一系列算法。相比于较为成熟的人脸检测任务，而弱小目标检测任务还很新颖。

■ 遥感目标检测

以 DOTA 数据集[14]为例，数据集中样本均为卫星遥感视图，视角开阔，样本通常绝对尺度很大。与本课题专注的弱小人体目标检测不同的地方在于遥感目标检测的类别多为舰船、飞机、汽车等，因此目标的长宽比相对来说较为极端，一定程度上会影响锚点框的先验设置。同时标注框的形式为旋转框，具有方向上的不确定性。

综上所述，弱小人体目标检测和以上四种目标检测任务在某些角度存在相似性，但是更多的是有着特定领域专有的特性。弱小人体目标检测相比于通用目标检测更加专注于单一的弱小尺度，相比行人检测在视角上具有更大的多样性，相比于

人脸检测不具备强先验的知识，相比遥感目标检测更专注于检测人体这一特定类别。弱小人体目标检测，作为弱小目标感知的一个子任务，目前对于这个任务的研究还远远不够。

2.1.2　基于深度学习的目标检测研究

得益于深度学习的发展，目标检测任务备受科研工作者的关注。首先需要声明的是，本文所涉及到的检测器均为基于锚点框的检测器，分为两种：一步法检测器（One Stage Detector）和二步法检测器（Two Stage Detector）。一步法检测器的工作原理为首先在输入图像上布置由超参数预先设定的锚点框，然后利用卷积神经网络（Convolutional Neural Networks，CNN）对样本进行特征提取。将锚点框和特征输入到区域候选网络（Region Proposal Network，RPN）中进行分类和回归任务，从而得到检测结果。然而，二步法检测器在一步法检测器的基础上，在得到回归后的候选包围框后，提取候选包围框对应的感兴趣区域（Region of Interesting，ROI）特征，再次进行分类和回归任务以得到最终的检测结果。通常情况下，一步法检测器胜在推理速度快，而二步法检测器具有更高的精度，占据了各个数据集的排行榜前列，但是性能的差距在逐渐地减少。在一步法检测器中，我们选取了 YOLO[15]、SSD[16]、RetinaNet[17]进行分析，而在二步法检测器中，Faster RCNN[18]和 FPN[19]较为经典，经常作为科研工作者的基线方法。

■　一步法检测器

YOLO 系列检测器发展至今已经迭代出了 5 个版本，博得了工业界和学术界的青睐，在这里仅以 YOLO v1 为例，后续简称为 YOLO。YOLO 的主要流程为将原始输入图片分为 x*x 的网格，每一个格子内的特征负责预测 y 个类别的各自可能性，并且每个格子都会对应 z 种不同类型的目标框。每个包围框包括能表达包围框信息的四个值及一个置信分数（Confidence Score），因而预测结果可以编码为 x*x*(5z+y)维的特征向量，然后进行分类与回归任务，从而得到最后的预测结果。通过上述的检测流程可以发现 YOLO 简单的网格化处理原始输入图像来得到预测结果，但是这样的建模方式会导致不同大小的目标如果落入同一个网格的情况会一定程度上忽视小目标的特征，不利于小目标的检测。

SSD 借鉴了 YOLO 的网格化机制和 Faster RCNN 的锚点框机制，引入了多尺

度特征层进行检测，为了能够实现对于任意大小目标的检测，去掉了 YOLO 中的全连接层。上述的修改进一步提高了一步法检测器的检测精度，其中对于较小尺度物体的提升更甚。SSD 与之前检测器的差别在于，前者通过引入多尺度特征层，在网络的不同特征层检测不同尺度的对象，而后者仅在神经网络的最深层进行检测。

RetinaNet 作为当下基于锚点框的一步法检测器的代表，其核心的创新点在于提出了 Focal Loss。Focal loss 建立在交叉熵损失（Cross Entropy Loss）的基础上，从给不同样本的损失函数赋予不同权重的角度出发，缓解目标检测中类别不平衡和正负样本不平衡在训练过程中造成的影响。在模型层面而言，RetinaNet 并没有很大的创新度，细节改动有如下几点：为了加速模型的推理，将 FPN 的输入特征层进行了后移；锚点框的设置机制更加丰富；模型分类和回归部分的子网络参数量较大；模型分类部分初始化更加合理等。

■ 二步法检测器

二步法的检测器通常会比一步法检测器具 有更复杂的网络结构。Faster RCNN 为经典的二步法检测器处理流程，首先使用卷积神经网络提取输入样本的特征，然后将预先设置好的先验包围框和样本特征送入区域候选网络得到被判定为前景的候选框，通过 RoIPooling 层提取候选框在特征上对应的候选框特征，最后根据候选框特征再次进行分类和回归以得到预测结果。相比于 Fast RCNN[20]，Faster RCNN 重要的改进点是摒弃选择性搜索（Selective Search），采用区域候选网络更加高效地提取候选区域。区域候选网络利用卷积层特征共享的特性，通过滑动窗口机制快速获得多尺度的候选区域特征。

为了解决 Faster RCNN 在处理待检测目标多尺度变化问题时的不足，后续科研工作者们提出了特征金字塔（Feature Pyramid Network，FPN）。以往的二步法检测器均基于单个高层特征进行目标检测，但是这种做法一个明显的缺陷就是不利于小目标检测。FPN 通过和主干卷积神经网络进行结合，将主干网络对应的多层次特征侧向提取，通过最近邻插值扩大当前特征图和上层特征保持大小一致，再以残差处理的形式和上层特征融合，最后在新的多层次特征上进行后续检测流程。FPN 多层特征融合的形式显著地增强了浅层特征的语义信息，在小目标检测精度上的提升尤为明显。现在 FPN 已经成为目标检测模型的标准配置，并且扩展到其他领域仍然

被验证有效。

对于面向海上快速救援的弱小人体目标检测的研究中，主要侧重于在算法层面，并没有专门弱小目标感知进行模型上的修改。因此，算法验证主要基于经典的二步法检测器 Faster R-CNN-FPN，同时为验证所提出的算法具有和检测器无关的特性，会再次选用一步法检测器 RetinaNet 作为验证。

2.1.3　小目标检测算法研究现状

通用目标检测的发展日益成熟，而近几年来，越来越多的科研工作者开始着手小目标检测的研究，对于小目标检测算法的研究主要分为神经网络结构、训练策略、锚点框补偿机制、数据增广四类，分别如下所述。

■　神经网络结构

小目标检测在神经网络结构方面的研究旨在通过设计特征融合的模型结构，使得神经网络学习到对于小目标识别更加鲁棒、更具有判别性的特征。如上节提及的 FPN，将顶层特征通过最近邻插值的方式和底层特征对齐并相加，而且相加融合后的每层分别进行检测，虽然影响了整体算法的推理速度，但是融合特征有利于检测小目标。除了融合神经网络不同层次的特征，TridentNet[21]还创造性地引入了多个不同感受野的分支以融合特征，同样有利于小目标检测。在 RFBNet[22]中，更加细致地集成了不同膨胀率（Dilation Rate）的空洞卷积（Dilated Convolution）[23]得到的特征以提高检测能力。特征融合的方法可以一定程度上增大特征的感受野（Reception Field），不仅仅提高了检测器在小目标上的性能，普遍提高了所有尺度上目标的性能。

■　训练策略

训练策略方面的研究旨在影响神经网络的学习过程，如果推理中不附加多尺度测试这类的涨点小技巧的话，并不会影响模型最后的推理时间。多尺度训练的思想是训练阶段使用随机的多尺度样本，从而使检测器在面对目标不同大小情况出现时维持稳定检测的特性，使其具有尺度不变性。作为多尺度训练的改进版本，SNIP[24]提出了在多尺度训练过程中对物体的尺度进行规范化，只有尺度在固定范围内的目标包围框参与训练。SNIPER[25]在 SNIP 上进行改进，论文中以适当的比例处理包围框的背景区域，并将其命名为 chips。训练过程中，chips 的数量会根据背景的多样

性而自适应地变化。

通过设计训练策略以平衡数据形式，Stitcher[13]提出了一个损失驱动的数据变换策略，根据小目标损失所占的比例来提供不同形式的训练数据。如果小目标损失所占比例过小，会通过拼接固定数量样本组成新的样本送入到检测器中，训练中会调整样本大小在固定范围内，从而人工造就了包含小目标的样本。同样的形式在YOLO 中作为数据增强，被称为马赛克。

■ 锚点框补偿策略

锚点框的补偿策略是为了解决小目标召回率低的问题而对锚点框预先设置的研究。Faceboxes[26]通过分析神经网络底层的锚点框过于稀疏，引入了锚点框密化策略来补偿小尺度锚点框的采样密度。策略具体的做法为对神经网络底层的小尺度锚点框进行稠密化，通过每个锚点框的中心根据不同特征图大小进行不同程度的偏移以提高召回率。S3FD[27]提出了一种尺度补偿的锚点框匹配机制，来改进目标检测网络的预测特征层，并且通过不同特征层等比例设置更加合理的锚点框。引入尺度补偿的锚点框匹配策略增加了正样本锚点框的数量，提高了人脸的召回率。

■ 数据增强

从小目标在数据集中样本数量少的角度出发，一种数据增强策略[28]被提出来提升小目标的检测性能，该方法通过利用 MS COCO 中实例分割的掩码（Mask）标注，在样本图像中多次复制粘贴来解决缺乏位置多样性的问题。

2.2 基于多模态互感的无人机跟踪

2.2.1 目标跟踪数据集

现有研究中，大多数跟踪器均基于可见光（RGB）信息，弱光条件下可能导致结果误判。其它部分研究虽然使用红外（TIR）信息，但受制于低分辨率，信息不足。因此，我们考虑将可见光和红外信息多模态光学信息进行结合，针对复杂环境下的无人机目标进行跟踪，并构建了首个反无人机多模态跟踪数据集 Anti-UAV。

学术界在针对无人机的检测跟踪领域的研究尚属空白，目前还不存在相关工作提出公开可用的高质量反无人机基准数据集，因此我们提出了反无人机多模态跟踪数据集 Anti-UAV。无人机跟踪属于目标跟踪（Object Tracking）的特定子领域，旨

在基于实时动态视频流数据精准定位无人机，完成反无人机的感知技术。

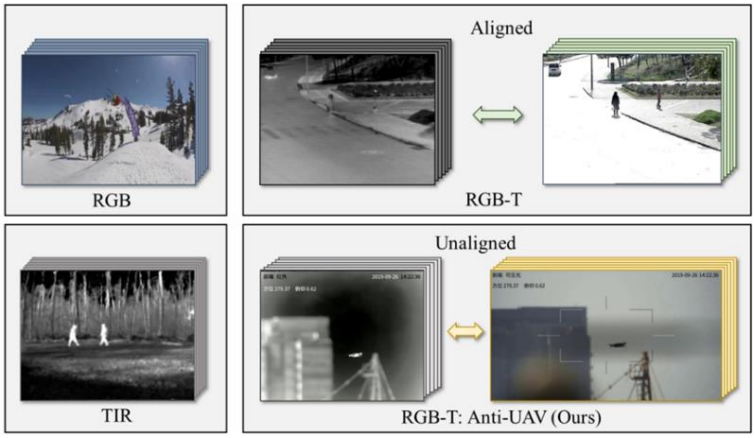


图 2.3　跟踪数据集的展示

Figure 2.3　Overview of tracking datasets

如图 2.3 所示，Anti-UAV 数据集和当前主流的目标跟踪数据集最大的区别在于：首先，Anti-UAV 针对无人机管控的应用背景只专注于跟踪无人机，没有像其他的目标跟踪数据集致力于通用目标的跟踪。其次，Anti-UAV 综合利用了可见光和红外信息的各自的优势，包含成对的无人机多模态光学信息的跟踪视频。除此之外，不同于以往的 RGB-T 目标跟踪数据集，可见光和红外的视频并没有进行数据对准，模拟了实际情况中光电摄像机和红外成像仪无法设置在同一中轴上的场景。

为了更好的阐述 Anti-UAV 和其余公开发布的单目标跟踪数据的区别，从数据模态的角度将数据集划分为三类，如表 2.2 所示：

1） 针对 RGB 目标跟踪数据集，选取了 OTB2013[29]、OTB2015[30]、VOT2014[31]、VOT2017[32]、VOT2019[33]、ALOV++[34]、TC128[35]、NUS_PRO[36]、OxUxA[37]、UAV123[38]、UAV20L[38]、Nfs[39]、LaSOT[40]、TrackingNet[41]、GOT-10k[42]。

2） 针对 TIR 目标跟踪数据集，选取了 OCU-T[43]、PDT-ATV[44]、BU-TIV[45]、ASL-TID[44]、TIV[45]、LTIR[46]、VOT-TIR16[47]、PTB-TIR[48]、LSOTB-TIR[49]。

3） 针对 RGB-T 目标跟踪数据集，选取了 OSU-CT[50]、LITIV[51]、GTOT[52]、RGBT210[53]、RGBT234[54]作为对比。

表 2.2 Anti-UAV 和其他单目标跟踪数据集的对比

Table 2.2 A comparison of Anti-UAV with other single object tracking (SOT) datasets

数据集		全部		训练集		测试集		属性
		序列	包围框	序列	包围框	序列	包围框	
可见光	OTB2013	50	29.4k	-	-	50	29.4k	11
	OTB2015	100	59k	-	-	100	59k	11
	VOT2014	25	10k	-	-	25	10k	5
	VOT2017	60	21k	-	-	60	21k	5
	VOT2019	60	19.9k	-	-	60	19.9k	5
	ALOV++	314	16k	-	-	314	16k	14
	TC128	128	55k	-	-	128	55k	11
	NUS_PRO	365	135k	-	-	365	135k	12
	OxUxA	366	155k	-	-	366	155k	6
	UAV123	123	113k	-	-	123	113k	12
	UAV20L	20	59k	-	-	20	59k	12
	Nfs	100	38k	-	-	100	38k	9
	LaSOT	1.4k	3.3M	1.1k	2.8M	280	685k	14
	TrackingNet	31k	14M	30k	14M	511	226k	15
	GOT-10k	10k	1.5M	9.3k	1.4M	420	56k	6
红外	OSU-T	10	0.2k	-	-	10	0.2k	-
	PDT-ATV	8	4k	-	-	8	4k	-
	BU-TIV	16	60k	-	-	16	60k	-
	ASL-TID	9	4.3k	-	-	9	4.3k	-
	TIV	16	63k	-	-	16	63k	-
	LTIR	20	11.2k	-	-	20	11.2k	5
	VOT-TIR16	25	14k	-	-	25	14k	10
	PTB-TIR	60	30k	-	-	60	30k	9
	LSOTB-TIR	1400	606k	1280	524k	120	82k	12
可见光-红外对	OSU-CT	6	17k	-	-	6	17k	-
	LITIV	9	6.3k	-	-	9	6.3k	-
	GTOT	50	15.8k	-	-	50	15.8k	7
	RGBT210	210	210k	-	-	210	210k	12
	RGBT234	234	233.8k	-	-	234	233.8k	12
	Anti-UAV(Ours)	318	585.9k	160	294.4k	91	168.4k	7

■ 基于 RGB 信息的目标跟踪数据集

RGB 图像作为人们生活中最常见的数据模态，在目标跟踪领域也同样为主流的研究方向。根据跟踪视频的特性，可以将 RGB 目标跟踪分为短时跟踪（Short-term Tracking）和长时跟踪（Long-term Tracking）两类。短时跟踪范围内的视频序列通常视频帧数较少，且跟踪目标极少会出现完全遮挡和消失视野内很久的情况，而长时跟踪的视频根据不同数据集最长可持续十几分钟。

短时跟踪：早期目标跟踪的视频序列大多都为较短的视频序列，作为现在针对短时跟踪的跟踪器必须验证性能的高质量基准数据集，OTB 和 VOT 得到广泛的认可，已经作为经典的目标跟踪基准数据集。除此之外，VOT 官方科研工作者每年还

在顶级会议上举办竞赛，鼓励越来越多的科研工作者关注目标跟踪这一领域，并且 VOT 数据集同时也会针对当前跟踪器难以解决的挑战不断更新。后续的 VOT 数据集引入了旋转框的标注，进一步减少了复杂背景的影响，更好地表达了跟踪目标的状态。后续科研工作者通过集成 314 个视频序列，带有更多种标注属性，组成了数据量更大的 ALOV++。

长时跟踪：相比于短时跟踪，长时跟踪要求跟踪器在较长时间跨度的视频序列内完成跟踪目标的定位，并且解决期间出现的消失视野、完全遮挡、急剧形变等巨大的挑战。因此，基于长时跟踪的跟踪器更侧重于决策跟踪目标是否仍存在视野内并且具备重新跟踪的能力。面向更加实际的应用背景，愈来愈多的大规模长时跟踪数据集被提出，并提供了大规模的训练集。TrackingNet 从 YouTube BB[55]中选择了大约 30000 个视频序列构成训练集。同时，TrackingNet 收集了 511 个视频构成测试集，测试集中视频序列的跟踪目标的类别和训练集相同。LaSOT 收集并手工标注了 14000 个视频用于构建高质量的长时单目标跟踪数据集，同时视频序列提供对应的自然语言描述，致力于探索视觉和语言的交互。

GOT-10k 涵盖了更广泛的跟踪目标类别，提供更加精细的额外标注。同时，它的评测标准要求所有跟踪器使用相同的训练集以保障公平的比较，并且在测试的时候保障训练集和测试集不会出现相同的类别，以此检验跟踪器的泛化能力。

■　基于 TIR 信息的目标跟踪数据集

由于弱光条件下 RGB 图像受到严重影响，科研工作者开始研究在这种条件下表现良好的基于 TIR 图像的目标跟踪。由 16 个红外视频序列组成的 BU-TIV 不仅仅适用于目标跟踪，还可以在运动估计、计数等其他 TIR 视觉任务。作为第一个标准的 TIR 目标跟踪基准数据集，LTIR 一共包含 20 个视频序列和 6 个目标类别，以及对应的一套评测工具箱。VOT-TIR16，是 LTIR 的扩展版本，具备更多的视频序列和目标类别，比 LTIR 更具挑战性。和 VOT 的 RGB 目标跟踪数据集一样，VOT-TIR16 同样具有针对 TIR 图像的属性标注，供跟踪器评测在各个指标上的性能。相比于以上的 TIR 目标跟踪数据集，LSOTB-TIR 作为一个大规模且多样性丰富的 TIR 目标跟踪基准，具备总数超过 600k 帧的 1400 个 TIR 视频序列，同时还提供专门的训练集。

■ 基于 RGB-T 信息的目标跟踪数据集

RGB-T 目标跟踪数据集通常由成对的 RGB 视频序列和 TIR 视频序列组成，其中 RGB 图像容易受光照条件的影响，却呈现出更多图像细节，红外图像不受光照条件的限制，缺失图像纹理信息。因此，一部分科研工作者利用 RGB 和 TIR 图像的互补性来进行目标跟踪，综合提高了目标跟踪的性能，同时可以削弱跟踪器对于光照、雨雾灯等因素影响。

作为最初的公开的 RGB-T 目标跟踪数据集，OSU-CT 数据集一共包含 6 对 RGB-T 视频序列，但是多样性较低，在这上面跟踪器的性能验证说服力并不有力。RGBT210 由通过移动平台录制的 210 个 RGB-T 视频序列对组成，充足的数据量大大丰富了数据集的多样性。作者后续又收集了 234 对 RGB-T 视频序列构建了更加大型的、更具有挑战性的 RGB-T 目标跟踪数据集 RGBT234，同样提供了基线算法的性能评测、特殊属性标注和相应评估指标。

与上述的 RGB-T 目标跟踪数据集相比，Anti-UAV 致力于反无人机感知的研究，因此数据集内物体的类别只有无人机；同时收集了更多的视频序列对来保证数据集的多样性，具备专门的无人机跟踪训练集以利于后续反无人机的研究；值得注意的是 Anti-UAV 中的视频序列对没有进行配准操作，相比于其他 RGB-T 目标跟踪数据集将具有更大的挑战性。

2.2.2 基于深度学习的单目标跟踪研究

依托于深度学习优秀的表征能力，Wang 等人[56]首次将深度学习引入到目标跟踪任务中，创造性地提出了第一个基于深度学习的目标跟踪框架。自此以后，基于深度学习的跟踪器百花齐放，发展至今已经占据了主流的地位。以下将根据不同的模态形式对跟踪器进行分析：

■ 基于 RGB 图像的跟踪器

虽然卷积神经网络已经被广泛地应用于视觉目标跟踪[57]，但是近年来，其他的网络结构的引入同样可以有效地提高跟踪器的效率和鲁棒性。因此，我们分别将其归纳为基于卷积神经网络、孪生神经网络（Siamese Neural Network，SNN）、循环神经网络（Recurrent Neural Network，RNN）和生成对抗网络（Generative Adversarial Network，GAN）的跟踪器组成。

基于 CNN 的跟踪器：通过 CNN 独立地对视频序列帧之间进行分层处理，来获得目标的特征表示。然而，传统的 CNN 有其固有的致命伤，例如需要大量带有标注的数据的支撑、忽略相邻帧之间的时序信息以及在线更新的计算复杂度较高。MDNet[58]的网络结构由共享和域相关的（Domain-specific）全连接层组成，因而将目标跟踪建模成一个多域学习的分类任务。训练中，使用大量视频序列训练网络的共享层驱使网络学到具备泛化性的特征表示，而在面对新的跟踪目标时，通过结合共享层和全新微调的全连接层将目标从背景中区分出来。为了解决 MDNet 速度慢的问题，RT-MDNet[59]通过修改网络结构并引入自适应的 RoI Align[60]加速网络的推理，拉大不同域实例级别的特征进一步促进了模型的表征学习。

基于 SNN 的跟踪器：给定目标特征模板和较大的搜索区域对，SiamFC[61]通过 SNN 计算互相关性来生成相似度的得分图，根据得分图响应最大的位置定位目标，充分利用端到端的学习达到了目标跟踪的实时运行。受目标检测的启发，SiamRPN[62]将目标跟踪建模为单样本学习（One-shot Learning）的目标检测任务从而引入了 RPN 网络。以第一帧的给定跟踪目标为模板，在后续视频帧内寻找与其相似的目标。从像素填充（Padding）和网络对称性的角度出发，SiamRPN++[63]重新设计了特征提取网络并引入了多层特征的融合，整体性能获得了较大的提升。

基于 RNN 的跟踪器：除利用相邻帧之间目标在时域和空域上的运动信息外，基于 RNN 的跟踪器同时避免了对于预训练好的 CNN 模型进行微调，一定程度上减少了训练时间并尽可能避免了过拟合的出现。这些方法的目的可分为利用上下文背景信息处理复杂背景[64]、使用多层级注意力机制[65]等。

基于 GAN 的跟踪器：GAN 可以被用来在特征空间中增强正样本或生成多样化的路径，来缓解训练过程中样本分布不平衡的现象[66]。

■ 基于 TIR 图像的跟踪器

随着 CNN 在计算机视觉领域的兴起，一些工作开始引入 CNN 来提高 TIR 跟踪器的性能。利用 VGGNet[67]的深层特征，MCFT[68]将大量数据学习到的更鲁棒的特征和相关滤波集成到一个 TIR 跟踪器中。充分挖掘视频序列的信息，LMSCO[69]整合了外观特征和运动特征进行跟踪。作为一种多级相似模型，MLSSNet[70]引入了 SNN 进行更鲁棒的 TIR 目标跟踪。基于 TIR 目标跟踪的粒子滤波框架，一种掩模

稀疏表示的深度外观模型[71]被提出以代替以往的随机抽样来搜索有用的候选区域。

■ 基于 RGB-T 图像的跟踪器

为了综合利用 RGB 图像和 TIR 图像的互补能力来提高多模态目标跟踪的性能，融合 RGB-T 图像进行目标跟踪的算法[72]应运而生。如图 2.4 所示，根据融合的位置不同。可以分为像素级别、特征级别、决策级别三种。

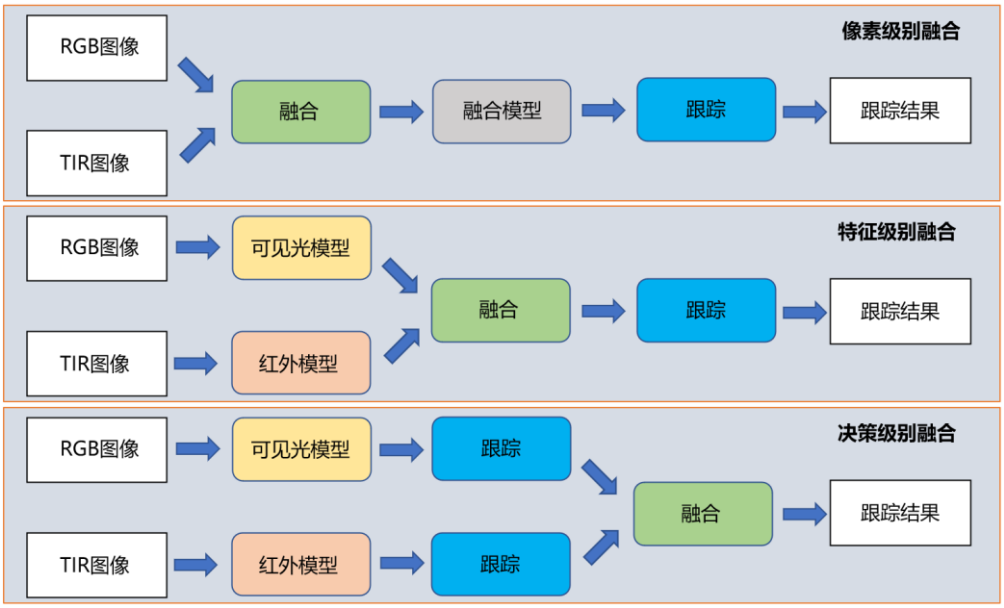


图 2.4 RGB-T 目标跟踪的不同融合方式

Figure 2.4 The different fusion levels of RGB-T object tracking

像素级别融合（Pixel-level Fusion Tracking）：通过对于不同形态的图像进行融合，产生具有更多信息量的图像，然后根据融合后的样本进行目标跟踪，如图 2.4 顶部所示，也称为跟踪前融合（Fusion-before- tracking）方法。像素级别融合跟踪易于实现，然而，图像融合方法的选取会很大程度上影响最后的跟踪结果。除此之外，像素级别的图像融合保留了源图像中的大部分信息，因此比较耗时，可能会大大减缓整个融合跟踪器的推理速度。

特征级别融合（Feature-level Fusion Tracking）：首先提取 RGB 图像和 TIR 图像的特征，然后根据预先设计的融合规则得到融合后的特征，最后使用融合特征进行后续跟踪，如图 2.4 中部所示。通常情况下，融合后的特征作为多模态的特征，比单独的特征信息更丰富，相比于像素级别融合跟踪更加直观。特征级别融合跟踪的关键在于 RGB 和 TIR 图像的提取和有效融合。

决策级别融合（Decision-level Fusion Tracking）：如图 2.4 底部所示，首先在各个模态图像中分别进行跟踪，然后融合结果以获得最终的跟踪结果，因此也被称为融合前跟踪（Tracking-before-fusion）方法。决策级别融合跟踪与像素级别和特征级别相比，计算量较小，跟踪速度更快，对 RGB 和 TIR 图像的区域性要求低。

2.3　本章小结

本章从基于无人机航拍的弱小人体目标检测和基于多模态互感的无人机目标跟踪两个角度出发，深入剖析当前关键弱小目标感知技术相关的内容。在面向海上快速救援的弱小人体目标检测方面，首先分析了共同发布的弱小人体目标检测数据集 TinyPerson 和相近公开数据集之间的异同，归纳了深度学习相关的目标检测经典模型，并就小目标检测的各个研究方向进行了全面详细分析。在面向无人机管制的无人机目标跟踪方面，描述了发布的反无人机目标跟踪数据集 Anti-UAV 同其他数据集间的关联，并对不同数据模态下基于深度学习的单目标跟踪进行了概述。

第3章　基于无人机航拍的弱小人体目标检测

目前在弱小目标检测领域的研究还远远不够，使用深度学习的方法需要数据的支持，而当前的公开数据集和研究所需的要求存在差距。依托于海上快速救援的应用背景，本章首先介绍了发布的公开的弱小人体目标检测数据集 TinyPerson 的具体情况及相应评测指标。从数据集体量的角度出发，探究了通过有效的预训练策略提高目标检测模型性能的方法。在此基础上，介绍了基于精细尺度匹配的弱小目标检测预训练策略，最后在 TinyPerson 数据集上评测算法的性能并做分析。

3.1　弱小人体目标检测数据集 TinyPerson

3.1.1　数据集介绍

■　数据收集

为了充分利用互联网的优势，根据 TinyPerson 数据集应用背景的定位，我们通过 bilibli、YouTube 等大型视频平台和百度、必应等搜索引擎，以沙滩、海滩等关键词，从中收集了大量相关的高清视频和图片，如图 3.1 所示，为了更好地诠释弱小人体目标的含义，对存在人体目标的区域进行了放大展示。为了尽可能地降低数据集的同质性，在收集的视频数据每隔 50 帧进行采样并加入 TinyPerson 的初筛版本内，然后再从整体角度对于同质性进行二次筛选。

经过多级筛选之后，TinyPerson 数据集包含 1610 张图片。在进行训练集、测试集划分的时候，首先按照视频序列进行划分，这样保证了属于同一个视频序列的图片只会在训练集或者测试集中之一出现，保持差异性，然后尽可能保证训练集和测试集的数据量相等。

■　数据处理

数据集以海边场景为主，设定了三种数据集标注类别：背景为陆地的人、背景为水面的人和忽略区域。为保证弱小人体目标数据集的质量，我们将较为棘手的场景标注为忽略区域：图像中观感近似像人但无法确定的目标、观感模糊聚集在一起无法很好区分的人群、水面人类的倒影、图像来源的标志。

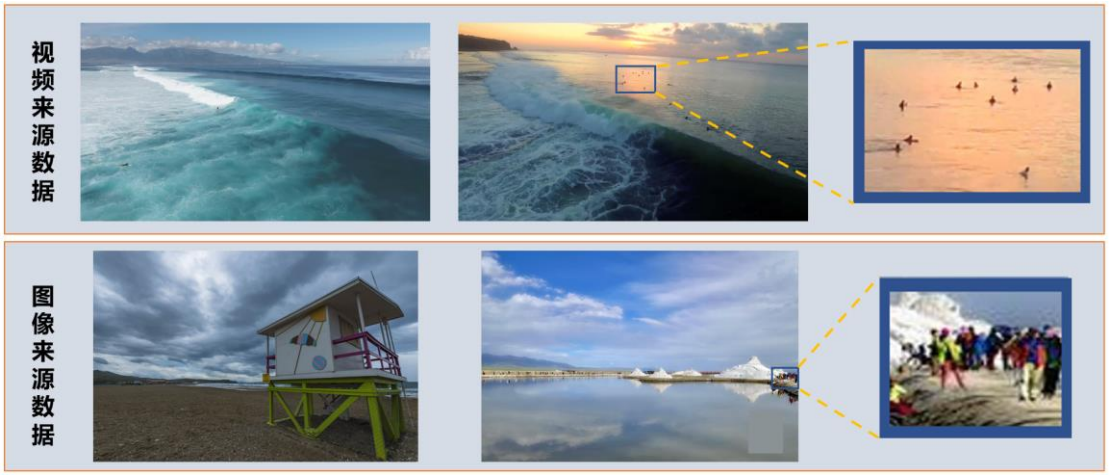


图 3.1　TinyPerson 数据集样本展示

Figure 3.1　The display of TinyPerson dataset

根据上述的标注规则，TinyPerson 的标注情况如表 3.1 所示，可以发现数据集中，背景为水面的人总数相对占据更大的比例，但是在训练集和测试集中数量分布相差较多，并不如背景为陆地的人这一类别均匀。

表 3.1　TinyPerson 数据集的标注情况

Table 3.1　The annotation of TinyPerson

标注类别	训练集	测试集	总计
背景为陆地的人	15k	15k	30k
背景为水面的人	26k	16k	42k
忽略区域	3k	2k	5k

由于数据集中图片较大，使用原图进行训练会出现显存溢出的现象；由于人体目标较小，采用缩小图片的策略也不可取。因此，我们采取了切图训练的策略：使用固定大小切割原图，并且相邻子图之间会有一定的重合度保证目标肯定会完整地出现在某一张子图中。

3.1.2　评测指标

为了综合地评估检测器的性能，我们需要确定检测器对于目标定位的能力的强弱，因此 TinyPerson 数据集选取了平均精度（Average Precision，AP）和丢失率（Miss Rate，MR）作为评测指标。在此之前需要引入交并比（Insertion Over Union，IOU）、精准率（Precision）和召回率（Recall）的概念。

表 3.2　TP、FP、TN 和 FN 的定义

Table 3.2　The definition of TP、FP、TN and FN

检测结果的包围框	包围框得分大于置信度	包围框得分小于置信度
与真值包围框的 IOU 大于设定阈值	TP	FN
与所有无重复的真值包围框的 IOU 均小于设定阈值	FP	TN

IOU 通过两个包围框的重叠区域和总区域的比值，表达两者之间的重叠程度。根据表 3.2 所示的真正例（True Positive，TP）、假正例（False Positive，FP）、真负例（True Negative，TN）和假负例（False Negative，FN）的概念，精确率表示的是判别为正例的样本（TP 和 FP 的总和）中真正例（TP）的比例：

$$Precision = \frac{TP}{TP + FP}. \tag{3.1}$$

召回率表示样本中的正样本（TP 和 FN 的总和）中真正例（TP）所占比例：

$$Recall = \frac{TP}{TP + FN}. \tag{3.2}$$

■　平均精度

根据上面的 Precision 和 Recall 的描述可以发现两者是互相矛盾的，为了综合评价检测器的性能，AP 应运而生。通过 Recall 为横轴，Precision 为纵轴定义 PR 曲线，而 PR 曲线下和横轴、纵轴围成的图形的面积即为 AP，因此 AP 越高表明检测器性能越好。

根据数据集中人体目标的大小（根号下相应面积大小），首先划分为三个区间：[2，20]为 tiny object 的尺度范围，[20，32]为 small object 的尺度范围，[2，+∞]为整个数据集内目标的尺度范围。由于大部分目标均处于 tiny object 的范围内，我们对该范围进行了更加细致地划分：尺度范围在[2，8]内的目标标号为 tiny1，在[8，12]内的目标标号为 tiny2，在[12，20]内的目标标号为 tiny3。考虑到特殊需要，判定正负例的 IOU 阈值除 0.5 外还有 0.25 和 0.75 的选项。

因此我们可以得到如下的 AP 评价指标，如表 3.3 所示。

表 3.3　评价指标 AP 的定义

Table 3.3　The definition of evaluation metric AP

评价指标	定义
AP_{50}^{tiny1}	尺度范围在[2，8]内、IOU 判定阈值为 0.5 的目标的平均精度
AP_{50}^{tiny2}	尺度范围在[8，12]内、IOU 判定阈值为 0.5 的目标的平均精度
AP_{50}^{tiny3}	尺度范围在[12，20]内、IOU 判定阈值为 0.5 的目标的平均精度
AP_{50}^{small}	尺度范围在[20，32]内、IOU 判定阈值为 0.5 的目标的平均精度
AP_{50}^{tiny}	尺度范围在[2，20]内、IOU 判定阈值为 0.5 的目标的平均精度
AP_{50}^{tiny}	尺度范围在[2，20]内、IOU 判定阈值为 0.5 的目标的平均精度
AP_{25}^{tiny}	尺度范围在[2，20]内、IOU 判定阈值为 0.25 的目标的平均精度
AP_{75}^{tiny}	尺度范围在[2，20]内、IOU 判定阈值为 0.75 的目标的平均精度

■　丢失率

MR 可以简单定义为$1-Recall$，但是在弱小目标检测数据集 TinyPerson 中由于存在大面积的忽略区域，使用 IOU 进行计算并不合适，因此引入了和检测结果的交占比（Insertion Over Detection）的概念，如图 3.2 所示。为了辅助理解，将 IOU 和 IOD 进行了对比，相比于 IOU 表达的是检测包围框和真值框之间的重叠程度，IOD 表示检测包围框和忽略区域的相交程度。

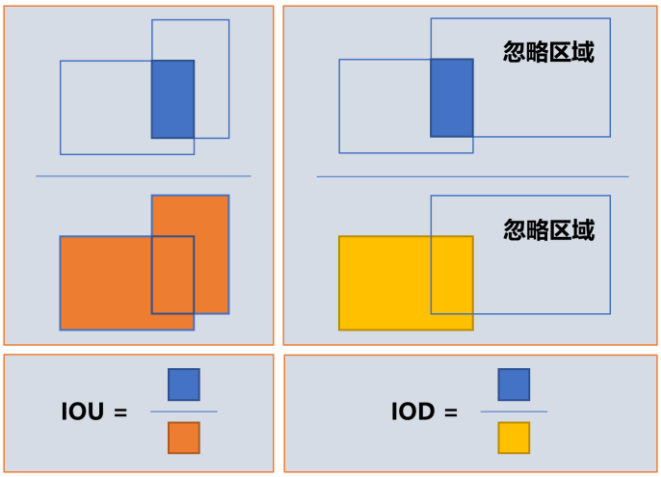


图 3.2　IOU 和 IOD 的定义

Figure 3.2　The definition of IOU and IOD

因此，根据 TinyPerson 数据集的目标尺度，我们同样可以得到如表 3.4 所示的 MR 评价指标。

表 3.4　评价指标 MR 的定义

Table 3.4　The definition of evaluation metric MR

评价指标	定义
MR_{50}^{tiny1}	尺度范围在[2，8]内、IOD 判定阈值为 0.5 的目标的丢失率
MR_{50}^{tiny2}	尺度范围在[8，12]内、IOD 判定阈值为 0.5 的目标的丢失率
MR_{50}^{tiny3}	尺度范围在[12，20]内、IOD 判定阈值为 0.5 的目标的丢失率
MR_{50}^{small}	尺度范围在[20，32]内、IOD 判定阈值为 0.5 的目标的丢失率
MR_{50}^{tiny}	尺度范围在[2，20]内、IOD 判定阈值为 0.5 的目标的丢失率
MR_{50}^{tiny}	尺度范围在[2，20]内、IOD 判定阈值为 0.5 的目标的丢失率
MR_{25}^{tiny}	尺度范围在[2，20]内、IOD 判定阈值为 0.25 的目标的丢失率
MR_{75}^{tiny}	尺度范围在[2，20]内、IOD 判定阈值为 0.75 的目标的丢失率

3.2　预训练策略研究现状

一个合理的预训练策略可以为下游任务提供更好的模型初始化。一般来说，现有的预训练策略可以采用两种形式：监督学习（Supervised Learning）和无监督学习（Unsupervised Learning），如图 3.3 所示。对于计算机视觉任务来说，图像存在很多的冗余信息，具有信息维度高、信息密度低的特点。在这过程中，神经网络就起到了一个信息压缩的作用。

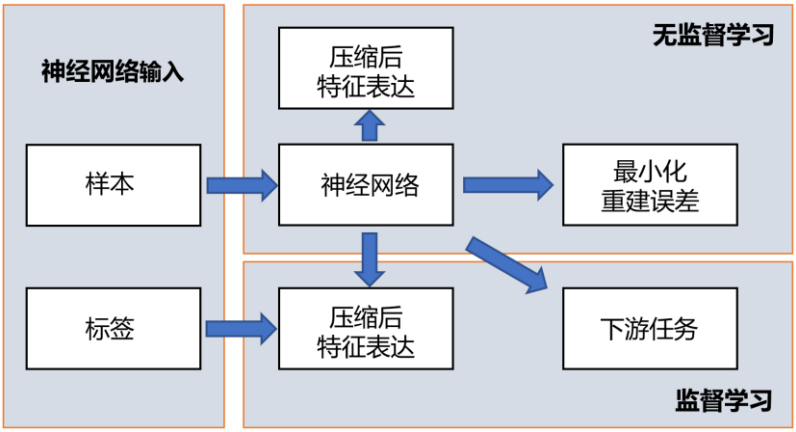


图 3.3　监督学习和无监督学习的区别

Figure 3.3　The difference between supervised learning and unsupervised learning

监督学习利用大量带有标注的数据来训练神经网络，将神经网络输出的特征表达向数据的真实标签进行修正，通过不断迭代地学习、修正来获得有效的压缩的特征表达，最后在当前的训练任务获得较高的性能及能够支持下游任务的特征表达模式。然而，无监督学习无需任何数据标注，通过对大量无标注数据潜在关联信息进

行探索，其中有一种比较流行的方法叫做自监督学习。自监督学习可以被看作是一种具备监督信号的无监督学习方法，这里的监督是由设计的代理任务（Proxy Task）产生的伪标签，通常使用数据集自身的各种信息来构造。通过最小化重建误差或者维持伪标签一致性来驱动神经网络学习特征表达。

■ 监督学习

在监督学习中，绝大多数算法直接加载 ImageNet 训练的模型作为模型初始化，以进行下游任务的微调。这样的做法也在 Facebook 研究人员的论文中[73]被验证有利于模型的收敛，同时减少模型的训练时间。FA-RPN[74]在人脸检测任务上使用 MS COCO 作为预训练数据集，验证有助于提高人脸检测器在 WIDER Face 上的性能。同时，Improved SRN[75]使用了相同的预训练策略也得到了验证。目前预训练使用的数据集都是已经公开的，然而，想要使用更大规模的数据集标注是耗时耗力的，许多研究人员试图构建大规模和多样化的合成数据集，来弥补真实数据集难以收集标注的缺点。需要注意的是，虽然合成数据可以较为容易地获得数据和标注，但是和真实数据集是存在一定的域（Domain）的差异。一些方法[76][77][78]在合成数据上进行预训练，然后在目标数据上对模型进一步微调。

■ 无监督学习

对于无监督学习中的预训练策略，主要采用自监督学习的形式。好的模型初始化相当于模型已经进行了常识性的学习，掌握了更加通用的特征表达。在这样的研究思想下，越来越多的预训练任务的设计方法被提出。Doersch 等人[79]将图片划分成不同的块，在给定其中一个块的位置条件下要求卷积神经网络学习预测第二个块相对于第一个块的位置。Larsson 等人[80]设计了一个着色相关的预训练任务，其中神经网络在给定 L 通道的条件下学习预测 a 和 b 通道。Kim 等人[81]将输入图片划分为块并打乱，驱使神经网络完成拼图任务来获得更加强大的多功能的特征表示。经过验证，以上方法成功地驱使模型学习到更好的特征表达，为下游任务提供了更好的模型初始化，有效地提升了模型在下游任务上的性能。

由于 TinyPerson 数据集中仅有近 800 张图片，即使经过切图处理训练的策略，数据集中样本仍然无法和其他大型公开数据集相比。考虑到 WIDER Face 和 TinyPerson 数据集间的尺度分布差异相近，即使 WIDER Face 这样体量较大的数据集

使用 MS COCO 作为预训练数据集也有性能上的提升，所以尝试探究如何在更大的数据集上进行预训练，以及更为有效的预训练方式是一个可行的技术路线。

3.3　基于精细尺度匹配的弱小人体目标检测算法

3.3.1　算法概述与创新点

从训练数据集体量的角度出发，可以使用更加契合下游任务的数据集和更合适的策略进行训练。本小节提出了基于精细尺度匹配的弱小人体目标检测预训练策略：相比于加载 ImageNet 训练模型，使用同为目标检测数据集的 MS COCO 作为预训练数据集会更加有效，使得神经网络学习到更多检测任务相关的知识；以数据集的尺度信息为出发点，尽可能减少预训练数据集和下游任务训练集之间的尺度分布上的差异，通过两个数据集间的尺度匹配算法，使得预训练数据集的尺度分布向下游任务训练集靠近，神经网络在预训练阶段学习到和下游任务更为相近的目标形式。通过将图像级别的尺度匹配推进到实例级别，更加有效地提升了检测器在下游任务上的性能。方法的创新点在于：

1）与上一节所述的预训练策略不同，提出的基于尺度匹配的预训练策略更加具有直接的指导性，探究了如何更好地利用预训练数据集，有效地促进了预训练数据集与目标数据集之间的相似性，来为下游任务提供更好的模型初始化。

2）基于数据集间尺度分布越相似越有利于神经网络学习到下游任务相关知识的角度，将共同发布的基于图像级别的尺度匹配算法上升到基于实例级别的尺度匹配算法，更加精细的尺度匹配算法进一步减少了匹配过程中的近似误差，为下游任务提供了更好的网络模型初始化。

3.3.2　算法介绍

为了减少预训练数据集和下游任务训练集的尺度分布上的差异，以下游任务训练集的尺度分布为目标，通过尺度匹配的方式将预训练数据集的尺度分布进行迁移，从而使检测器在迁移过后的新的预训练数据集上进行训练，然后再在下游任务训练集上进行训练。

值得注意的是，这里说的尺度为绝对尺度而非相对尺度，以防产生歧义混淆将进行一个简单的说明。绝对尺度为根号下目标包围框的面积的大小，如下所示，

$$as(w,h) = \sqrt{w \times h}. \tag{3.3}$$

其中，w 和 h 分别表示为目标包围框的宽和高，as 为目标的绝对尺度，后续将绝对尺度 as 简写为 s。相对尺度定义为根号下目标包围框的面积和所在图片的面积的比值的大小，如下公式 3.4 所示，

$$rs(w,h,W,H) = \sqrt{\frac{w \times h}{W \times H}}. \tag{3.4}$$

其中，w 和 h 的定义同上，W 和 H 为该目标包围框所属的图片的宽度和高度，rs 为目标的相对尺度大小。

因此，在给定额外预训练数据集 E 和目标任务数据集 T 时，可以根据统计信息得到两者各自的尺度分布$P_{size}(s;E)$和$P_{size}(s;D)$，通过施加尺度匹配的变换 T 在额外预训练数据集 E 上，使得新预训练数据集 $T(E)$的尺度分布向目标任务数据集靠近，如下公式所示，

$$P_{size}\big(s;T(E)\big) \approx P_{size}(s;D). \tag{3.5}$$

提出的精细尺度匹配算法从图像和实例出发，分为图像级别的尺度匹配（Image-level Scale Match）和实例级别的尺度匹配（Instance-level Scale Match）两种思想。由于实例级别的尺度匹配通过减少匹配过程中的近似操作进一步减小两个数据集间的尺度分布差异，所以将图像级别的尺度匹配简记为 SM 算法，而将实例级别的尺度匹配简记为 SM+算法。在此之前，需要声明一下 SM 和 SM+算法在尺度匹配过程均可以采取随机尺度匹配（Random Scale Match，RSM）和单调尺度匹配（Monotone Scale Match，MSM）两种实现方式。

■ 尺度匹配实现方式

依据机器学习的假设条件，数据集训练集的分布近似于实际的现实场景分布，所以通过统计下游目标任务训练集的尺度分布$P_{size}(s;D_{train})$可以近似于得到了$P_{size}(s;D)$。因为在计算机中数据均以离散的形式进行存储，在实际匹配过程中对尺度分布进行了量化，借用了离散直方图的形式以辅助尺度匹配。由于数据集的尺度分布呈现长尾分布的形式，为了更加高效地建立直方图，将尺度分布中统计量过

小的较小头部尺度部分和较大尾部尺度部分分别进行了合并，以得到了修正过后的尺度直方图 H。

在预训练过程中，每次获得目标尺度 s 将从尺度直方图 H 中进行采样，采样可以采用随机尺度匹配和单调尺度匹配两种形式。其中，由于随机采样而带来的多样性，可能会使得尺度较小的目标采样到较大的尺度而产生较为模糊的训练样本，反之亦然，所以单调尺度匹配的形式应运而生。两者的区别如下所示：

随机尺度匹配：在匹配的过程中，首先在尺度直方图 H 众多矩形区间种进行第一次采样，获得采样到的矩形区间的索引。根据索引可以得到对应区间的上限和下限，然后再次在区间内进行均匀采样获得最后目标的尺度大小$\hat{s}$以进行尺度变换。

单调尺度匹配：通过预先统计的方式，采用单调函数的形式将额外预训练数据集 E 的每一个尺度就恒等映射到下游任务数据集训练集D_{train}上，如下公式所示，

$$\int_{min(s)}^{s'} P_{size}(s;E)ds = \int_{f(min(s))}^{f(s')} P_{size}(\hat{s};D_{train})d\hat{s}. \tag{3.6}$$

$f(\cdot)$表示单调映射函数，s'表示额外预训练数据集上的任一尺度。替代了上述随机尺度匹配中的第一次采样，通过单调函数直接得到对应矩形区间的索引，然后根据索引在该矩形区间上下限内进行均匀采样，获得尺度大小$\hat{s}$以进行尺度变换。

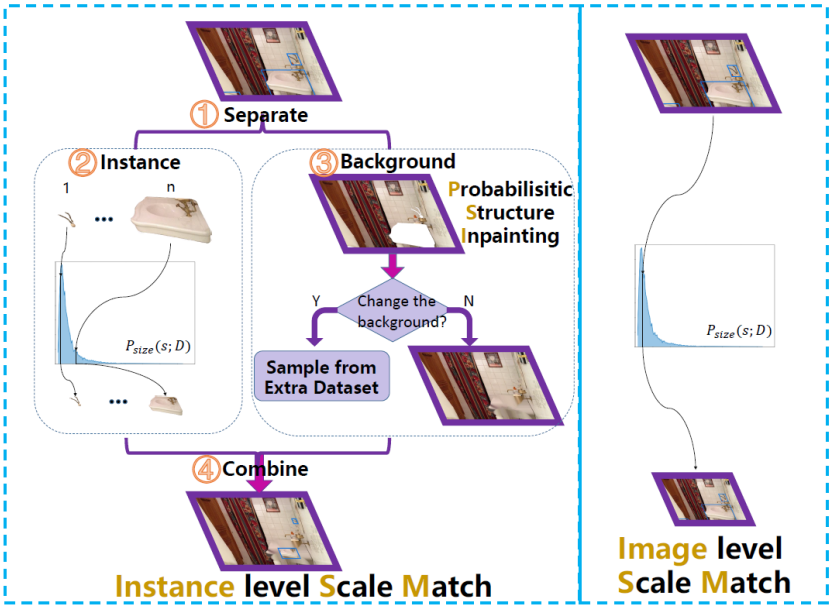


图 3.4　图像级别和实例级别的尺度匹配的区别阐述

Figure 3.4　The illustration of difference between Image-level SM and Instance-level SM

根据上述的两种尺度匹配的实现方式，可以配合基于图像级别的尺度匹配和基于实例级别的尺度匹配两种思想，两者的思路和具体区别如图 3.4 所示。

表 3.5　图像级别尺度匹配算法的细节

Table 3.5　The detail of image-level scale match algorithm

算法 1 图像级别尺度匹配算法 SM

输入：下游任务训练集D_{train}；额外预训练数据集 E；尺度直方图的区间数 k；

输出：尺度变换后的新预训练数据集 *T(E)*，记作$\hat{E}$；

注意：I_i, G_i 分别为 E 中采样得到的第 i 张图片和对应的所有标注；$ModifiedHistogram(\cdot)$的作用为根据数据集尺度分布D_{train}和设定直方图区间数 k，得到修正后的离散尺度直方图 H 和对应的区间尺度范围 R；$GetMean(\cdot)$为根据当前图片所有包围框G_i的尺度，得到平均尺度大小s的函数；$GetUniform(\cdot)$的作用为根据区间范围，得到均匀采样后的尺度$\hat{s}$；$ScaleImage(\cdot)$为根据目标缩放比例c将图片I_i和标注G_i进行缩放操作的函数；

1: $\hat{E} \leftarrow \emptyset$

2: $(H, R) \leftarrow ModifiedHistogram(D_{train}, k)$

3: **for** (I_i, G_i) *in* E **do**

4: $s \leftarrow GetMean(G_i)$

5: 采样 $k \sim H$

6: 采样 $\hat{s} \sim GetUniform(R[k]^-, R[k]^+)$

7: $c \leftarrow \hat{s}/s$

8: $(\widehat{I_i}, \widehat{G_i}) \leftarrow ScaleImage(I_i, G_i, c)$

9: $\hat{E} \leftarrow \hat{E} \cup (\widehat{I_i}, \widehat{G_i})$

■ 基于图像级别的尺度匹配

算法的细节如表 3.5 所示，给定额外预训练数据集、下游目标任务训练集和一系列相关超参数，首先根据下游目标任务训练集建立离散尺度分布直方图。然后，在每个训练循环当中，每次采样得到一张样本图片和对应的全部标注。根据样本图片内的所有实例的包围框标注，计算所有实例的尺度，从而得到整张样本图片上的平均尺度作为原尺度大小。在尺度分布直方图上通过 RSM 或者 MSM 的实现形式得到某一矩形区间的索引，根据索引获得该矩形区间的尺度范围，再次在尺度范围内进行均匀采样以获得理想尺度大小$\hat{s}$，从而获得目标的缩放尺度 c。最后根据缩放尺度对样本图片直接进行缩放，得到缩放后对应的所有标注，使检测器在经过尺度

变换后的新样本图片上进行预训练。

■　基于实例级别的尺度匹配

为了消除 SM 算法中尺度变化的近似操作以达到更加精准的尺度匹配，利用额外数据集中实例分割的掩码标注，提出了基于实例级别的尺度匹配算法 SM+，可以将其分为提取和分离、实例级别尺度直方图匹配、概率结构修补（Probabilistic Structure Inpainting，PSI）和前后景合并四个步骤。

提取和分离：根据掩码标注，每张样本图片将进行前景和背景的分离。由于掩码标注采用边界点的存储形式，直接使用相应标注进行分离会得到锯齿形的前景实例。为了得到更加精细的前景，使用了 matting 方法[82]使得分离出来的前景的边缘轮廓线更加平滑。分离之后，得到该样本图片上的分离出来的每一个前景实例和不完整的背景如图 3.4 左侧所示，这两部分接下来将被分别进行处理。

实例级别尺度直方图匹配：首先获得预先建立好的尺度直方图，对该样本图片内的每个实例依据相应的包围框标注计算各自的原尺度s_{ij}。对于每一个实例都在尺度直方图上通过 RSM 或者 MSM 的实现形式获得某一个矩形区间的索引，从而在索引对应的尺度范围内进行均匀采样得到每个实例的理想尺度大小$\widehat{s_{ij}}$。根据比例对每个实例进行尺度变换，得到新的前景实例和相应的标注。尺度变换公式如下所示，

$$\mathrm{T} = \begin{bmatrix} c & 0 & t_x \\ 0 & c & t_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{3.7}$$

这里 c 表示理想尺度大小和原尺度大小的比例，t_x和t_y分别表示实例在变换过程中在 x 轴、y 轴上各自的偏移。

概率结构修补：为了修补背景中被抠除前景实例而带来的空洞，受 InstaBoost[83] 启发，首先尝试了采用 inpainting 策略[84]来填补背景中的空白区域。然而实际上，由于采用的预训练数据集 MS COCO 和下游目标任务数据集 TinyPerson 的尺度分布差异较大，大部分实例在尺度匹配算法的指导下会进行较大程度地缩小操作，造成背景的大面积空白。在这样的情况下，直接使用传统的 inpainting 策略进行修补的结果并不太理想，修补结果如图 3.5 顶部所示。通过可视化图像，可以发现图片的结构信息被严重地破坏，尤其当前景实例在图片中占据较大面积的时候。

为了缓解这种由实例级别目标缩放而带来的图片结构失真，引入了 PSI 策略，

即通过采样额外的新背景来代替原本结构失真的背景。这种策略通过替换新背景解决了背景结构失真的问题如图 3.5 底部所示，但是同样会引发一个新的问题：目标的背景信息会完全不相同，引发语义上的歧义性，一定程度上可能会影响网络的学习。因此，通过引入一个预先定义好的概率超参数 p 来决定是否要进行背景的替换，凭借这种方式来权衡两种背景修补方式间的最佳平衡点。

因此，具体的实现细节为当随机数大于预先设定好的概率 p 会从预训练数据集中采样一张图片作为新背景，而随机数小于等于 p 的时候则会使用传统的 inpainting 策略。值得注意的是，新背景图片上的实例被默认为背景，并不会参与训练。

前后景合并：最后得到了修复后的背景和尺度变换后的所有实例，根据新的标注信息将所有实例按照位置粘贴在背景上。算法整体流程如表 3.6 所示。

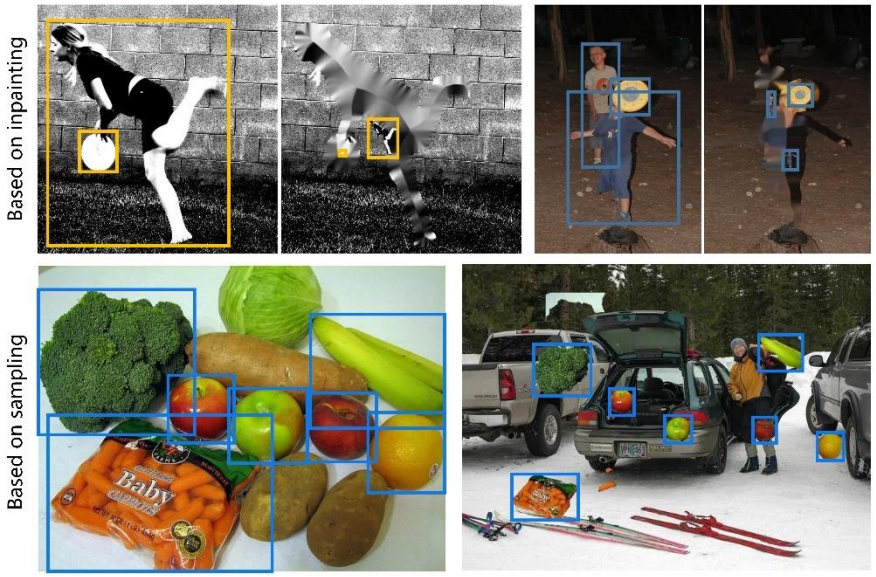


图 3.5　基于 inpainting 策略和背景采样策略的训练图片的可视化

Figure 3.5　The visualization of the training image based on different strategy

表 3.6　实例级别尺度匹配算法的细节

Table 3.6　The detail of instance-level scale match algorithm

算法 2 实例级别尺度匹配算法 SM+

输入：下游任务训练集D_{train}；额外预训练数据集 E；尺度直方图的区间数 k；

输出：尺度变换后的新预训练数据集 T(E)，记作$\hat{E}$；

注意：I_i和G_i分别为 E 中采样得到的第 i 张图片和对应的所有标注；F_{ij}和G_{ij}分别为第 i 张图片中第 j 个前景实例和对应的标注；$ModifiedHistogram(\cdot)$的作用为根据数据集D_{train}尺度分布和设定直方图区间数 k，得到修正后的离散尺度直方图 H 和对应的区间尺度范围 R；$Separate(\cdot)$的作用为根据标注G_i将图片I_i分离为前景B_i和背景F_i；$GetInstanceScale(\cdot)$为根据实例的标注G_{ij}得到对应尺度的函数；$GetUniform(\cdot)$的作用为根据区间范围，得到均匀采样后的尺度c_{ij}；$ScaleInstance(\cdot)$为根据目标缩放比例c_{ij}将前景实例F_{ij}和标注G_{ij}进行缩放操作的函数；$ProbabilisticStructureInpainting(\cdot)$是依靠概率 p 来决策是否要更换背景的函数，若随机数大于 p 则从整个数据集$I_{:}$ 中采样一张新图片作为背景，小于等于则使用原背景B_i；$Merge(\cdot)$为根据标注$\widehat{G_i}$将前景$\widehat{F_i}$和背景$\widehat{B_i}$进行合并得到新图片$\widehat{I_i}$的函数。

$\hat{E} \leftarrow \emptyset$

$(H, R) \leftarrow ModifiedHistogram(D_{train}, k)$

for (I_i, G_i) *in* E **do**

　$B_i, F_i \leftarrow Separate(I_i, G_i)$

　$\widehat{F_i} \leftarrow \emptyset$

　$\widehat{G_i} \leftarrow \emptyset$

　for (F_{ij}, G_{ij}) *in* (F_i, G_i) **do**

　　$s_{ij} \leftarrow GetInstanceScale(G_{ij})$

　　采样 $k \sim H$

　　采样 $\widehat{s_{ij}} \sim GetUniform(R[k]^-, R[k]^+)$

　　$c_{ij} \leftarrow \widehat{s_{ij}} / s_{ij}$

　　$(\widehat{F_{ij}}, \widehat{G_{ij}}) \leftarrow ScaleInstance(F_{ij}, G_{ij}, c_{ij})$

　　$\widehat{F_i} \leftarrow \widehat{F_i} \cup \widehat{F_{ij}}$

　　$\widehat{G_i} \leftarrow \widehat{G_i} \cup \widehat{G_{ij}}$

　$\widehat{B_i} \leftarrow ProbabilisticStructureInpainting(B_i, I_{:}, p)$

　$\widehat{I_i} \leftarrow Merge(\widehat{B_i}, \widehat{F_i}, \widehat{G_i})$

　$\hat{E} \leftarrow \hat{E} \cup (\widehat{I_i}, \widehat{G_i})$

3.4 实验验证

通过大量实验验证了两种基于精细尺度匹配的弱小人体目标检测算法的有效性，尤其是基于实例级别的尺度匹配算法可以更加精确地减少尺度分布的差异性。接下来将从实验配置、结果分析和消融实验三个方面来进行深入的剖析，消融实验将主要以基于实例级别尺度匹配算法 SM+为主要分析对象。

3.4.1 实验配置

基于精细尺度匹配的弱小目标检测预训练策略分为基于图像级别的尺度匹配 SM 算法和基于实例级别的尺度匹配 SM+算法两种，在每种算法的具体实现方式又可分为随机尺度匹配 RSM 和单调尺度匹配 MSM 两种，所以具体算法可以划分为 RSM、RSM+、MSM、MSM+四种。

超参数设置：实验的超参数共分为预训练阶段和微调阶段两套参数，为保证实验的公平比较，上述的四种算法均使用相同的训练参数。

在预训练阶段，加载 ImageNet 的训练模型作为模型初始化。训练过程一共有 45k 次迭代，初始学习率为 0.02，然后分别在第 30k 次和第 40k 次迭代的时候降为 0.002 和 0.0002。批大小设置为每张 GPU 上为 4 张图片，共使用 8 张 NVIDIA GeForce GTX 2080Ti。

在微调阶段，使用上一阶段的最终模型作为当前初始化。学习率起初为 0.01，整个训练包含 12 个循环，学习率将分别在第 6 个循环和第 8 个循环降为十分之一。这时的批大小设置为每张 GPU 上一张图片，共使用了 2 张 NVIDIA GeForce GTX 1080Ti。TinyPerson 数据集中单张内目标超过 200 个的图片被标定为 dense，目前这类图片并不参与训练和测试，所以将检测器最大输出的预测包围框数目设定为 200。由 TinyPerson 数据集介绍得知训练过程中为原图的子图形式，这时直接使用子图进行训练。

两个训练阶段中，分类均采用交叉熵损失，回归均采用平滑 L1 损失（Smooth L1 Loss）。锚点框的大小被设定为由聚类算法得到的（8.31，12.5，18.55，30.23，60.41）五组，长宽比设定为（0.5，1.3，2）三组。

训练数据：在预训练阶段使用 MS COCO 训练集和测试集在内的全部数据，虽然下游目标任务为检测人体目标，但是 80 个类别的所有目标均被使用参与预训练。

SM 算法会改变训练数据图像的大小，而 SM+算法只改变样本图片内前景实例的大小，因此使用原图大小进行训练。在下游任务的微调阶段只使用了弱小人体目标检测集 TinyPerson 的训练集。由于显存的限制，TinyPerson 通过切分原图的策略使用子图进行训练，在训练过程中不改变子图的大小。

检测器：为了更好地体现基于精细尺度匹配的弱小人体目标检测算法的有效性，选取了多种 state-of-the-art 检测器进行比较，如 FCOS[85]、Reppoints[86]、RetinaNet、FreeAnchor[87]、GCNet[88]、Libra RCNN[89]、Double Head[90]、Cascade RCNN[91]、Faster RCNN-FPN、SCRDet[92]、DFSD[93]。

由于 TinyPerson 数据集中人体目标过小，Reppoints 和 RetinaNet 的性能不佳，通过分析发现是检测特征分辨率过大的原因，因此将两者的 FPN 层进行了前移，由 P3-P7 前移到了 P2-P6，取得了较大的性能提升。将进行调整过后的 Reppoints 和 RetinaNet 分别命名为 Reppoints*、RetinaNet*。其他检测器保持原有参数。

3.4.2　实验结果及分析

表 3.7　各检测器在 TinyPerson 上 MR 的性能比较

Table 3.7　Comparisons of detectors in terms of MRs (%) on TinyPerson

检测器	MR_{50}^{tiny1}	MR_{50}^{tiny2}	MR_{50}^{tiny3}	MR_{50}^{tiny}	MR_{25}^{tiny}	MR_{75}^{tiny}
FCOS	99.96	99.77	97.68	99.00	97.24	99.89
Reppoints*	95.89	91.20	85.64	93.08	85.73	98.88
RetinaNet	94.52	88.24	86.52	92.66	81.95	99.13
FreeAnchor*	88.93	80.75	83.63	89.63	78.21	98.77
GCNet	90.57	85.57	82.56	89.67	84.16	98.50
Libra RCNN	90.93	84.64	81.62	89.22	82.44	98.39
RetinaNet*	89.65	81.03	81.08	88.31	76.33	98.76
Double Head	88.00	83.35	79.45	88.26	77.76	98.37
Cascade RCNN	88.70	82.87	79.11	88.26	79.62	98.40
Faster RCNN-FPN	87.86	82.02	78.78	87.57	76.59	98.39
SCRDet	98.23	94.62	89.65	95.31	88.23	99.63
DSFD	96.41	88.02	86.84	93.47	78.02	99.48
Faster RCNN-FPN-RSM	87.14	79.60	76.14	86.22	74.16	98.28
Faster RCNN-FPN-RSM+(ours)	86.81	79.87	76.85	86.26	74.29	**98.22**
Faster RCNN-FPN-MSM	86.54	79.20	76.86	85.86	74.33	98.23
Faster RCNN-FPN-MSM+(ours)	**86.47**	**78.12**	**75.83**	**85.60**	**74.13**	98.25

表 3.8 各检测器在 TinyPerson 上 AP 的性能比较

Table 3.8 Comparisons of detectors in terms of APs (%) on TinyPerson

检测器	AP_{50}^{tiny1}	AP_{50}^{tiny2}	AP_{50}^{tiny3}	AP_{50}^{tiny}	AP_{25}^{tiny}	AP_{75}^{tiny}
FCOS	0.99	2.82	6.20	3.26	13.28	0.14
Reppoints*	15.00	30.28	44.33	30.54	50.79	3.84
RetinaNet	12.24	38.79	47.38	33.53	61.51	2.28
FreeAnchor*	25.13	47.41	52.77	41.41	63.38	4.58
GCNet	28.68	45.76	53.05	43.09	61.33	5.32
Libra RCNN	27.08	49.27	55.21	44.68	64.77	6.26
RetinaNet*	27.08	52.63	57.88	46.56	69.60	4.49
Double Head	30.33	50.08	58.15	46.88	67.52	6.17
Cascade RCNN	30.89	50.75	57.83	46.97	67.01	6.00
Faster RCNN-FPN	30.25	51.58	58.95	47.35	68.43	5.83
SCRDet	4.19	18.54	36.51	21.95	51.15	1.46
DSFD	13.85	37.24	49.31	33.65	63.18	1.94
Faster RCNN-FPN-RSM	33.91	55.16	62.58	51.33	71.55	6.46
Faster RCNN-FPN-RSM+(ours)	33.74	55.32	62.95	51.46	72.38	6.62
Faster RCNN-FPN-MSM	33.79	55.55	61.29	50.89	71.28	6.66
Faster RCNN-FPN-MSM+(ours)	**34.20**	**57.60**	**63.61**	**52.61**	**72.54**	**6.72**

上述各个检测器在 TinyPerosn 上的性能如表 3.7 和表 3.8 所示，其中 MR 越小，性能越好；AP 与之相反。加粗数字表示同类评测指标内性能最好的选项。实验选取了 Faster RCNN-FPN 作为基线检测器。从 AP 的性能表可以发现 RetinaNet* 在调整了 FPN 结构后比原生的 RetinaNet 在AP_{50}^{tiny}上高 13.03 个点，通过分析认为合适的锚点框设置机制和带有丰富细节信息的高分辨率特征有利于更好地进行弱小目标检测。除此之外，增加输入图像的分辨率同样可以提升检测器的性能，但是同时会增加大量的计算量。接下来将从同其他 state-of-the-art 检测器的比较、尺度匹配算法内部的比较入手。

■ 同其他 state-of-the-art 检测器的比较

如表 3.7 和表 3.8 所示，顶部区域内的检测器均为通用目标检测器，这些检测器的主干网均使用 ImageNet 作为预训练数据集。中部区域的 SCRDet 为遥感目标检测器，其在 DOTA 数据集上取得了 state-of-the-art 的性能，而人脸检测器 DFSC 在 WIDER Face 上获得了最佳的性能。

以 AP 为例，在通用目标检测器中经过精调的 RetinaNet*和 Faster RCNN-FPN 的性能十分可观，分别在AP_{50}^{tiny}上取得了 46.56%和 47.35%。值得注意的是，最近发布的二步法通用目标检测器即使没有经过精调，但是在AP_{75}^{tiny}这个评测指标上仍高于 Faster RCNN-FPN。这个现象和这些通用目标检测器侧重于大目标定位任务上面的提升不谋而合，对于弱小目标检测任务来说，一个 IOU 在 0.75 以上的紧致的包围框对检测器要求十分高，同样这样严苛的要求使得在评测时更加关注于 IOU 阈值为 0.5 的评测指标。因此，这些检测器的优势在弱小目标检测上并没有凸显出来。对于 DSFD 和 SCRDet 没有在 TinyPerson 上取得较好的性能，通过分析认为这是由于两者网络结构针对各自目标任务而设计，并不适合于弱小人体目标检测任务。

■　尺度匹配算法内部的比较

基线方法设置为使用 ImageNet 分类模型作为检测器主干网的预训练模型，如表 3.7 所示，SM+算法在基线方法上提升明显，Faster RCNN-FPN-MSM+在 tiny 尺度上的性能均为最佳，在MR_{50}^{tiny1}、MR_{50}^{tiny2}、MR_{50}^{tiny3}和MR_{50}^{tiny}上分别取得了 86.47%、78.12%、75.83%和 85.60%。如表 3.8 所示，基线方法 Faster RCNN-FPN 在AP_{50}^{tiny}上取得了 47.35%的性能。相比于基线方法，SM 算法提升的效果较为明显，Faster RCNN-FPN-RSM 和 Faster RCNN-FPN-MSM 分别在AP_{50}^{tiny}上提高了 3.98 个点和 3.54 个点。在此基础上，SM+算法进一步提升了性能，在表中所有的评测指标均取得了最佳的性能，MSM+在AP_{50}^{tiny1}、AP_{50}^{tiny2}、AP_{50}^{tiny3}和AP_{50}^{tiny}上分别取得了 34.20%、57.60%、63.61%和 52.61%。Faster RCNN-FPN-RSM+和 Faster RCNN-FPN-MSM+分别在AP_{50}^{tiny}上比基线方法提升了 4.11 个点和 5.26 个点。这样性能提升的比较策略是包含引入额外大规模数据集而带来的增益的。由于 SM+算法更大程度符合算法的前提假设，接下来将主要以实例级别的尺度匹配算法 SM+为主要的分析对象。

以 AP 为例，从表 3.8 可以发现 MSM+算法对于检测器的提升要比 RSM+大很多。相比于 RSM，RSM+在AP_{50}^{tiny}上提升了 0.13 个点，性能提升较为有限。而在 MSM 方面，MSM+在AP_{50}^{tiny}上显著提升了 1.72 个点。从数据形式的角度出发，通过分析认为这是因为实例级别尺度变换的不确定性导致的。由于 RSM+在尺度分布上进行随机的采样，因此得到的理想尺度大小会发生较为剧烈的波动。比如，一个十分小的前景实例可能会采样到一个较大的理想尺度大小，因而放大的实例会在观感上十

分模糊；反之亦然，较大的前景实例可能会采样到较小的理想尺度，使得实例被很大程度上缩小。除此之外，同样一个实例在不同的训练循环中可以采样到不同的尺度大小，再次增加了尺度变换的不确定性。然而，MSM+相对来说通过单调匹配的机制，按照前景实例尺度从小到大的排序，依次单调进行匹配，而且每次训练循环过程中结果一致，能够很大程度上避免这种随机性。

从另一个角度出发，RSM+更像是一种数据增强的形式，通过尺度变换的不确定性来扩充样本以增加数据集的多样性，防止检测器的过拟合。在基于实例级别的尺度匹配中引入这样的不确定性并不合适，会放大数据观感上的不合理性。MSM+则像是一种以减少数据集间尺度分布差异性的训练策略，通过单调匹配的方式减少不适当情况的出现。接下来将从定性和定量分析两个角度更好地剖析 SM+算法。

定性分析：为了更好地展示所提出的尺度匹配算法，以 RSM 和 RSM+为例，通过绘制数据集匹配前后的尺度分布呈现了减少数据集间分布差异的有效性。如图 3.6 中左图所示，RSM 算法有效地迁移了预训练数据集 MS COCO 的尺度分布，使得尺度分布向着下游目标任务数据集 TinyPerson 靠拢。由于图像级别的尺度匹配直接对图片大小进行操作，从而在匹配过程中产生了近似误差，一定程度上阻碍了尺度分布的对齐效果，具体的差异可以如图 3.6 左图放大区域所示。RSM+算法的分布对齐效果如图 3.6 中右图所示，可以发现 MS COCO 在经过 RSM+算法的分布对齐后和下游目标数据集 TinyPerson 的尺度分布十分相似，更大程度上达成了尺度匹配的目标，在同样区域的放大部分可以看到 RSM+算法有效地填补了这些空白，进一步减少了数据集间的尺度分布差异。

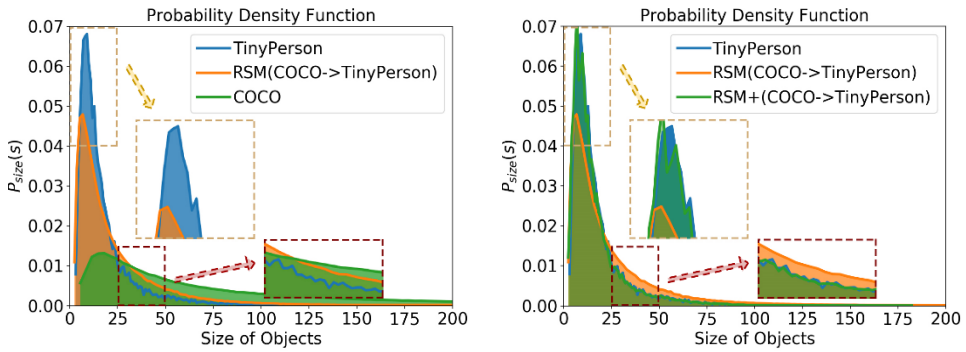


图 3.6　通过 RSM 和 RSM+进行尺度分布对齐的效果

Figure 3.6　The effect of scale distribution alignment by RSM and RSM+

定量分析：除尺度分布的可视化之外，引入了 JS 散度（Jensen-Shannon Divergence）[94]通过定量的方式来更好地度量两个尺度分布之间的相似性。这里，定义 *p(x)*和 *q(x)*均为一个离散随机变量 *x* 的概率密度分布，各自的和相加为 1，并且对于取值范围 *X* 内的任一 *x* 均有 *p(x)*>0 和 *q(x)*>0。为了更好地引入 JS 散度此之前需要先说明 KL 散度（Kullback-Leibler Divergence）[95]，其数学形式可以等价于一个交叉熵减去一个信息熵，如下公式 3.8 所示，

$$D_{KL}\big(p(x) \parallel q(x)\big) = \sum_{x \in X} p(x) ln \frac{p(x)}{q(x)}. \tag{3.8}$$

JS 散度解决了 KL 散度非对称的问题，如下公式所示，

$$D_{JS}\big(p(x) \parallel q(x)\big) = \sum_{x \in X} [\frac{1}{2} D_{KL}\left(p(x) \parallel \frac{p(x)+q(x)}{2}\right) + \frac{1}{2} D_{KL}\left(q(x) \parallel \frac{p(x)+q(x)}{2}\right)]. \tag{3.9}$$

根据 JS 散度的定义，可以得到进行对齐后的尺度分布和下游目标数据集的尺度分布之间的相似性，SM 和 SM+算法的对齐效果如表 3.9 所示。从表中可以看出 RSM 和 MSM 算法有效地对齐了两个数据集之间的分布，对齐过后各自和 TinyPerson 的 JS 散度为 0.0091 和 0.0133。同时，RSM+和 MSM+更加有效地减少了两个数据集之间的尺度分布差异，分别在 JS 散度上取得了 0.0020 和 0.0013，并将数据集间尺度分布相似性提高了 78.02%和 90.23%。

表 3.9　不同尺度分布对齐方法的效果

Table 3.9　The effect of different scale distributions alignment methods

T	$D_{JS}\left(P_{size}(s;T(E)) \parallel P_{size}(s;D)\right)$
RSM	0.0091
RSM+	0.0020
MSM	0.0133
MSM+	0.0013

3.4.3　消融实验

为了分析基于精细尺度匹配的弱小目标检测算法背后的原理，进行了大量的消融实验。接下来将从检测器、加载权重、概率阈值三方面进行算法特性层面的分析，

同时对关键策略 PSI 能提升性能的缘由进行了剖析。

表 3.10 Faster RCNN-FPN 使用不同预训练数据集的性能比较

Table 3.10 Comparisons of different pre-training dataset on Faster RCNN-FPN

预训练数据集	$MR_{50}^{tiny}(\downarrow)$	$AP_{50}^{tiny}(\uparrow)$
ImageNet	87.57	47.
COCO800	86.85	49.76
RSM(COCO)	86.22	51.33
RSM+(COCO)	86.26	51.46
MSM(COCO)	85.86	50.89
MSM+(COCO)	**85.60**	**52.61**

表 3.11 RetinaNet*使用不同预训练数据集的性能比较

Table 3.11 Comparisons of different pre-training dataset on RetinaNet*

预训练数据集	$MR_{50}^{tiny}(\downarrow)$	$AP_{50}^{tiny}(\uparrow)$
ImageNet	88.31	46.56
COCO800	89.42	45.03
RSM(COCO)	88.87	48.48
RSM+(COCO)	87.64	50.59
MSM(COCO)	88.39	49.59
MSM+(COCO)	**87.09**	**51.25**

表 3.12 Faster RCNN-FPN-MSM+加载不同预训练模型的性能比较

Table 3.12 Comparisons of loading different model weights on Faster RCNN-FPN-MSM+

模型加载权重	$MR_{50}^{tiny}(\downarrow)$	$AP_{50}^{tiny}(\uparrow)$
RPN-none	**85.58**	52.41
RPN-cls	85.60	**52.61**
RPN-reg	85.83	52.24
RPN-all	85.94	51.97

表 3.13 Faster RCNN-FPN-MSM+使用不同的概率 p 的性能比较

Table 3.13 Comparisons of different probability p on Faster RCNN-FPN-MSM+

概率 p	0	0.2	0.4	0.6	0.8	1
$MR_{50}^{tiny}(\downarrow)$	85.75	85.70	**85.60**	85.88	85.97	86.48
$AP_{50}^{tiny}(\uparrow)$	51.42	51.85	**52.61**	51.53	51.25	50.69

■ 算法特性

检测器：如表 3.10 和表 3.11 所示，为了验证 SM 和 SM+算法的有效性，除了上一节二步法检测器选取 Faster RCNN-FPN 作为基线方法，一步法检测器中也选取了 RetinaNet 作为基线方法，可以发现算法在 Faster RCNN-FPN 和 RetinaNet 取得了一致性的提升，证明了算法的提升是和检测器无关的。需要注意的是 COCO 和

COCO800 是不同的概念，COCO 表示使用 MS COCO 数据集原图大小进行预训练，预训练和微调阶段使用同一套锚点框设置。COCO800 表示在预训练时会跟据图片大小将图片的长短边控制在（800，1333）的范围内，这也是通用目标检测器在 MS COCO 上进行训练的标准配置，此时预训练和微调阶段使用两套锚点框设置。

以 COCO800 为基线方法，以评测指标AP_{50}^{tiny}为例，图像级别的尺度匹配 RSM 和 MSM 相比基线方法分别在 Faster RCNN-FPN 上取得了 1.57%和 1.13%，在 RetinaNet*上各自取得了 3.45%和 4.56%。实例级别的尺度匹配性能提升更为明显，RSM+和 MSM+分别在 Faster RCNN-FPN 上取得了 1.7%和 2.85%，在 RetinaNet*上各自取得了 5.56%和 6.22%。在 RetinaNet*上的提升要远远超过 Faster RCNN-FPN，原因主要是因为使用 COCO800 作为预训练数据集反而会带来负面的影响。

加载权重：如表 3.12 所示，采用 Faster RCNN-FPN-MSM+作为基线，对检测器模型中可以拆解的 RPN 分类和回归模块进一步分析。使用在微调阶段加载的模块作为命名，以 RPN-cls 和 RPN-all 为例，RPN-cls 表示在 RPN 中只加载了分类模块的预训练参数，RPN-all 表示在 RPN 中分类和回归模块的预训练参数均被加载。

从表中可以发现，以 AP 为例，RPN-cls 在四种加载权重方式中取得了最好的性能，在AP_{50}^{tiny}上为 52.61%，然而 RPN-all 的实验性能严重下降到 51.97%。由于 SM+算法会调整样本图片上的每个实例的尺度大小，在这种情况下，经调整的预训练样本图片中的目标形式将会更接近于下游目标任务数据集中的目标形式。因而，在预训练阶段所学习到的知识，对检测器在微调阶段区分前景和背景具有一定的参考意义。因此在分类方面，与 SM 算法相比，SM+算法能够提供更好的模型初始化，所以加载 RPN 的分类模块有利于性能的提升。然而，值得注意的是，尺度迁移后样本的图像结构会受到一定程度的破坏，如图 3.5 所示。类似地，RPN-none 比 RPN-reg 的性能高是因为背景的形式不同。经 PSI 策略修补的背景会带有结构信息损失或者背景语义歧义，这种现象在下游任务数据集上几乎不会出现，因此检测器在此类预训练样本上学习的回归参数并不具备参考意义。同时，由于传统的 inpainting 策略修补形式单一，容易引起神经网络的过拟合。因此，加载回归模块权重的性能比重新训练 RPN 模块都要低。同时全部加载 RPN 模块权重会加剧分类和回归的歧义性，从而引发性能的剧烈下降，使得AP_{50}^{tiny}降为 51.97%。

概率阈值：当概率设置为 0 的时候，PSI 策略退化为全部使用替换背景的背景修补策略，由于上下文背景信息可能不同而带来语义上的歧义性；当概率设置为 1 时，PSI 策略退化为传统的 inpainting 策略，无法有效地修补图片结构上的破坏。单纯分开使用这两种极端的背景修补模式均会带来性能的衰退，如表 3.13 所示，相比于最佳的性能 52.61%，分别会降低 1.19%和 1.92%。然而，综合利用两种背景修补策略的优势，一个合适的概率阈值（p=0.4）可以很好地平衡背景的语义歧义性和图片的结构损失。

■ 关键策略

通过实验发现，仅仅在实例级别对齐预训练数据集和下游目标任务数据集的尺度分布，并不能够大幅地提升弱小人体目标检测性能。因为 SM+算法会显著地缩小前景实例，从而破坏原来样本图片的结构信息。采用传统的 inpainting 策略会导致一些伪影的出现，无法有效地进行样本背景修补。PSI 策略的引入至关重要，综合借鉴了两种背景修补策略各自的优势，和实例级别的尺度对齐相结合，形成的 SM+算法有效地提升了检测器的性能。为了验证 PSI 策略的有效性，消融实验如表 3.14 所示，可以发现没有 PSI 策略的辅助，在AP_{50}^{tiny}上 RSM+和 MSM+分别会显著降低 1.34%、1.92%，甚至性能比图像级别尺度匹配算法 SM 还低。通过分析认为这是由于不符合实际样本的图像结构和不合理的伪影模式使得神经网络过拟合，从而影响了实例级别的样本缩放而带来的优势，导致了不理想的结果。

相比于传统 inpainting 策略单一的修复模式，PSI 策略中的背景替换更加多样化，因而 PSI 策略也可以被认为是背景相关的数据增强策略。为了更好地探究 PSI 策略的性能提升是来源于数据增强还是更好地修复了背景，进一步设计了消融实验以验证 PSI 策略提升性能的根本原因。如表 3.15 所示，为了探究背景选择策略带来的影响，引入了 CP（COCO）和 CP+（COCO）作为预训练数据集的实验。CP 和 CP+表示根据概率选择直接复制并粘贴前景实例到新的背景图像上，或者使用原图的背景选择策略，其中 CP 和 CP+变换的区别在于，在 CP+变换中新背景图片上面的实例被判定为前景，因此这些前景实例的标注会参与预训练，而 CP 变换中这些实例会被判定为背景，从而不参与预训练。COCO 作为预训练数据集表示直接使用 MS COCO 原图进行预训练。

如表 3.15 所示，以AP_{50}^{tiny}为例，CP（COCO）和 CP+（COCO）分别为 50.66%和 50.46%，性能较为接近。相比于 COCO 原图作为预训练数据集的基线方法，CP（COCO）和 CP+（COCO）带来的性能提升较为有限，分别提升了 0.70%和 0.50%，说明背景替换策略只能带来有限的性能提升。如表 3.14 所示，没有 PSI 策略的 MSM+（COCO）仅仅获得了 50.69%的性能，MSM+（COCO）结合了背景替换策略 CP 和实例级别的前景目标缩放 MSM+，相比基线方法显著提升了 2.65%，证明了 PSI 策略带来提升的原因不在于数据增强，而在于和实例级别的目标尺度匹配进行了深度地结合，在实现更加精细的尺度分布对齐外，更好解决了由实例级别的尺度匹配引发的背景问题，通过概率更好地平衡了图像结构损失和背景歧义现象。

表 3.14　Faster RCNN-FPN 使用不同预训练策略的性能比较

Table 3.14　Comparisons of different pre-training strategies on Faster RCNN-FPN

预训练策略	$MR_{50}^{tiny}(\downarrow)$	$AP_{50}^{tiny}(\uparrow)$
RSM+(w/o PSI)	86.39	50.12
RSM+	86.26	51.46
MSM+(w/o PSI)	86.48	50.69
MSM+	**85.60**	**52.61**

表 3.15　Faster RCNN-FPN 上性能增益的消融实验

Table 3.15　The ablation study of performance improvement on Faster RCNN-FPN

预训练数据集	$MR_{50}^{tiny}(\downarrow)$	$AP_{50}^{tiny}(\uparrow)$
COCO	86.96	49.96
CP(COCO)	87.00	50.66
CP+(COCO)	86.47	50.46
MSM(COCO)	85.86	50.89
MSM+(COCO)	**85.60**	**52.61**

3.5　本章小结

本章从各个维度介绍了面向海上快速救援的弱小人体目标检测数据集 TinyPerson 及相应的评测指标。综合分析当前预训练策略的研究进展，结合目前 TinyPerson 数据集体量的现象，提出了基于精细尺度匹配的弱小人体目标检测预训练策略，从图像级别和实例级别两个思路出发，驱使预训练数据集的尺度分布向下游目标任务数据集的尺度分布迁移，有效地减少了尺度分布间的差异性。最后本章在 TinyPerson 上对提出的算法进行了验证，结果显示本章提出的 SM 和 SM+算法有效

提升了检测器的性能，并且 PSI 策略成功缓解了 SM+算法引发的背景歧义问题，和实例级别目标缩放相辅相成，促进了检测器预训练阶段的学习。

第4章　基于多模态互感的无人机跟踪

无人机的管控需求日益增加，而目前学术界在无人机跟踪领域缺少公开的高质量基准数据集、基线方法和评测体系。面对亟待解决反无人机感知的需求，本章首先介绍了面向无人机管控的无人机目标跟踪数据集 Anti-UAV 的详细信息、评测体系和指标等内容。考虑到数据集内只有无人机这一通用类别，对基于图像信息交互的训练策略进行了剖析，提出了基于双流语义一致性的训练策略，针对 Anti-UAV 搭建了大量跟踪器的基线方法，并在上面对于提出的算法进行验证和分析。

4.1　无人机跟踪数据集 Anti-UAV

4.1.1　数据集介绍

- 数据收集

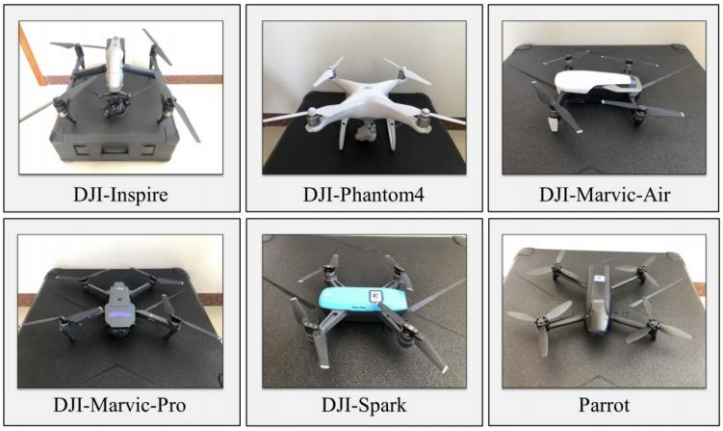


图 4.1　拍摄多模态跟踪数据的无人机总览

Figure 4.1　Overview of UAVs for capturing multi-modal tracking data

为了构建复杂环境下的无人机目标跟踪数据集，从数据模态、场景和无人机类型三个角度着手以保证数据集的多样性。为了集成不同光学信息条件下数据的互补的优势，采用光电跟踪设备进行 RGB 和 TIR 视频的录制。在场景方面，选取了白天和夜晚两种光照条件，以及楼宇、天空、树木等多种场景，充分保证了数据集的场景多样性。在实际拍摄过程中，一共采用了六种不同大小的无人机类型进行拍摄

以保证样本的多样性，如图 4.1 所示。图中展示的每种无人机均被进行了数据采集，每段视频以可见光和红外两种模式存储，存储的数据格式为 mp4 文件。

为保证数据集场景的多样性，采用了多种采集策略以增加数据集的跟踪难度，加大对于跟踪器的挑战性。同时，在采集的过程中遇到了诸多困难。

采集策略：无人机在飞行中采用高速飞行、低速飞行和悬停三种飞行速度；制造消失视野、遮挡等现象，无人机大部分时间在跟踪视野内，少数时间处于遮挡状态，遮挡状态又分为完全遮挡和部分遮挡；收集鸟类等干扰物存在的场景；保障场景多样性，在高低不同的楼宇、塔吊、树木、天空、云层、雾霾等各种复杂背景中试飞；为增加数据集的跟踪难度，无人机在飞行时忽远忽近，覆盖不同的尺度；采用直线、折线和曲线等多种飞行方式、多种飞行高度、多种飞行距离等策略，加大了对于跟踪器的挑战性。

采集难点：无人机在飞行过程中，风速会对无人机的操控有一定的影响；光电跟踪设备的位置固定，使得无人机的一次拍摄中无法进行超远距离的移动；数据采集员和无人机飞手不默契的配合会导致有时候在采集视野中丢失无人机目标；无人机的续航能力有限，短暂飞行后需要补充电源，一定程度上降低了采集效率。

■ 数据处理

为了保证构建一个高质量的数据集，在数据处理过程中采取了一种渐进式的策略，分为粗标注、精标注、检查调整三个阶段。

粗标注：在原始视频数据上，首先对每段视频进行序列级别的属性和场景标注，例如，无人机目标体型（大、中、小）、光照条件（白天、黑夜）、数据光学模态（红外、可见光）、场景标签（有无云雾遮挡、有无楼宇遮挡、有无树林遮挡等）和干扰物标签（有无鸟类、有无民航飞行等）。然后对每段视频每隔 25 帧进行帧级别的粗略标注，在标注过程中需要注明无人机目标的状态和位置信息，若无人机在视野内，使用紧致的包围框将其框出。

精标注：对粗标注后的数据进行筛选，根据视频序列的场景标注，选出视频序列中每个场景复杂度较高的视频进行标注。然后根据粗标注的结果分别对每段视频每帧进行精细标注，标注规则和粗标注相同。

检查调整：精标注之后，可能仍然存在不合理的标注现象：目标包围框不够紧

致，没有无人机目标或者无人机被大面积遮挡也被认定为存在在视野内；数据采集中由于转台移动速度过快导致无人机目标运动模糊等等。首先需要进行进一步的检查和调整，移除以上所述的不合理的标注现象。为了更好地验证跟踪器的泛化能力，将所有视频序列分为两部分，一部分构成训练集和验证集，另一部分构成测试集。然后，对视频序列进行拆分，每 1000 帧拆分为一个子视频序列。最后，Anti-UAV 中一共有 318 个视频序列对（包含一个 RGB 视频序列和一个 TIR 视频序列），其中 160 个视频序列构成训练集，91 个视频序列构成测试集。

■　数据分布

为了更好地了解 Anti-UAV 数据集的特性，以 TIR 视频序列为例，从位置信息和尺度信息两个方面对数据集进行了统计。

位置分布：利用标注中所有包围框的位置进行统计，如图 4.2 所示，训练集、验证集和测试集中无人机的包围框都集中在图像的中央部分。相比于垂直方向上的波动，三个子集的位置分布在水平方向上波动更大。

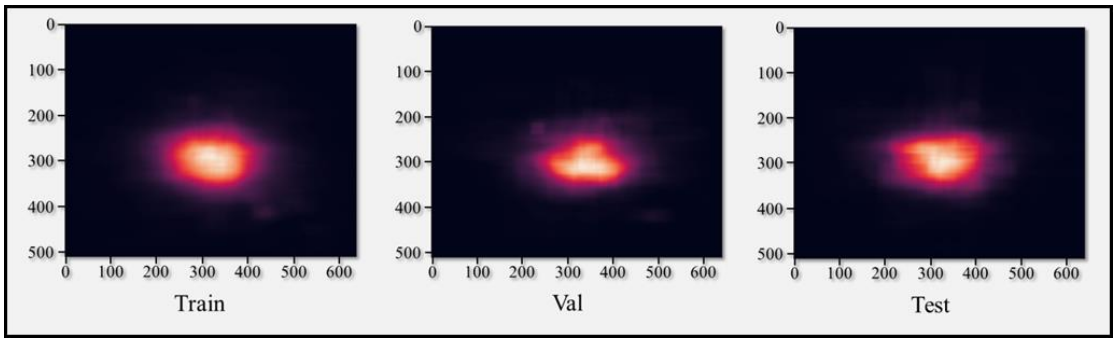


图 4.2　Anti-UAV 数据集的位置分布

Figure 4.2　The position distribution of Anti-UAV

尺度分布：目标的尺度的定义和弱小目标检测中的定义相同，为根号下包围框的面积，因此根据统计信息可以得到三个子集的尺度分布，如图 4.3 所示。可以看出，训练集、验证集和测试集三者的尺度分布较为相似，平均尺度大小均在 40 个像素以下，其中测试集的尺度分布更为尖锐。

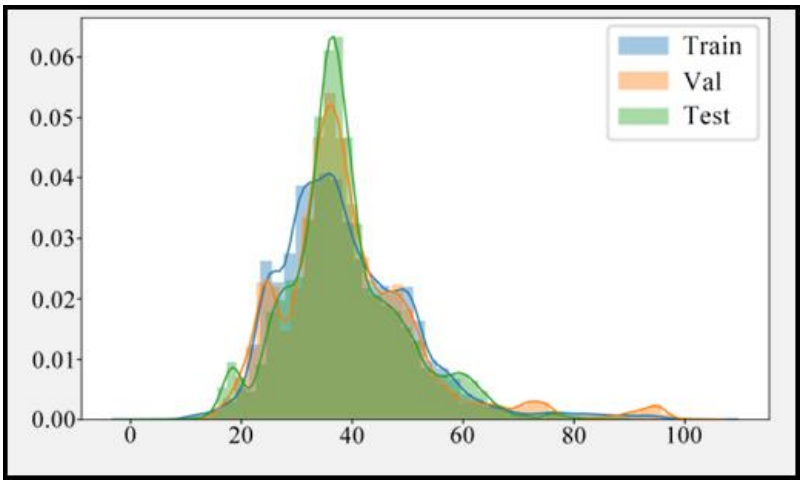


图 4.3 Anti-UAV 数据集的尺度分布

Figure 4.3 The scale distribution of Anti-UAV

■ 属性标注

为了更加全面地评估跟踪器的性能，将跟踪过程中可能遇到的难点问题作为属性，对整个数据集进行了序列级别的归类。总共分为消失视野（OV）、遮挡（OC）、快速移动（FM）、尺度变化（SV）、弱光条件（LI）、红外交迭（TC）和低分辨率（LR）七个属性，定义如表 4.1 所示。为了更好地表达每个属性的含义，从数据集中特地挑选了每个属性对应的场景，如图 4.4 所示。作为多模态光学反无人目标跟踪数据集，从 RGB 视频序列和 TIR 视频序列中均进行了属性的挑选。各种场景的属性标注有助于辅助科研工作者了解每个跟踪器的优势和劣势，从而推动反无人机感知技术的发展。

除此之外，针对属性标注同样进行了统计分析，如图 4.5 所示。根据统计信息可以得到如下结论：相比于其他属性，OV 中测试集占据了更大的比例；遮挡出现的频次最低；TC、FM 和 LI 是整个数据集中出现较多的属性，其中 TC 的频次最高等等。由于视频序列中 TC 属性出现的频次过高，根据跟踪器 SiamRPN++LT[61]在 TIR 视频序列上的性能，将其进一步划分为TC_{easy}、TC_{med}和TC_{hard}，统计结果如图 4.6 所示。

表 4.1　Anti-UAV 属性标注的含义

Table 4.1　Illustration of attribute annotation in Anti-UAV

属性	属性描述
OV	Out-of-View: the UAV leaves the view.
OC	Occlusion: the UAV is partially or heavily occluded.
FM	Fast Motion: the ground-truth's motion between two adjacent frames is larger than 60 pixels.
SV	Scale Variation: the ratio of the bounding boxes of the first frame and the current frame is out of the range [0.66, 1.5].
LI	Low Illumination: the illumination in the target region is low.
TC	Thermal Crossover: the UAV has a similar temperature with other objects or background surroundings.
LR	Low Resolution: the number of pixels inside the bounding box is below 400 pixels.

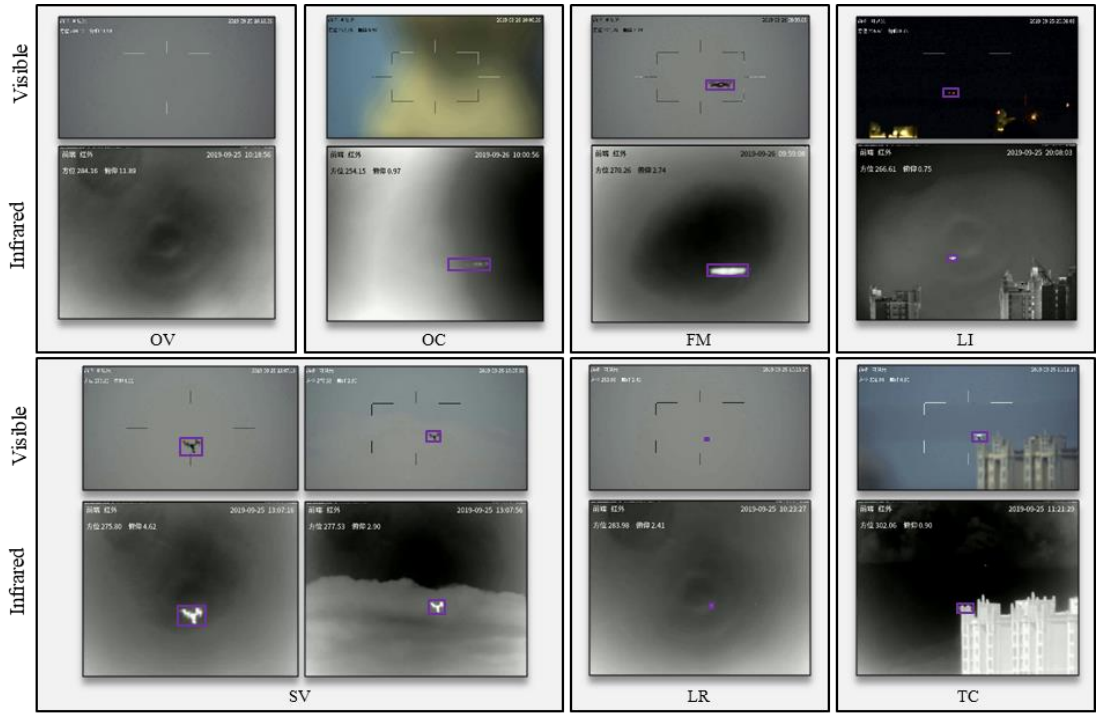


图 4.4　多模态数据集 Anti-UAV 的视频截图

Figure 4.4　Screenshots taken from Anti-UAV multi-modal dataset

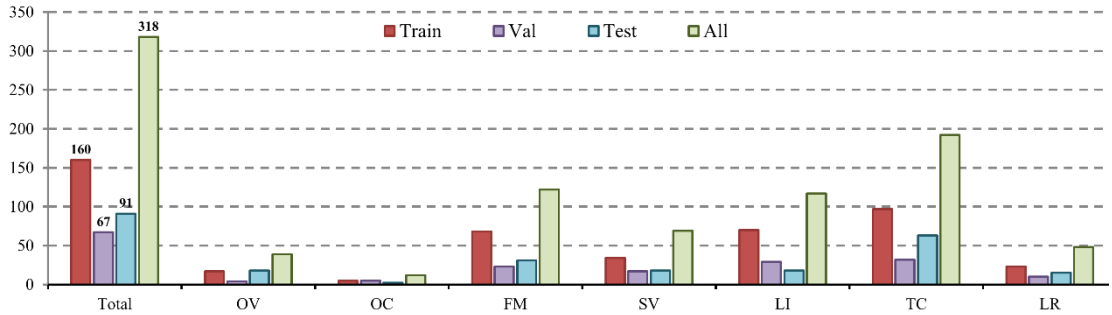


图 4.5 Anti-UAV 的属性标注分布

Figure 4.5 The distribution of attribute annotation in Anti-UAV

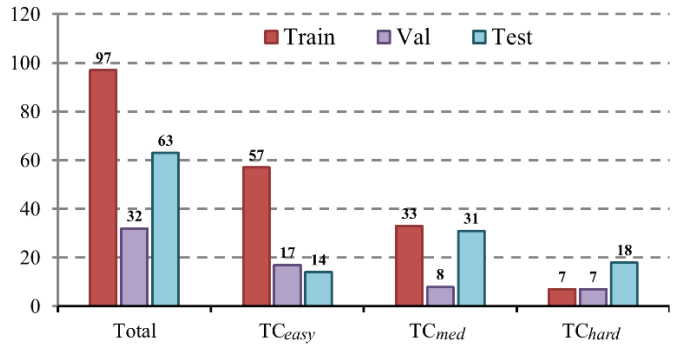


图 4.6 TC 属性的进一步划分

Figure 4.6 Further division of attribute TC

■ 评测规则

为了适应于不同的评测条件，针对反无人机感知的特性，一共制定了如下的三种评测的规则：

1）Protocol I 使用 Anti-UAV 单模态光学视频进行无人机跟踪。不使用包含无人机通用类别的数据进行训练，然后通过 RGB 或者 TIR 视频序列分别单独进行无人机跟踪，然后对跟踪器的性能进行评价。因为在训练期间跟踪器见过相关类别的条件下，普遍会在推理期间对这些类别取得更好的性能，所以该 Protocol 的目的是验证在训练过程不使用含无人机通用类别的数据的情况下，跟踪器对无人机跟踪的性能表现。

2）Protocol II 旨在提供一种独特的无人机跟踪单模态评估准则。允许跟踪器使用 Anti-UAV 中的 RGB 或 TIR 视频序列参与训练，具体的使用方式可以是

用来微调或者从头训练跟踪器。训练数据中可以包括其他包含无人机目标的数据，但是会引起由无人机相关的数据量不同带来的不公平比较，在这里需要建立公平的比较体系。

3）Protocol III 鼓励研究人员探索如何充分利用 Anti-UAV 中非配准的 RGB 和 TIR 进行融合跟踪，这也是未来的工作重点之一。

4.1.2　评测指标

实验中所有跟踪器都使用成功率（Success Rate）[27]、精确率（Precision Rate）[27]和状态准确率（State Accuracy，SA）评估。在此引入平均像素误差（Average Pixel Error，APE）和平均重叠率（Average Overlap Rate，AOR）。

跟踪过程中，当前帧中存在无人机目标，APE 会根据预测的无人机包围框和真值包围框两者的中心点的距离进行累积，最后依据总有效的帧数进行平均；AOR 根据两者的包围框的重叠程度（可以参照目标检测的 IOU 定义）进行累积，最后同样求平均作为结果。然而，单单使用 APE 和 APR 作为评测指标在跟踪器丢失目标的情况下，无法准确地对跟踪器的性能进行评估，从而产生了基于一次通过评测（One Pass Evaluation，OPE）的成功率图（Success Plot）和精确率图（Precision Plot）。

成功率：以横坐标为给定 IOU 阈值范围、纵坐标为帧数占比构成成功率图，成功率图的含义为跟踪器预测的包围框和真值包围框在给定的 IOU 阈值内的帧数占比。成功率即为成功率图的曲线下面积。

精确率：以横坐标为给定距离阈值范围、纵坐标为帧数占比构成精确率图，精确率图的含义为跟踪器预测的包围框和真值包围框的中心点在给定的像素值内的帧数占比。精确率即为精确率图的曲线下面积。

状态准确率：由于无人机的移动速度较快和复杂的背景环境，需要一个和无人机状态有关的新的评价指标，因此，引入了状态准确率这个评价指标：

$$SA = \sum_{t} \frac{IOU_t \times \delta(v_t > 0) + p_t \times \left(1 - \delta(v_t > 0)\right)}{T}. \tag{4.1}$$

其中，T 为视频序列的总帧数，IOU_t 是第 t 帧上跟踪器预测的包围框与相应的真值包围框的重叠程度，v_t 为第 t 帧是否存在无人机目标，p_t 为跟踪器对于当前帧无人机状态的判断。所有序列 SA 的平均值 mSA 为评测结果。

4.2 基线实验方法

4.2.1 实验配置

基于 Protocol I 的测试过程中，直接使用跟踪器原有的配置，并没有在 Anti-UAV 上进行微调等操作。实验选取的跟踪器主要分为以下两类：

1）基于深度学习的跟踪器：SiamFC、SiamRPN、SiamRPN++、SiamRPN++LT[63]、SiamMask[96]、SiamDW[97]、SiamBAN[98]、SiamFCOS[33]、SiamCAR[99]、SiamRCNN[100]、SPM-AlexNet[101]、SPM-Res18[101]、ATOM[102]、Dimp[103]、PrDimp[104]、Super-Dimp[1]、ATOM-MU[105]、Ocean-Online[106]、Ocean-offline[106]、MDNet、RT-MDNet、ECO[107]、KYS[108]、SPLT[109]、LTDSE[33]、GlobalTrack[110]。

2）基于相关滤波的跟踪器：MOSSE[111]、DAT[112]、CSK[113]、Staple[114]、Staple-CA[115]、MCCTH[116]、DCF[117]、KCF[117]、CN[118]、STRCF[119]、LDES[120]、DSST[121]、CSRDCF[122]、BACF[123]、MKCFup[124]、ECO-HC[107]。

4.2.2 实验结果及分析

上述跟踪器在 Anti-UAV 验证集上的性能如表 4.2 所示，测试集上的性能如表 4.3 所示。测试集中跟踪器依据 TIR 视频序列上的性能进行排序，验证集中跟踪器的顺序和测试集保持一致。由于跟踪器数量较多，序列中分别用红色、蓝色、绿色来表示在特定条件下跟踪器排名的第一名、第二名、第三名。接下来将从总体性能和属性性能两个角度对这些基线跟踪器进行深入分析。

总体性能：如表 4.2 和表 4.3 所示，SiamRCNN 在验证集和测试集上均取得了最佳的性能，TIR 序列的状态准确率$\mathrm{mSA_{TIR}}$分别为 74.33%、65.41%，RGB 序列的状态准确率$\mathrm{mSA_{RGB}}$分别为 74.32%和 70.83%。GlobalTrack 在验证集和测试集上的表现同样稳定，均取得了排名第二的表现，$\mathrm{mSA_{TIR}}$分别为 72.00%和 63.86%，mSA_{RGB}分别为 67.28%和 66.24%。然而，位列第三名的跟踪器发生了波动，验证集上 SiamRPN++LT 在 RGB 序列和 TIR 序列上一致取得了第三名的成绩。测试集上，Super-Dimp 在 TIR 序列上位列第三名，LTDSE 在 RGB 序列上排名第三。 通过分

[1] Super-Dimp 集成了 PrDimp 和 Dimp 各自的创新点，代码网址：https://github.com/visionml/pytracking.

析可以发现，普遍基于长时跟踪的跟踪器由于考虑到目标会消失视野、完全遮挡等现象而设计多种重跟踪机制，因而会取得较高的性能。同时，基于深度学习的跟踪器普遍会在排行榜上位列较前。

表 4.2　基于 Protocol I 的 Anti-UAV 验证集上跟踪器的性能 mSA（%）

Table 4.2　The performance mSA(%) of tracker on Anti-UAV validation set under Protocol I

跟踪器	红外											可见光
	OV	OC	FM	SV	LI	TC				LR	All	All
						TC_{easy}	TC_{med}	TC_{hard}	TC_{all}			
MOSSE	23.29	4.63	5.26	8.08	28.64	48.70	27.19	8.17	34.45	6.20	29.40	21.06
DAT	15.98	6.77	8.30	9.63	30.76	36.06	20.62	1.49	24.64	3.00	29.43	38.27
CSK	21.33	8.90	6.02	8.06	40.32	53.32	44.70	7.51	41.14	1.96	37.67	35.89
Staple-CA	21.33	10.65	7.09	14.09	36.96	55.63	46.81	14.54	44.43	10.48	37.51	37.23
MCCTH	26.51	20.61	8.65	17.80	39.75	54.28	42.12	8.81	41.29	10.29	38.10	39.49
Staple	33.92	13.82	7.54	17.64	36.20	54.80	40.54	13.01	42.09	9.81	37.62	38.15
CN	24.74	10.12	6.93	14.74	40.78	58.34	46.32	11.22	45.03	10.25	39.82	36.04
DCF	21.15	6.93	6.95	13.32	37.57	53.64	42.68	9.02	41.14	11.44	36.65	36.48
KCF	22.04	8.85	7.50	14.21	37.91	53.41	42.94	9.67	41.23	11.92	36.82	38.11
STRCF	39.25	25.34	21.56	27.04	48.48	47.92	41.02	4.47	36.69	18.34	41.65	44.92
LDES	21.76	8.91	19.78	13.55	46.63	54.01	42.90	5.05	40.52	8.92	41.41	48.98
DSST	23.66	10.46	7.05	14.03	43.25	55.31	45.76	13.55	43.79	9.29	40.54	37.48
CSRDCF	40.46	21.13	25.45	27.84	50.87	58.66	52.17	14.53	47.39	24.77	47.73	41.54
BACF	19.51	9.00	18.21	20.53	44.72	56.47	44.43	7.06	42.65	21.99	43.16	43.87
SiamFC	42.10	27.99	30.28	22.83	53.41	66.36	45.80	10.45	48.99	16.59	49.34	44.08
Ocean-Online	16.21	21.02	19.65	24.11	45.77	52.66	39.21	6.09	39.11	20.33	41.56	46.45
MKCFup	44.60	20.13	8.64	22.57	40.81	56.95	43.91	13.77	44.24	10.25	41.31	40.21
SiamMask	40.77	27.28	24.22	24.16	47.10	54.82	38.76	12.50	41.54	14.62	44.34	44.26
SiamDW	43.93	39.66	29.40	35.14	56.84	59.89	42.31	8.29	44.21	27.87	49.46	44.90
SiamBAN	20.02	31.73	19.03	24.84	48.75	57.63	49.44	6.44	44.39	13.38	43.60	39.90
RT-MDNet	43.88	20.82	18.60	27.35	46.97	64.97	44.42	13.15	48.50	13.52	45.99	44.93
SPM-AlexNet	39.84	24.75	30.35	27.70	51.15	56.73	34.98	8.20	40.68	14.41	46.71	46.51
ECO-HC	23.16	14.96	25.38	28.66	52.66	60.61	48.51	16.63	47.96	30.29	49.26	43.67
Ocean-Offline	26.63	42.86	32.09	24.75	55.05	59.61	41.66	20.76	46.62	18.60	48.74	45.74
MDNet	45.88	24.16	23.47	31.13	50.42	65.96	48.59	14.84	50.43	24.85	49.49	45.17
SiamRPN++	35.32	41.11	28.17	28.39	55.85	57.95	44.55	6.94	43.44	19.71	48.60	46.12
SiamRPN	32.07	31.37	25.73	30.56	52.57	65.60	45.24	10.83	48.53	21.97	48.16	46.63
SPM-Res18	39.75	30.78	32.25	29.56	54.90	59.16	42.55	12.45	44.79	17.95	49.56	46.09
SiamFCOS	41.97	42.43	29.01	34.66	53.09	58.24	37.02	10.07	42.40	21.94	47.71	44.28
ECO	31.87	32.03	38.43	38.47	55.77	69.24	45.90	21.67	53.00	34.25	54.44	46.31
SiamCAR	25.05	46.12	40.84	29.96	63.19	68.27	55.99	17.55	54.11	21.67	56.70	46.52
KYS	36.07	55.52	57.64	52.68	64.70	66.30	35.35	32.05	51.07	48.00	60.50	59.79
ATOM	63.01	55.82	56.04	46.91	65.05	65.36	50.03	25.73	52.86	38.53	60.87	58.79
Dimp	41.96	59.85	59.95	55.78	66.33	70.19	51.62	30.16	56.79	47.52	63.51	61.54
ATOM-MU	42.73	55.51	58.74	47.16	64.83	68.58	46.61	26.26	53.83	39.18	61.27	60.45
SiamRPN++LT	50.02	**68.16**	63.39	54.06	**70.25**	**76.71**	**61.76**	19.25	**60.40**	42.66	**65.84**	**67.15**
SPLT	33.39	58.42	55.22	42.49	63.18	73.09	58.73	19.30	57.73	45.46	60.73	57.32
PrDimp	**65.22**	63.89	62.85	55.59	66.95	69.47	55.79	29.04	57.21	51.12	64.54	62.95
LTDSE	64.39	50.17	59.04	48.67	62.69	67.18	55.89	27.43	55.66	42.54	61.27	66.64
Super-Dimp	52.35	68.07	**65.30**	**64.80**	67.93	68.87	59.85	**34.22**	59.03	**61.45**	65.76	63.05
GlobalTrack	**69.21**	**78.62**	**73.35**	**66.11**	**76.33**	**76.47**	**63.08**	**43.45**	**65.90**	**60.26**	**72.00**	**67.28**
SiamRCNN	**73.46**	**78.24**	**73.98**	**67.97**	**76.19**	**78.21**	**69.55**	**55.48**	**71.07**	**67.93**	**74.33**	**74.32**

表 4.3 基于 Protocol I 的 Anti-UAV 测试集上跟踪器的性能 mSA（%）

Table 4.3 The performance mSA(%) of tracker on Anti-UAV test set under Protocol I

跟踪器	红外											可见光
	OV	OC	FM	SV	LI	TC				LR	All	All
						TC_{easy}	TC_{med}	TC_{hard}	TC_{all}			
MOSSE	8.89	24.16	6.02	4.06	3.56	15.23	10.34	5.80	10.13	3.80	13.47	15.23
DAT	8.01	21.94	5.33	13.53	3.11	40.57	13.28	6.50	17.41	3.76	22.68	27.19
CSK	11.51	26.97	9.56	12.35	2.71	46.51	15.63	5.30	19.54	3.29	24.26	28.38
Staple-CA	15.60	41.11	13.29	9.03	3.64	46.25	18.27	7.38	21.37	5.53	25.44	31.40
MCCTH	11.58	33.21	9.84	9.08	4.95	41.09	20.33	6.61	21.02	5.13	25.85	29.96
Staple	14.74	44.09	11.56	11.67	3.70	44.56	21.51	6.82	22.44	5.26	26.50	29.71
CN	14.41	39.75	10.66	15.18	3.54	59.75	28.02	7.93	29.33	4.81	31.72	28.65
DCF	14.85	36.89	12.28	11.55	3.18	60.32	30.26	8.80	30.80	4.25	32.55	38.39
KCF	16.14	37.56	12.60	11.66	3.55	60.23	30.80	9.17	31.16	4.36	32.88	39.40
STRCF	15.72	44.39	14.69	20.18	7.31	59.49	28.26	10.89	30.23	7.84	33.77	45.19
LDES	16.83	40.97	17.86	16.33	8.40	60.88	28.07	10.46	30.33	7.54	34.46	49.13
DSST	14.45	41.19	12.59	15.88	3.57	69.31	30.86	9.37	33.27	4.85	35.18	35.67
CSRDCF	13.70	46.26	12.96	19.14	4.97	61.55	32.05	10.22	32.37	7.26	35.29	47.10
BACF	16.17	41.91	15.66	16.70	4.13	70.16	33.02	8.53	34.28	4.91	36.78	47.52
SiamFC	18.59	60.83	21.46	23.58	13.55	63.82	29.00	11.02	31.60	10.36	36.97	45.69
Ocean-Online	17.98	41.56	14.72	17.35	3.73	68.51	32.02	9.27	33.63	4.45	37.22	48.11
MKCFup	16.44	43.35	15.60	14.15	3.48	70.74	35.55	8.92	35.76	4.88	37.41	39.52
SiamMask	29.09	53.55	18.73	22.03	8.14	59.32	30.42	17.91	33.27	10.08	37.44	45.92
SiamDW	19.36	38.14	17.65	22.18	8.52	57.68	36.28	13.73	34.60	9.44	38.01	49.86
SiamBAN	14.92	33.72	16.42	18.84	4.78	72.67	39.33	16.90	40.33	6.15	40.86	44.71
RT-MDNet	19.66	50.00	20.88	21.38	12.42	65.38	37.37	16.21	37.55	7.98	41.05	42.59
SPM-AlexNet	28.82	54.65	22.99	21.89	10.96	72.10	35.75	16.00	38.19	10.86	41.33	54.09
ECO-HC	20.48	50.46	20.72	24.69	6.74	**77.18**	41.74	14.76	41.91	10.21	42.39	48.91
Ocean-Offline	25.11	53.63	23.74	21.07	12.22	71.65	40.67	9.26	38.58	10.74	42.51	47.45
MDNet	28.90	73.29	24.19	20.60	12.95	65.92	42.13	15.44	39.79	12.14	42.95	43.94
SiamRPN++	22.96	48.88	20.44	21.27	9.63	75.16	40.24	16.46	41.21	10.76	43.01	51.69
SiamRPN	25.35	46.57	24.50	21.90	14.92	74.95	39.37	11.54	39.32	11.18	43.39	47.96
SPM-Res18	27.02	55.47	23.73	26.13	10.85	76.39	40.14	14.16	40.77	12.92	44.06	50.42
SiamFCOS	28.74	53.54	23.39	26.24	10.81	73.10	42.28	17.89	42.16	12.40	44.37	48.54
ECO	24.38	45.92	23.90	23.36	11.38	75.76	48.21	14.72	44.76	7.97	46.51	47.31
SiamCAR	28.90	48.06	29.03	27.63	17.82	**78.88**	45.27	13.01	43.52	12.52	47.82	54.79
KYS	40.80	55.25	35.23	35.70	33.30	73.71	46.77	17.58	44.42	24.61	49.32	55.85
ATOM	40.17	53.91	36.45	34.39	36.70	73.24	54.37	23.58	49.77	25.84	52.19	55.68
Dimp	40.87	55.29	36.57	40.03	32.42	73.53	48.82	29.93	48.91	25.57	52.47	58.25
ATOM-MU	38.67	53.84	35.37	35.73	35.44	74.30	52.16	26.81	49.84	24.85	52.61	54.02
SiamRPN++LT	45.50	71.71	47.09	43.61	44.70	75.88	52.75	18.66	48.15	32.35	54.34	61.17
SPLT	49.92	51.73	**51.75**	41.69	54.76	72.46	50.82	26.39	48.65	37.89	54.63	53.10
PrDimp	**57.43**	**79.52**	49.40	**50.05**	49.09	74.29	53.68	31.43	51.90	37.42	56.50	57.02
LTDSE	56.25	75.55	53.65	49.89	**56.72**	71.19	52.04	**37.79**	52.22	**48.70**	56.51	**64.29**
Super-Dimp	53.37	**78.79**	46.59	47.45	46.77	75.39	**56.50**	31.99	**53.69**	35.58	**57.72**	59.49
GlobalTrack	**68.98**	**79.47**	**63.42**	**57.34**	**67.78**	74.38	**60.24**	**43.02**	**58.46**	**58.48**	**63.86**	**66.24**
SiamRCNN	**68.17**	78.49	**67.66**	**57.23**	**73.92**	**78.78**	**61.89**	**42.48**	**60.10**	**64.04**	**65.41**	**70.83**

基线跟踪器的成功率图和精确率图如图 4.7 所示，图中只选取了排序前 20 的跟踪器以保证清晰的感官。SiamRCNN 在除了 TIR 测试集的精确率外的七个评价指标上均取得了位列第一的性能，在 TIR 验证集上的精确率得分为 95.70%、成功率为 71.52%，在 TIR 测试集上取得了最高的 63.60%的成功率，而 Globaltrack 超过

SiamRCNN 取得了最好的精确率——87.13%。在 RGB 序列上，SiamRCNN 分别在成功率和精确率上超过了第二名 GlobalTrack 4.68%和 3.06%。

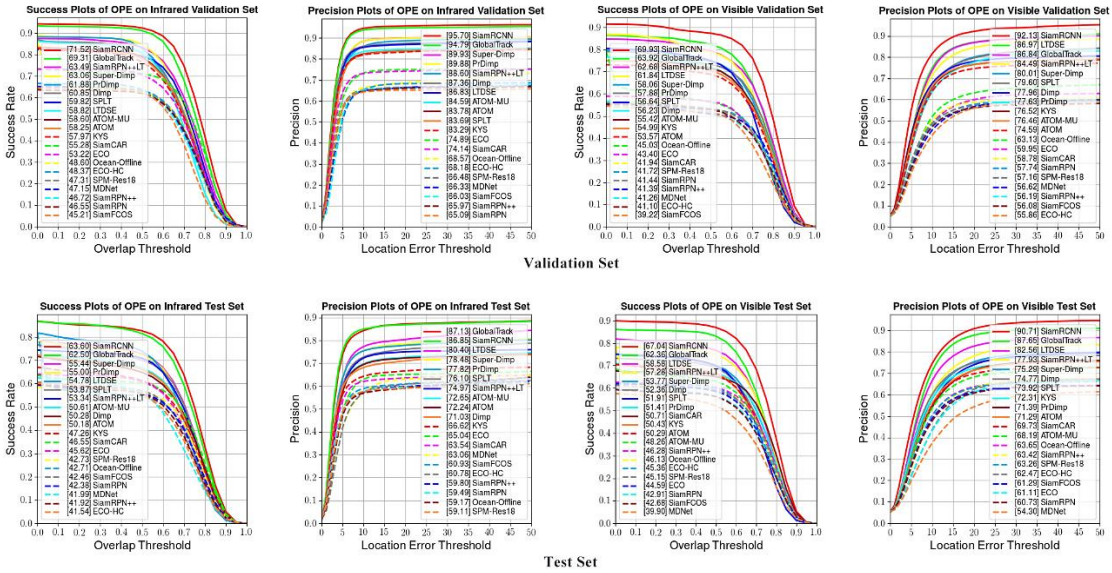


图 4.7　基于 Protocol I 的 Anti-UAV 的成功率图和精准率图

Figure 4.7　The success plot and precision plot on Anti-UAV under Protocol I

所有跟踪器中，最好的带有在线更新的机制的跟踪器是 Super-Dimp。除基于深度学习的跟踪器外，ECO-HC 是性能最好的基于相关滤波的跟踪器。以 TIR 测试集为例，Super-Dimp 的成功率为 55.44%、精确率为 78.48%，ECO-HC 在基于相关滤波的跟踪器中取得了最佳的成功率（41.54%）和精确率（60.78%）。

属性性能：为了更加综合地分析现有跟踪器面临的不同挑战时的性能，各种属性的评测同样如表 4.2 和表 4.3 所示。以 TIR 测试集为例，在大部分属性上，SiamRCNN 和 GlobalTrack 都取得了较好的性能，两者主要的属性性能差异在 FM、LR 和 LI 三个属性上。SiamRCNN 分别在 FM、LR、LI 上比 GlobalTrack 高 4.24%、6.14%、5.56%，具有较明显的优势。然而，在其他属性方面，GlobalTrack 大都处于劣势或者以微弱优势领先。各个属性性能的前二基本被 SiamRCNN 和 GlobalTrack 包揽。

由于 OV 通常不会出现在短时跟踪视频序列当中，所以在这个属性上可以看出基于长时跟踪的跟踪器性能更好。由于多样的采集飞行策略，Anti-UAV 中的快速移动、尺度变化和远处无人机的低分辨率同样为具有挑战性的属性。由于数据集内

带有 TC 属性的视频序列最多，因而根据跟踪难度分为TC_{easy}、TC_{med}和TC_{hard}三类，其中TC_{hard}为最具难度的属性。以 TIR 测试集为例，TC_{hard}性能最好的跟踪器为 GlobalTrack，仅为 43.02%。这也是 Anti-UAV 数据集收集发布的原因之一，缺乏红外交迭场景的数据不利于深度学习在反无人机感知该领域的发展。后面的基于 protocol II 的实验也验证了在 Anti-UAV 训练上进行微调或者从头训练的跟踪，在面对红外交迭场景会具有更好的鲁棒性。

4.3 基于信息交互的训练策略研究现状

■ 基于图像内信息交互的训练策略

利用单张图像内的多个实例，探索如何合理并且高效地使用这些实例以帮助提高跟踪器的性能成为当下研究方向之一。由于二维人体姿态估计（Human Pose Estimation）数据集中遮挡等具有挑战性的样本过少而不利于神经网络学习，一些研究工作者提出了一种通过粘贴背景块覆盖人体关键点来人工制造遮挡现象和构造歧义部位的数据增强方法[125]。

SPM 提出了一种基于 SiamFC 的串行和并行结合的目标跟踪匹配框架，以满足对更强的判别性能力和鲁棒性的需求。在粗匹配阶段注重鲁棒性，以减少同一目标的类内多样性为首要需求，即尽管目标在表观上发生了巨大的变化，但仍寄希望于它能够被成功跟踪。因此，在粗匹配阶段，要求跟踪器对于与跟踪目标同类的实例均有高响应，而在后续匹配阶段进行过滤。GlobalTrack 提出了交叉查询损失（Cross Query Loss）函数，通过使用样本内共同存在的不同的目标特征以搜索同一幅图像并平均它们的预测损失，从而提高了跟踪器对于实例级别干扰物的判别性能力。提出的训练策略利用的图片对在特征提取部分特征共享，因此共享绝大部分计算量。

■ 基于图像间信息交互的训练策略

利用图像内所有实例的训练策略已经取得了一定成效，但仅仅单个图像上的所有实例信息仍然存在一定的局限性。因此，一些工作开始关注基于图像间信息交互的训练策略。在三维人体姿态估计数据集上，从正则化的角度出发，部分科研工作者使用来自额外数据集的随机放置的遮挡物合成遮挡训练数据[126]。在图像分类任务中，Mixup[127]按一定比例随机融合两个甚至是多个样本的像素值，而 Cutmix[128]

则是随机选取样本的一部分区域，并在整个训练样本中采样，以采样数据的区域像素值进行填充，两者的分类标签均按照混合的比例进行分配。这样的图像间信息交互的训练策略一定程度上扩充了数据集，达到了数据增强的效果，以极少计算量为代价增加了模型的分类性能。

与上述方法不同，DaSiamRPN[129]摒弃了在图像层面进行信息交互的做法，转而移在特征层面进行。从样本不均衡的角度出发，作者认为训练数据中缺少正样本类别以及难例负样本进一步阻碍了跟踪器的学习。因此，在训练阶段引入公开检测数据集生成的正样本解决模型的泛化能力，构建难负例样本，提高跟踪器的判别性能力。

基于图像内信息交互的跟踪器训练策略需要训练样本中包含多个目标的标注，而单目标跟踪数据集中单张图像仅有一个跟踪目标的标注，所以上述训练策略需要引入目标检测数据集以辅助训练过程，在单目标跟踪的训练集上无法使用。

4.4　基于双流语义一致性的无人机跟踪训练策略

4.4.1　算法概述与创新点

特定针对于无人机目标跟踪，利用 Anti-UAV 数据集中只有无人机这一通用类别，设计更加高效的训练策略帮助跟踪器学习。本小节将跟踪器的双流语义学习分为鲁棒性学习和判别性学习两个阶段，通过更加高效地利用不同序列的无人机目标的特征，在第一阶段保持类别级别的语义流一致性加强跟踪器的鲁棒性，而在第二阶段保持实例级别的语义流一致性以增强跟踪器的判别性。方法的创新点在于：

1）提出了基于双流语义一致性的训练策略，驱使跟踪器在不同的训练阶段采用不同组合的模板-搜索图像对来分别增强跟踪器的鲁棒性和判别力。

2）没有引入额外数据集，更加高效地利用单目标跟踪数据集序列间无人机目标的特征，通过信息交互的方式改善跟踪器的学习方式，且几乎没有增加计算量，丝毫不影响测试时跟踪器的推理时间。

4.4.2　算法介绍

基于双流语义一致性（Dual-Flow Semantic Consistency，DFSC）的训练策略选

取了 GlobalTrack 作为基线跟踪器，我们提出的跟踪器的推理过程采用一种检测式的跟踪方式（Tracking by Detection）：给定首帧无人机目标的包围框，将首帧图像输入到卷积神经网络中，通过 ROI Align 获得模板特征，利用模板特征对接下来的每一帧进行调制，将调制后的特征输入区域候选网络获得候选框；然后选取一定数量的候选框输入 RCNN 部分进行第二次具体的分类和回归；最后得分最高的预测框即为无人机的位置，将得分和预先设定好的阈值进行比较，若低于阈值即表示无人机不在当前帧内存在，不对包围框进行输出。

训练过程如图 4.8 所示中，训练过程中输入为不同视频序列的两帧（在选定视频序列中进行随机采样选出某一帧，然后再在选定帧的前后一定范围内继续进行采样），跟踪器的训练过程主要分为两个阶段：类别级别的语义调制（Class-level Semantic Modulation，CSM）和实例级别的语义调制（Instance-level Semantic Modulation，ISM）。

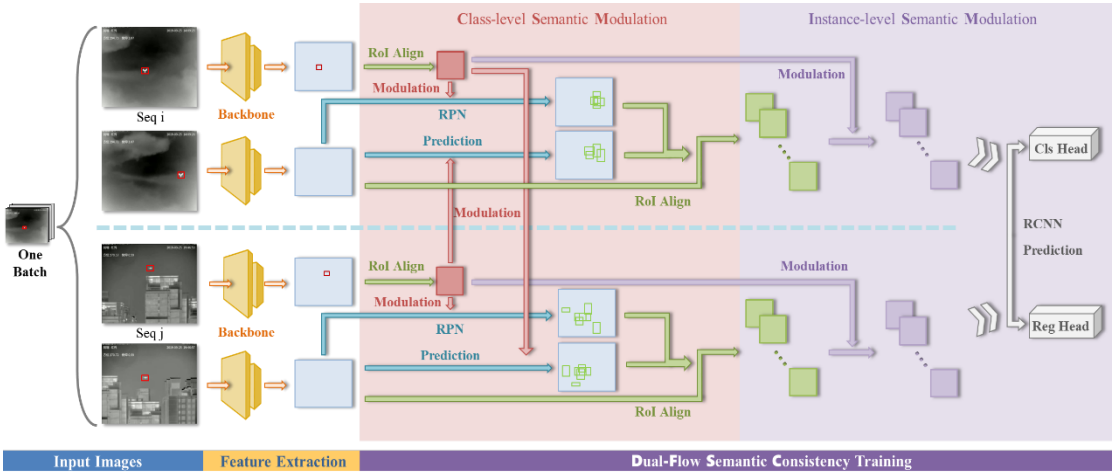


图 4.8　DFSC 训练策略的流程图

Figure 4.8　The pipeline of DFSC training strategy

■　类别级别的语义调制

在这个阶段通过不同视频序列间的信息交互，驱使跟踪器学习到无人机这一通别的鲁棒的特征。在标注包围框和视频帧输入到卷积神经网络后，获得前后模板帧和搜索帧两帧的特征，然后通过 ROI Align 获得前一帧给定包围框对应的模板特征（也可称为查询特征，Query），使用不同视频序列的模板特征对后一帧进行同序列的调制和跨序列的交叉调制如下所示，

$$\hat{t}_{ij} = f_{CSM}(z_i,\ x_i) = f_{out}\left(\left(f_z(z_i) \otimes f_x(x_j)\right)\right). \tag{4.2}$$

其中z_i表示第 i 个视频序列的模板帧的模板特征，x_j表示第 j 个视频序列的搜索帧的区域特征，$f_z(\cdot)$和$f_x(\cdot)$分别表示z_i、x_j进行映射的函数，$\otimes$为类别级别语义调制的卷积操作（以$f_z(z_i)$作为卷积核对$f_x(x_j)$进行卷积），$f_{out}(\cdot)$为将前一步得到的调制区域特征进行通道数上的调整的函数，以上的操作步骤简称为$f_{CSM}(\cdot)$函数，$\hat{t}_{ij}$表示 CSM 阶段得到的经过类别级别调制后的区域特征。若批大小（Batch Size）为 n，那么$i,j \in [0,n)$。

根据定义，可以发现当 i 和 j 相等时为同序列的模板特征进行调制，当 i 和 j 不相等的时候为跨序列的模板特征进行调制。得到调制后的区域特征，将会和预先设定好的锚点框一起输入 RPN，进行分类和回归任务，具体的损失函数如下所示：

$$\begin{aligned} L_{CSM}(z_i, x_i, z_j, x_j) &= L_{same} + \alpha L_{cross} \\ &= \sum_{i,\, j \in n, i=j} L_{rpn}(\hat{t}_{ij}) + \alpha \sum_{i,\, j \in n, i \neq j} L_{rpn}(\hat{t}_{ij}). \end{aligned} \tag{4.3}$$

其中，L_{CSM}为 CSM 阶段的损失函数，L_{same}表示将同序列模板特征调制后的区域特征送入 RPN 得到的损失函数，L_{cross}表示将跨序列模板特征调制后的区域特征送入 RPN 得到的损失函数，α为调节L_{same}和L_{cross}的权重因子，L_{rpn}表示在每一个区域特征上分类和回归的损失函数，具体如下定义：

$$L_{rpn}(\hat{t}_{ij}) = \frac{1}{N_{cls}} \sum_n L_{cls}(s_n,\ s_n^*) + \beta \frac{1}{N_{reg}} \sum_n L_{reg}(p_n,\ p_n^*). \tag{4.4}$$

其中，β是一个用来调节L_{cls}和L_{reg}的权重因子，s_n和s_n^*分别表示第 n 个预设锚点框的分类得分和对应的真值标注，p_n和p_n^*分别表示第 n 个预设锚点框的位置编码和对应的真值标注，N_{cls}和N_{reg}分别对应需要进行平均化的总数。

■　实例级别的语义调制

上一阶段驱使跟踪器对于无人机这一通用类别有着更强的鲁棒性，而在这一阶段需要迫使跟踪器更多的关注于每个序列对应的特定的无人机实例，使得跟踪器在遇到复杂背景或者是相似干扰物时能够准确判断，从而达到增强跟踪器的判别能力。在获得上一阶段输出的每个序列的高分候选包围框后，将对这些候选包围框进

行对应序列模板特征的实例级别语义调制，如下所示：

$$\hat{t}_K = f_{ISM}(z,\ x_k) = f'_{out}\left(\left(f'_z(z)\odot f'_x(x_k)\right)\right). \tag{4.5}$$

其中，x_k表示经过类别语义调制后的第 k 个候选包围框对应的 ROI 特征，z 为x_k所属的视频序列的模板特征，$f'_z(\cdot)$和$f'_x(\cdot)$分别表示 z、x_k进行映射的函数，$\odot$为实例级别语义调制的哈达玛积(Hadamard product)，$f'_{out}(\cdot)$为将前一步得到的调制 ROI 特征进行通道数上的调整的函数，以上的操作步骤简称为$f_{ISM}(\cdot)$函数，$\hat{t}_k$表示 ISM 阶段得到的经过实例级别调制后的第 k 个 ROI 特征。

在得到经过实例级别语义调制的 ROI 特征后，将输入 RCNN 以进行第二次分类和回归任务，具体的损失函数如下所示：

$$L_{ISM}(z, x) = \frac{1}{N_{pnum}}\sum_k L_{rcnn}(t_k). \tag{4.6}$$

其中，L_{ISM}为 ISM 阶段的损失函数，N_{pnum}为输入的所有候选包围框的总数，L_{rcnn}为在每一个调制后的 ROI 特征上得到的分类和回归损失函数，如下所示：

$$L_{rcnn}(\hat{t}_K) = L'_{cls}(s'_n,\ s'^*_n) + \beta L'_{reg}(p'_n,\ p'^*_n). \tag{4.7}$$

其中，s'_n和s'^*_n分别表示第 n 个预测包围框的分类得分和对应的真值标注，p'_n和p'^*_n分别表示第 n 个预测包围框的位置编码和对应的真值标注，β是一个用来调节L'_{cls}和L'_{reg}的权重因子。

4.5 实验验证

接下来将从实验配置、结果分析和消融实验三个方面来进行深入的剖析。

4.5.1 实验配置

超参数设置：由于 Anti-UAV 为多模态光学无人机目标跟踪数据集，在基于 Protocol II 的设置下，对于 RGB 和 TIR 视频序列各自设置了一套超参数设定。

对于 RGB 序列而言，使用了 GlobalTrack 提供的预训练模型作为提出算法的初始化模型。整个训练过程中一共有 12 个循环，初始的学习率设置为 0.02，然后分别在第 8 次循环和第 11 次循环的时候降十分之一。在 TIR 序列上，使用了 Faster RCNN 迁移过来的模型参数作为初始化。整个训练的过程一共有 18 次循环，初始的学习率参数仍设置为 0.02，并且分别在第 12 次循环和第 15 次循环的时候降为

0.002 和 0.0002。

在 CSM 和 ISM 的训练过程中，分类和回归分别采用了交叉熵损失和平滑 L1 损失（Smooth L1 Loss），批大小设置为每个 GPU 上为 2，使用的 GPU 为 NVIDIA GeForce GTX 1080Ti。

训练数据：Anti-UAV 的 TIR 序列被用来训练 TIR 无人机跟踪器，Anti-UAV 的 RGB 序列被用来训练 RGB 无人机跟踪器，而由于 RGB 无人机跟踪器使用了 GlobalTrack 提供的预训练模型，该预训练模型在训练过程中使用了 MS COCO、LaSOT 和 GOT-10k 三个数据集。

4.5.2　实验结果及分析

表 4.4 为不同训练策略之间在 mSA 上的性能比较，large-scale 训练策略即基于 Protocol I 下使用 MS COCO、LaSOT 和 GOT-10k 三个大型数据集训练的方法，nomal 训练策略即基于 Protocol II 只使用 Anti-UAV 的对应光学模态训练集进行训练的方法，DFSC 训练策略即为本小节提出的在 Protocol II 设定下基于双流语义一致性的训练策略。接下来依旧从总体性能和属性性能两个方面进行分析。

表 4.4　不同训练策略在 mSA（%）上的性能比较

Table 4.4　Comparisons in terms of mSA (%) with different training methods

训练策略	类型	红外											可见光
		OV	OC	FM	SV	LI	TC				LR	All	All
								TCmed	THard	TCall			
large-scale		69.21	78.62	73.35	66.11	76.33	76.47	63.08	43.45	65.90	60.26	72.00	67.28
normal	val	**78.09**	81.42	78.66	76.61	79.95	81.23	76.56	74.04	78.49	73.55	79.60	73.25
DFSC (Ours)		77.72	**82.70**	**79.34**	**77.58**	**80.31**	**81.48**	**76.83**	**75.65**	**79.04**	**74.33**	**80.09**	**73.73**
large-scale		68.98	**79.47**	**63.42**	**57.34**	**67.78**	74.38	60.24	43.02	58.46	**58.48**	63.86	66.24
normal	test	**70.44**	68.94	60.66	55.48	59.78	77.07	64.64	44.61	61.68	52.94	65.36	69.27
DFSC (Ours)		70.16	68.07	60.95	55.55	60.13	**77.96**	**65.85**	**45.59**	**62.75**	53.10	**66.04**	**69.84**

总体性能：如表 4.4 所示，加粗的数字表示当前表格里同属性评测里性能位列第一，因此，可以发现 DFSC 算法在验证集和测试集上 RGB 序列和 TIR 序列均取得了最好的性能。同样的，DFSC 在四项评测中成功率和精确率也是最高的，获得了一致的提升。和基线方法 normal 训练策略相比，在 TIR 序列上，DFSC 在验证集和测试集分别获得了 0.49%、0.68%的 mSA 的性能提升；而在 RGB 序列上，DFSC 在验证集和测试集上各自取得了 0.48%、0.57%的性能增益。和 large-scale 方法相比，normal 和 DFSC 在验证集上的性能提升更加的明显，这是因为训练集和验证集

是同源的，也即两者的视频序列可能是来自于同一视频序列的不同切分子序列，具有较高的同质性，因而会获得更大的性能提升。

属性性能：分析方法在不同属性上的性能可以帮助理解训练策略各自的优势和劣势。如表 4.4 所示，在验证集上 DFSC 方法几乎在绝大部分属性上均超越了基线方法，尤其是在 OC 属性上。和 normal 训练策略相比，DFSC 分别在 OC、SV 和 LR 上在 mSA 上超过了 1.28%、0.97%和 0.78%，而在其他属性上增益的性能有限。

在测试集上，DFSC 策略在TC_{all}上取得了 62.75%的性能，足足比 normal 训练策略高出 1.07%。除此之外，以 normal 训练策略为基线，DFSC 分别在TC_{easy}、TC_{med}和TC_{hard}上有着 0.89%、1.21%和 0.98%的增益。large-scale 训练策略在除了 TC 和 OV 两个属性上均取得了最高的性能，然而整体的性能却不如 DFSC 训练出来的跟踪器。经过分析可能一方面由于 large-scale 方法使用了远远超出 DFSC 的大体量的数据集，由大量数据驱动促使跟踪器能够学到更加具有泛化性的特征，从而有益于无人机目标跟踪。另一方面，由于在 Anti-UAV 训练集上进行了训练，使得跟踪器在面对 TC 场景时更加稳定，增加了跟踪器的判别能力，能够有效地区分复杂背景和相似干扰物。

为了更好地展示 DFSC 训练策略的优势，从 Anti-UAV 测试集中挑选出来成功的跟踪序列进行了可视化，如图 4.9 所示。由可视化可知，提出的 DFSC 算法可以帮助跟踪一定程度上有效地解决无人机目标跟踪过程中的重大挑战。

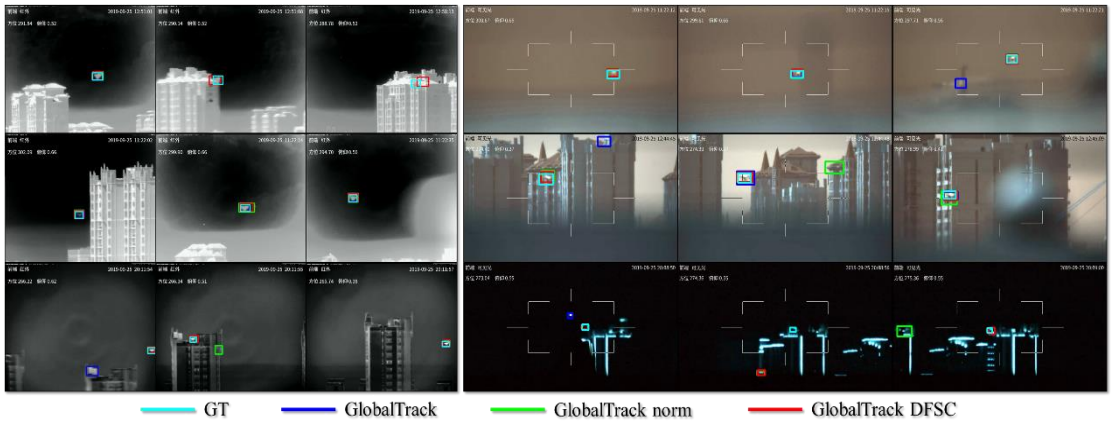


图 4.9　不同训练策略下跟踪序列的可视化

Figure 4.9　Visualization of successful tracking sequences with different training methods

4.5.3　消融实验

这节主要分析 DFSC 训练策略中监督任务以及跨序列信息交互在训练过程中带来的影响，所有消融实验以 Anti-UAV 测试集为例，性能如表 4.5 和表 4.6 所示。

监督任务：本消融实验意在分析 DFSC 训练过程中不同监督任务带来的影响。DFSC 分为 CSM 和 ISM 两个阶段，在 CSM 阶段同序列、跨序列模板特征调制同时进行，跨序列调制后的区域特征进行分类和回归两个监督任务，而在 ISM 阶段为保持实例级别的类内差异性，只有同序列的模板调制。接下来对设置的消融实验进行阐述：DFSC-all 表示在 ISM 阶段采用了同序列和跨序列模板特征调制，DFSC-cls 表示在 CSM 阶段跨序列模板调制后的区域特征只进行分类任务，DFSC-reg 表示在 CSM 阶段跨序列模板调制后的区域特征只进行回归任务。

表 4.5　DFSC 算法在测试集上有关 mSA 的消融实验

Table 4.5 The ablation study of DFSC algorithm in terms of mSA (%) on the test set

训练策略	红外											可见光
	OV	OC	FM	SV	LI	TC				LR	All	All
						TCeasy	TCmed	TChard	TCall			
DFSC-all	68.26	66.23	58.66	54.60	56.96	77.38	63.65	44.14	61.12	50.51	63.99	67.94
DFSC-cls	70.22	**68.47**	60.78	55.49	60.02	**78.17**	65.75	45.07	62.60	52.29	**66.12**	69.53
DFSC-reg	**70.31**	68.26	60.70	55.31	59.05	77.52	65.52	**46.74**	**62.82**	51.76	66.06	68.95
DFSC	70.16	68.07	**60.95**	**55.55**	**60.13**	77.96	**65.85**	45.59	62.75	**53.10**	66.04	**69.84**
DFSC-α0.25	**70.63**	**69.18**	60.80	**55.74**	59.61	**78.56**	65.70	**46.26**	**63.00**	52.75	**66.33**	**70.31**
DFSC-α0.5	70.37	68.32	60.20	55.01	58.86	77.17	64.49	45.28	61.82	51.52	65.25	69.55
DFSC-α1	70.16	68.07	**60.95**	55.55	**60.13**	77.96	**65.85**	45.59	62.75	**53.10**	66.04	69.84
DFSC-α2	69.76	68.19	60.46	55.20	59.53	77.57	65.43	45.42	62.41	52.85	65.61	69.98

表 4.6　DFSC 在测试集上有关精确率和成功率的消融实验

Table 4.6 The ablation study of DFSC in terms of precision and success on the test set

训练策略	红外		可见光	
	精确率	成功率	精确率	成功率
Large-scale	85.34	61.13	88.53	63.00
Normal	87.61	62.88	93.30	65.22
DFSC	87.77	63.24	**93.87**	65.87
DFSC-all	86.97	61,59	91.87	64.20
DFSC-cls	87.47	63.14	93.72	65.56
DFSC-reg	**87.87**	63.35	93.26	65.24
DFSC-α 0.25	87.52	**63.40**	93.65	**66.24**
DFSC-α 0.5	87.08	62.53	93.84	65.66
DFSC-α 1	87.77	63.24	93.83	65.50
DFSC-α 2	87.83	62.81	93.74	65.94

如表 4.5 上部所示，以 TIR 序列为例，DFSC-cls 取得了最高的 mSA，但是性

能差异并不大，同时在表 4.6 中却发现成功率和精准率相对较低。与之相反，DFSC-reg 的 mSA 最低，但是却在其中有着最高的成功率和精准率。可以发现 DFSC-cls 使得跟踪器对于无人机的状态判断能力更强，但是对于无人机的定位能力相对偏弱，说明了 DFSC 中分类的监督任务驱使跟踪器更关注于无人机状态的感知，而回归任务的监督使得跟踪器更侧重于无人机的定位能力，两者的综合使用使得跟踪器的性能更加全面、更加稳定。

在全部的评测指标下，DFSC-all 的性能均表现很差，说明了在 ISM 阶段的类别级别语义调制会对神经网络的学习造成恶劣的影响，使得跟踪器对不同无人机实例间产生疑惑，因此，验证了在 ISM 阶段只使用实例级别的语义调制有利于增强跟踪器的判别性。

α比例影响：本消融实验旨在说明 CSM 阶段同序列、跨序列监督所占权重比例大小的影响。α为公式 4.3 中所示，CSM 阶段跨序列损失函数L_{cross}和同序列损失函数L_{same}的比值。当α等于 1 时，即为基线方法 DFSC；当α等于 0 的时候，即为 normal 训练策略。如表 4.5 和表 4.6 的底部所示，当α由小到大变动时，跟踪器的性能产生了较大的变动。

随着α的增加，mSA 的性能整体先升再降，在α等于 0.25 的时候达到顶峰。以 TIR 序列为例，这时 DFSC-α0.25 的 mSA 为 66.33%，高 normal 训练策略接近一个点。这时α过高或者过低都会引起性能的下降，通过分析认为这是因为这α过大的时候跟踪器在学习时会过多地关注于跨序列的信息交互，一定程度上阻碍了跟踪器学习类别级别的语义特征，从而影响了跟踪器的鲁棒性。而当α过小的时候，DFSC 会退化为 normal 训练策略，并不具备图像间信息交互的优势。一个合适的比例（α等于 0.25）可以很好地权衡同序列和跨序列的语义特征。

4.6 本章小结

本章从各方面综合地介绍了面向无人机管控的无人机目标跟踪数据集 Anti-UAV，包括数据获取、处理、属性分布等，详细介绍了实验的评测规则和评测指标。在 Anti-UAV 上基于 Protocol I 搭建了超过 40 个跟踪器的大型高质量基线方法，全面分析了这些跟踪器的优势和劣势。通过分析基于信息交互的训练策略的研究现

状，并结合 Anti-UAV 无人机跟踪数据的特性，提出了基于双流语义一致性的 DFSC 训练策略，充分利用跨序列语义的信息交互，合理地分阶段地提升了跟踪器的鲁棒性和判别能力。在 Protocol II 的条件下，对提出的 DFSC 算法进行了深入分析和可视化展示，并从监督任务和权重因子两个角度剖析了算法的性能，验证了 DFSC 训练策略的高效性。

第5章　总结与展望

关键弱小目标感知作为计算机视觉领域新兴的研究内容，涉及到多个领域的知识。随着无人机技术的研究和发展，无人机作为一柄双刃剑，虽可便利大众生活，但是也要防范相应的安全隐患，这方面得到了越来越多科研工作者的注意。本文从弱小人体目标检测和无人机目标跟踪两个任务出发，对基于无人机监控系统的关键弱小目标感知技术进行了深入研究。本章中，将对上述章节讨论的内容进行全面的归纳，并对日后的研究方向进行规划。

5.1　全文总结

基于无人机监控系统的关键弱小目标感知技术目前还是一个新兴的领域，在面对真实的复杂场景时，目前的感知技术距离实际的应用还有一定的差距。本文面向无人机相关的应用背景，依托无人机监控系统，以关键弱小目标感知算法为核心，延伸出了基于无人机航拍的弱小人体目标检测技术和基于多模态互感的无人机跟踪技术，针对如上任务的创新点总结如下：

基于无人机航拍的弱小人体目标检测技术：本文在第三章首先介绍了联合提出的弱小人体目标检测数据集 TinyPerson，并对数据集的情况进行了简要的介绍。从数据驱动的角度出发，引入额外目标检测数据集以驱使检测器学习到更多检测任务相关的知识。由于预训练数据集和下游目标任务数据集尺度分布存在明显差异，提出了基于精细尺度匹配的弱小人体目标检测预训练策略，通过图像级别和实例级别两个思路实现具有直接指导性的尺度迁移，进一步减少预训练数据和下游目标数据间尺度分布的差异性。在此基础上进行预训练，为下游目标任务提供更加合适的初始化模型。最后，经过大量实验验证了该预训练策略的有效性。

基于多模态互感的无人机跟踪技术：在第四章中搭建了面向无人机管控的多模态无人机目标跟踪数据集 Anti-UAV，从多个角度对 Anti-UAV 进行了较为详细的介绍。通过充分利用训练数据间的信息交互，提出了基于双流语义一致性的训练策略。将跟踪器的目标识别解耦为两个阶段：在类别级别语义调制阶段，引入不同视频序

列间的特征交叉调制，旨在降低特征的类内差异性，加强跟踪器的鲁棒性；而在后续的实例级别语义调制阶段，更加侧重于跟踪器的判别能力，仅使用同视频序列的模板特征进行调制。同时，该方法仅作用于训练阶段，在训练过程中由于特征共享，只增加了很小的计算量，并且完全不影响测试阶段的推理速度。

本文中采集的弱小目标感知数据集和搭建的高基准评测平台均会被公开，以供后续科研工作者使用，更好地推动相关领域的发展。

5.2 未来工作展望

以产学研结合为目标，追求计算机视觉技术落地有很长的路要走，从当前的数据形式和实验结果来看，可以发现距离实际的落地应用还有不小的差距。未来可以从如下方面入手进行后续研究，进一步完善基于无人机监控的弱小目标感知技术。

- 基于精细尺度匹配的弱小人体目标检测

数据集形式：和通用目标检测、人脸检测等任务的经典数据集相比，TinyPerson 数据量较小，而深度学习的方法需要依靠大量数据驱动，目前的弱小人体目标检测数据集还远远不够，尝试通过网络数据收集更多的相关视频、图片以扩充数据样本。除此之外，TinyPerson 数据集依托于海上快速救援的应用背景而搭建，后续可以面向城市等更多具有实际意义的背景以丰富场景多样性。从人力、物力角度而言，尝试利用游戏场景合成数据是一种简单易行的方法。例如，通过 GTA5 游戏提供的接口，对游戏场景进行布置，控制游戏人物的位置、着装、动作等各种状态以截取合成数据样本。这样的收集方式具备标注获取容易、数据量极大等各种优势，但是会和真实场景有域层面的差异。

当前方法推进：目前，基于精细尺度匹配的弱小人体目标检测算法只关注于绝对尺度方面的差异，尚没有对相对尺度分布在整个预训练过程中产生的影响进行分析。除此之外，当前的尺度匹配策略操作的对象为输入图像，直接对前景实例的尺度缩放会严重破坏图片的背景结构，尝试探究在特征层面进行尺度匹配的方式。为了更好地验证尺度匹配算法的完备性，可以尝试在飞机、车辆等其他关键弱小目标数据集上验证，以佐证算法的泛化性能。

■　基于多模态互感的无人机跟踪

数据集形式：Anti-UAV 是针对于单无人机目标而搭建的无人机跟踪数据集。在现阶段的反无人机探测识别和跟踪中，以雷达为主、光电为辅的复合策略已经成为主流。然而，这样的探测手段面对无人机群时将会显得束手无策。针对这样的技术难点，可以尝试通过拍摄并搭建大型的多无人机跟踪数据集，通过视觉的解决方案来弥补技术上的不足之处。

当前方法推进：由于基于双流语义一致性的训练策略目前在无人机跟踪数据集上得到了很好的验证，下一步可以尝试推广到更加大型的通用目标跟踪数据集上，验证算法的泛化性。利用某些通用目标跟踪数据集中视频序列的类别标签，在训练时按照类别进行视频序列的采样，然后进行前后帧的采样。

非对齐数据的多模态融合跟踪：本文中的无人机跟踪实验均建立在单模态的基础上，由于 Anti-UAV 中相应的 RGB 序列和 TIR 序列没有进行配准操作，在实际的多模态融合跟踪过程中，不建议使用像素级别融合的方案，尝试探索特征级别和决策级别的融合跟踪方案，充分利用数据集中多模态的优势。

参考文献

[1] Hinton G, Osindero S and Teh Y. A Fast Learning Algorithm for Deep Belief Nets[J]. Neural Computation, 2014, 18(7): 1527-1554.

[2] Rumelhart D, Hinton G and Williams R. Learning Representations by Back Propagating Errors[J]. Nature, 1986, 323(6088): 533-536.

[3] Krizhevsky A, Sutskever I and Hinton G. ImageNet Classification with Deep Convolutional Neural Networks[C]. In Proceedings of Neural Information Processing Systems, 2012: 1106-1114.

[4] Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2009: 248-255.

[5] Silver D, Huang A, Maddison C, et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484–489.

[6] Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge[J]. Nature, 2017, 550(7676): 354-359.

[7] 孟琭,杨旭.目标跟踪算法综述[J]. 自动化学报, 2019, 45(07): 1244-1260.

[8] 张慧,王坤峰,王飞跃.深度学习在目标视觉检测中的应用进展与展望[J]. 自动化学报, 2017, 43(08): 1289-1305.

[9] Yu X, Gong Y, Jiang N, et al. Scale Match for Tiny Person Detection[C]. In Proceedings of IEEE Winter Conference on Applications of Computer Vision, 2020: 1246-1254.

[10] Lin T, Maire M, Belongie S, et al. Microsoft COCO: Common Objects in Context[C]. In Proceedings of European Conference on Computer Vision, 2014: 740-755.

[11] Zhang S, Benenson R and Schiele B. CityPersons: A Diverse Dataset for Pedestrian Detection[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2017: 4457-4465.

[12] Yang S, Luo P, Loy C, et al. WIDER FACE: A Face Detection Benchmark[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2016: 5525-5533.

[13] Chen Y, Zhang P, Li Z, et al. Stitcher: Feedback-driven Data Provider for Object Detection[J]. ArXiv, abs/2004.12432, 2020.

[14] Xia G, Bai X, Ding J, et al. DOTA: A Large-scale Dataset for Object Detection in Aerial Images[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2018: 3974-3983.

[15] Redmon J, Divvala S, Girshick R, Et al. You Only Look Once: Unified, Real-Time Object Detection[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2016: 779-788.

[16] Liu W，Anguelov D，Erhan D, et al. SSD: Single Shot MultiBox Detector[C]. In Proceedings of European Conference on Computer Vision, 2016: 21-37.

[17] Lin T, Goyal P, Girshick R, et al. Focal Loss for Dense Object Detection[C]. In Proceedings of IEEE International Conference on Computer Vision, 2017: 2999-3007.

[18] Ren S，He K，Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[C]. In Proceedings of Neural Information Processing Systems, 2015: 91-99.

[19] Lin T, Dollar P, Girshick R, et al. Feature Pyramid Networks for Object Detection[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2017: 936-944.

[20] Girshick R. Fast R-CNN[C]. In Proceedings of IEEE International Conference on Computer Vision, 2015: 1440-1448.

[21] Li Y, Chen Y, Wang N, et al. Scale-Aware Trident Networks for Object Detection[C]. In Proceedings of IEEE International Conference on Computer Vision, 2019: 6053-6062.

[22] Liu S, Huang D, and Wang Y. Receptive Field Block Net for Accurate and Fast Object Detection[C]. In Proceedings of European Conference on Computer Vision, 2017: 404-419.

[23] Yu F and Koltun V. Multi-Scale Context Aggregation by Dilated Convolutions[C]. In Proceedings of International Conference on Learning Representations, 2016.

[24] Singh B and Davis L. An Analysis of Scale Invariance in Object Detection - SNIP[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2018: 3578-3587.

[25] Singh B, Najibi W, and Davis L. SNIPER: Efficient Multi-Scale Training[C]. In Proceedings of Neural Information Processing Systems, 2018: 9333-9343.

[26] Zhang S, Zhu X, Lei Z, et al. FaceBoxes: A CPU Real-time Face Detector with High Accuracy[C]. In Proceedings of IEEE International Joint Conference on Biometrics, 2017: 1-9.

[27] Zhang S, Zhu X, Lei Z, et al. S3FD: Single Shot Scale-invariant Face Detector[C]. In Proceedings of IEEE International Conference on Computer Vision, 2017: 192-201.

[28] Kisantal M, Wojna Z, Murawski J, et al. Augmentation for small object detection[J]. ArXiv, abs/1902.07296, 2019.

[29] Wu Y, Lim J and Yang M. Online Object Tracking: A Benchmark[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2013: 2411-2418.

[30] Wu Y, Lim J and Yang M. Object Tracking Benchmark[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1834-1848.

[31] Kristan M, Pflugfelder R, Leonardis A. The Visual Object Tracking VOT2014 Challenge Results[C]. In Proceedings of European Conference on Computer Vision Workshops, 2014: 191-217.

[32] Kristan M, Leonardis A, Matas J, et al. The Visual Object Tracking VOT2017 Challenge Results[C]. In Proceedings of IEEE International Conference on Computer Vision Workshops, 2017: 1949-1972.

[33] Kristan M, Berg A, Zheng L, et al. The Seventh Visual Object Tracking VOT2019 Challenge Results[C]. In Proceedings of IEEE International Conference on Computer Vision Workshops, 2019: 2206-2241.

[34] Smeulders A, Chu D, Cucchiara R, et al. Visual Tracking: An Experimental Survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(7): 1442-1468.

[35] Liang P, Blasch E, Ling H. Encoding Color Information for Visual Tracking: Algorithms and Benchmark[J]. IEEE Transactions on Image Processing, 2015, 24(12): 5630-5644.

[36] Li A, Lin M, Yang M, et al. NUS-PRO: A New Visual Tracking Challenge[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(2): 335-349.

[37] Valmadre J, Bertinetto J, Henriques J, et al. Long-Term Tracking in the Wild: A Benchmark[C]. In Proceedings of European Conference on Computer Vision, 2018: 692-707.

[38] Mueller M, Smith N, and Ghanem B. A Benchmark and Simulator for UAV Tracking[C]. In Proceedings of European Conference on Computer Vision, 2016: 445-461.

[39] Galoogahi H, Fagg A, Huang C, et al. Need for Speed: A Benchmark for Higher Frame Rate Object Tracking[C]. In Proceedings of IEEE International Conference on Computer Vision, 2017: 1134-1143.

[40] Fan H, Lin L, Yang F, et al. LaSOT: A High-Quality Benchmark for Large-Scale Single Object

Tracking[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2019: 5374-5383.

[41] M¨uller M, Bibi A, Giancola S, et al. TrackingNet: A Large-Scale Dataset and Benchmark for Object Tracking in the Wild[C]. In Proceedings of European Conference on Computer Vision, 2018: 310-327.

[42] Huang L, Zhao X, and 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] Davis J and Keck M. A Two-Stage Template Approach to Person Detection in Thermal Imagery[C]. In Proceedings of IEEE Winter Conference on Applications of Computer Vision Workshops, 2005: 364-369.

[44] Portmann J, Lynen S, Chli M, et al. People detection and tracking from aerial thermal views[C]. In Proceedings of IEEE International Conference on Robotics and Automation, 2014: 1794-1800.

[45] Wu Z, Fuller N, Theriault D, et al. A Thermal Infrared Video Benchmark for Visual Analysis[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition Workshops, 2014: 201-208.

[46] Felsberg M, Berg A, Häger G, et al. The Thermal Infrared Visual Object Tracking VOT-TIR2015 Challenge Results[C]. In Proceedings of IEEE International Conference on Computer Vision Workshops, 2015: 639-651.

[47] Felsberg M, Kristan M, Matas J, et al. The Thermal Infrared Visual Object Tracking VOT-TIR2016 Challenge Results[C]. In Proceedings of European Conference on Computer Vision Workshops, 2016: 824-849.

[48] Liu Q, He Z, Li X, et al. PTB-TIR: A Thermal Infrared Pedestrian Tracking Benchmark[J]. IEEE Transactions on Multimedia, 2020, 22(3): 666-675.

[49] Liu Q, Li X, He Z, et al. LSOTB-TIR: A Large-Scale High-Diversity Thermal Infrared Object Tracking Benchmark[C]. In Proceedings of ACM International Conference on Multimedia, 2020: 3847-3856.

[50] Davis J and Sharma V. Background-subtraction using contour-based fusion of thermal and visible imagery[J]. Computer Vision and Image Understanding, 2007, 106(2): 162-182.

[51] Torabi A, Massé G, and Bilodeau G. An iterative integrated framework for thermal-visible image registration, sensor fusion, and people tracking for video surveillance applications[J]. Computer

Vision and Image Understanding, 2012, 116(2): 210-221.

[52] Li C, Cheng H, Hu S, et al. Learning Collaborative Sparse Representation for Grayscale-Thermal Tracking[J]. IEEE Transactions on Image Processing, 2016, 25(12): 5743-5756.

[53] Li C, Zhao N, Lu Y, et al. Weighted Sparse Representation Regularized Graph Learning for RGB-T Object Tracking[C]. In Proceedings of ACM International Conference on Multimedia, 2017: 1856-1864.

[54] Li C, Liang X, Lu Y, et al. RGB-T object tracking: Benchmark and baseline[J]. Pattern Recognition, 2019, 96.

[55] Real E, Shlens J, Mazzocchi S, et al. YouTube-BoundingBoxes: A Large High-Precision Human-Annotated Data Set for Object Detection in Video[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2017: 7464-7473.

[56] Wang N and Yeung D. Learning a Deep Compact Image Representation for Visual Tracking[C]. In Proceedings of Neural Information Processing Systems, 2013: 809-817.

[57] Marvasti-Zadeh S, Cheng L, Ghanei-Yakhdan H, et al. Deep Learning for Visual Tracking: A Comprehensive Survey[J]. ArXiv, abs/1912.00535, 2019.

[58] Nam H and Han B, Learning Multi-domain Convolutional Neural Networks for Visual Tracking[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2016: 4293-4302.

[59] Jung I, Son J, Baek M, et al. Real-time MDNet[C]. In Proceedings of European Conference on Computer Vision, 2018: 89–104.

[60] He K, Gkioxari G, Dollár P, et al. Mask R-CNN[C]. In Proceedings of IEEE International Conference on Computer Vision, 2017: 2980-2988.

[61] Bertinetto L, Valmadre J, Henriques J, et al. Fully-Convolutional Siamese Networks for Object Tracking[C]. In Proceedings of European Conference on Computer Vision Workshops, 2016: 850-865.

[62] Li B, Yan J, Wu W, et al. High Performance Visual Tracking With Siamese Region Proposal Network[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2018: 8971-8980.

[63] Li B, Wu W, Wang Q, et al. SiamRPN++: Evolution of Siamese Visual Tracking With Very Deep Networks[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2019: 4282-4291.

[64] Ma D, Bu W and Wu X. Multi-Scale Recurrent Tracking via Pyramid Recurrent Network and Optical Flow[C]. In Proceedings of British Machine Vision Conference, 2018: 242.

[65] Chen B, Li P, Sun C, et al. Multi attention module for visual tracking[J]. Pattern Recognition, 2019, 87: 80-93.

[66] Song Y, Ma C, Wu X, et al. VITAL: VIsual Tracking via Adversarial Learning[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2018: 8990–8999.

[67] Simonyan K and Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[C]. In Proceedings of International Conference on Learning Representations, 2015.

[68] Liu Q, Lu X, He Z, et al. Deep convolutional neural networks for thermal infrared object tracking[J]. Knowledge Based Systems, 2017, 134: 189-198.

[69] Gao P, Ma Y, Song K, et al. Large Margin Structured Convolution Operator for Thermal Infrared Object Tracking[C]. In Proceedings of International Conference on Pattern Recognition, 2018: 2380-2385.

[70] Liu Q, Li X, He Z, et al. Learning Deep Multi-Level Similarity for Thermal Infrared Object Tracking[J]. IEEE Transactions on Multimedia, 2020.

[71] Li M, Peng L, Chen Y, et al. Mask Sparse Representation Based on Semantic Features for Thermal Infrared Target Tracking[J]. Remote Sensing, 2019, 11(17): 1967.

[72] Zhang X, Ye P, Leung H, et al. Object fusion tracking based on visible and infrared images: A comprehensive review. Information Fusion[J]. 2020, 63: 166-187.

[73] He K, Girshick R and Dollár P. Rethinking ImageNet Pre-training[C]. In Proceedings of IEEE International Conference on Computer Vision, 2019: 4917-4926.

[74] Najibi M, Singh B and Davis L. FA-RPN: Floating Region Proposals for Face Detection[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2019: 7723-7732.

[75] Zhang S, Zhu R, Wang X, et al. Improved Selective Refinement Network for Face Detection[J]. ArXiv, abs/1901.06651, 2019.

[76] Richter S, Hayder Z and Koltun V. Playing for Benchmarks[C]. In Proceedings of IEEE International Conference on Computer Vision, 2017: 2232-2241.

[77] Richter S, Vineet V, Roth S, et al. Playing for Data: Ground Truth from Computer Games[C]. In

Proceedings of European Conference on Computer Vision, 2016: 102-118.

[78] Ros G, Sellart L, Materzynska J, et al. The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2016: 3234-3243.

[79] Doersch C, Gupta A and Efros A. Unsupervised Visual Representation Learning by Context Prediction[C]. In Proceedings of IEEE International Conference on Computer Vision, 2015: 1422-1430.

[80] Larsson G, Maire M, and Shakhnarovich G. Colorization as a Proxy Task for Visual Understanding[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2017: 840-849.

[81] Kim D, Cho D, Yoo D, et al. Learning Image Representations by Completing Damaged Jigsaw Puzzles[C]. In Proceedings of IEEE Winter Conference on Applications of Computer Vision, 2018: 793-802.

[82] He K, Rhemann C, Rother C, et al. A global sampling method for alpha matting[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2011: 2049-2056.

[83] Fang H, Sun J, Wang R, et al. InstaBoost: Boosting Instance Segmentation via Probability Map Guided Copy-Pasting[C]. In Proceedings of IEEE International Conference on Computer Vision, 2019: 682-691.

[84] Bertalmío M, Bertozzi A, Sapiro G. Navier-Stokes, Fluid Dynamics, and Image and Video Inpainting[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2001: 355-362.

[85] Tian Z, Shen C, Chen H, et al. FCOS: Fully Convolutional One-Stage Object Detection[C]. In Proceedings of IEEE International Conference on Computer Vision, 2019: 9626-9635.

[86] Yang Z, Liu S, Hu H, et al. RepPoints: Point Set Representation for Object Detection[C]. In Proceedings of IEEE International Conference on Computer Vision, 2019: 9656-9665.

[87] Zhang X, Wan F, Liu C, et al. FreeAnchor: Learning to Match Anchors for Visual Object Detection[C]. In Proceedings of Neural Information Processing Systems, 2019: 147-155.

[88] Cao Y, Xu J, Lin S, et al. GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond[C]. In Proceedings of IEEE International Conference on Computer Vision Workshops, 2019:1971-1980.

[89] Pang J, Chen K, Shi J, et al. Libra R-CNN: Towards Balanced Learning for Object Detection[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2019: 821-830.

[90] Wu Y, Chen Y, Yuan L, et al. Rethinking Classification and Localization for Object Detection[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2020: 10183-10192.

[91] Cai Z and Vasconcelos N. Cascade R-CNN: Delving Into High Quality Object Detection[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2018: 6154-6162.

[92] Yang X, Yang J, Yan J, et al. SCRDet: Towards More Robust Detection for Small, Cluttered and Rotated Objects[C]. In Proceedings of IEEE International Conference on Computer Vision, 2019: 8231-8240.

[93] Li J, Wang Y, Wang C, et al. DSFD: Dual Shot Face Detector[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2019: 5060-5069.

[94] Lin J. Divergence measures based on the Shannon entropy[J]. IEEE Transactions on Information Theory, 1991, 37(1): 145-151.

[95] Kullback S and Leibler R. On Information and Sufficiency[J]. Annals of Mathematical Statistics, 1951, 22(1): 79-86.

[96] Wang Q, Zhang L, Bertinetto L, et al. Fast Online Object Tracking and Segmentation: A Unifying Approach[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2019: 1328-1338.

[97] Zhang Z and Peng H. Deeper and Wider Siamese Networks for Real-Time Visual Tracking[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2019: 4591-4600.

[98] Chen Z, Zhong B, Li G, et al. Siamese Box Adaptive Network for Visual Tracking[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2020: 6667-6676.

[99] Guo D, Wang J, Cui Y, et al. SiamCAR: Siamese Fully Convolutional Classification and Regression for Visual Tracking[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2020: 6268-6276.

[100] Voigtlaender P, Luiten J, Torr P, et al. Siam R-CNN: Visual Tracking by Re-Detection[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2020: 6577-6587.

[101] Wang G, Luo C, Xiong Z, et al. SPM-Tracker: Series-Parallel Matching for Real-Time Visual

Object Tracking[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2019: 3643-3652.

[102] Danelljan M, Bhat G, Khan F, et al. ATOM: Accurate Tracking by Overlap Maximization[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2019: 4660-4669.

[103] Bhat G, Danelljan M, Gool L, et al. Learning Discriminative Model Prediction for Tracking[C]. In Proceedings of IEEE International Conference on Computer Vision, 2019: 6181-6190.

[104] Danelljan M, Gool L and Timofte R. Probabilistic Regression for Visual Tracking[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2020: 7181-7190.

[105] Dai K, Zhang Y, Wang D, et al. High-Performance Long-Term Tracking With Meta-Updater[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2020: 6297-6306.

[106] Zhang Z, Peng H, Fu J, et al. Ocean: Object-Aware Anchor-Free Tracking[C]. In Proceedings of European Conference on Computer Vision, 2020: 771-787.

[107] Danelljan M, Bhat G, Khan F, et al. ECO: Efficient Convolution Operators for Tracking[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2017: 6931-6939.

[108] Bhat G, Danelljan M, Gool L, et al. Know Your Surroundings: Exploiting Scene Information for Object Tracking[C]. In Proceedings of European Conference on Computer Vision, 2020: 205-221.

[109] Yan B, Zhao H, Wang D, et al. 'Skimming-Perusal' Tracking: A Framework for Real-Time and Robust Long-Term Tracking[C]. In Proceedings of European Conference on Computer Vision, 2019: 2385-2393.

[110] Huang L, Zhao X and Huang K. GlobalTrack: A Simple and Strong Baseline for Long-Term Tracking[C]. In Proceedings of National Conference on Artificial Intelligence, 2020: 11037-11044.

[111] Bolme D, Beveridge J, Draper B, et al. Visual object tracking using adaptive correlation filters[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2010: 2544-2550.

[112] Possegger H, Mauthner T, and Bischof H. In defense of color-based model-free tracking[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2015: 2113-2120.

[113] Henriques J, Caseiro R, Martins P, et al. Exploiting the Circulant Structure of Tracking-by-Detection with Kernels[C]. In Proceedings of European Conference on Computer Vision, 2012: 702-715.

[114] Bertinetto L, Valmadre J, Golodetzet S, et al. Staple: Complementary Learners for Real-Time

Tracking[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2016: 1401-1409.

[115] Mueller M, Smith N and Ghanem B. Context-Aware Correlation Filter Tracking[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2017: 1387-1395.

[116] Wang N, Zhou W, Tian Q, et al. Multi-Cue Correlation Filters for Robust Visual Tracking[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2018: 4844-4853.

[117] Henriques J, Caseiro R, Martins P, et al., High-Speed Tracking with Kernelized Correlation Filters[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2015, 37(3): 583-596.

[118] Danelljan M, Khan F, Felsberg M, et al. Adaptive Color Attributes for Real-Time Visual Tracking[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2014: 1090-1097.

[119] Li F, Tian C, Zuo W, et al. Learning Spatial-Temporal Regularized Correlation Filters for Visual Tracking[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2018: 4904-4913.

[120] Li Y, Zhu J, Hoi S, et al. Robust Estimation of Similarity Transformation for Visual Object Tracking[C]. In Proceedings of National Conference on Artificial Intelligence, 2019: 8666-8673.

[121] Danelljan M, Häger G, Khan F, et al. Accurate Scale Estimation for Robust Visual Tracking[C]. In Proceedings of British Machine Vision Conference, 2014.

[122] Lukezic A, Vojir T, Zajc LA, et al. Discriminativ e Correlation Filter with Channel and Spatial Reliability[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2017: 4847-4856.

[123] Galoogahi H, Fagg A and Lucey S. Learning Background-Aware Correlation Filters for Visual Tracking[C]. In Proceedings of IEEE International Conference on Computer Vision, 2017: 1144-1152.

[124] Tang M, Yu B, Zhang F, et al. High-Speed Tracking With Multi-Kernel Correlation Filters[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition, 2018: 4874-4883.

[125] Ke L, Chang M, Qi H, et al. Multi-Scale Structure-Aware Network for Human Pose Estimation[C]. In Proceedings of European Conference on Computer Vision, 2018: 731-746.

[126] Sárándi I, Linder T, Arras K, et al. Synthetic Occlusion Augmentation with Volumetric Heatmaps for the 2018 ECCV PoseTrack Challenge on 3D Human Pose Estimation[J]. ArXiv, abs/1809.04987, 2018.

[127] Zhang H, Ciss´e M, Y, Dauphin Y, et al. mixup: Beyond Empirical Risk Minimization[C]. In Proceedings of International Conference on Learning Representations, 2018.

[128] Yun S, Han D, Chun S, et al. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features[C]. In Proceedings of IEEE International Conference on Computer Vision, 2019: 6022-6031.

[129] Zhu Z, Wang Q, Li B, et al. Distractor-Aware Siamese Networks for Visual Object Tracking[C]. In Proceedings of European Conference on Computer Vision, 2018: 103-119.

致 谢

曾几何时，心中科研的包袱很重，开启了一场以顶会顶刊为目标的旅行、一场慌慌张张的旅行。论文的发表像是一场整个世界范围内心态的零和游戏，所以每当某个实验室工位存在一个意气风发的顶会顶刊中稿玩家，其他角落往往可能存在一个及以上郁郁寡欢的拒稿玩家，甚至某个阴暗的角落还蹲着若干没有 idea、入场门票都羞于支付的场外吃瓜群众。诚惶诚恐之际，旅途草草以遗憾而收场。过程中渐渐明白，并不是每个研究生都可以手握顶会顶刊毕业，论文的多少与否也不能和科研能力直接画等号。顶会顶刊固然是理想的目标，但是也要循序渐进。与此同时，更为重要的是在研究生期间形成自主思考的模式，形成属于自己的方法论，关注细节却不拘泥于细节。

本文是在焦建彬教授、韩振军副教授、叶齐祥教授和秦飞副教授的悉心指导下完成的。首先感谢第一导师焦建彬教授和第二导师韩振军副教授在我攻读硕士学位期间，从生活、科研等方面无微不至的帮助。正所谓师者，所以传道、授业、解惑也。感谢焦老师对我各方面的支持，严谨的治学态度深深感染着我。感谢韩老师对于科研的热爱程度和高尚的人格极大地激励了我，鞭策着我不断向前。感谢叶老师若干次的论文点评与见解，诲人不倦的教授风范使我受益匪浅。感谢秦老师的言传身教，认真细致的做事风格促使我不断前行。

感谢我的家人，对我的选择给予无条件的支持，是我在砥砺前行中坚强的后盾。感谢实验的师兄师姐师弟师妹们，和我一同前行，怀念我们一起度过的岁月。感谢我的女朋友督促我尽快完成盲审论文。感谢参加开题、中期和毕业答辩的各位指导老师和专家，丰富的阅历和细微的指导对我的科研工作带来了巨大的帮助。

须知少日拏云志，曾许人间第一流。一晃我这模型训练已三载，是时候测试一下性能了。（感谢 Matebook 和 Word 的默契配合，自动重启且无保存，给了我第二轮致谢的机会）

蒋楠

2021 年 5 月

作者简历及攻读学位期间发表的学术论文与研究成果

姓名：蒋楠　　性别：男　　出生日期：1995 年 12 月 15 日　　籍贯：内蒙古

2014 年 8 月——2018 年 6 月，在西安交通大学电子与信息工程学院获得学士学位。

2018 年 9 月——2021 年 6 月，在中国科学院大学微电子学院攻读硕士学位。

攻读硕士学位期间论文发表：（*Equal Contribution）

[1] **Jiang N**, Yu X, Peng X, Gong Y and Han Z. SM+: Refined Scale Match for Tiny Person Detection[C]. In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2021.

[2] **Jiang N**, Zhang Y, Luo D, Liu C, Zhou Y, Han Z. Feature Hourglass Network for Skeleton Detection[C]. In Proceedings of IEEE Computer Vision and Pattern Recognition Workshops, 2019: 1172-1176. (CVPR2019 Geometry shape understanding Workshop Point SkelNetOn Track：2nd)

[3] Li Z*, Zhou Z*, **Jiang N**, Han Z, Xing J, Jiao J. Spatial Preserved Graph Convolution Networks for Person Re-identification[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2020.

[4] Yu X, Gong Y, **Jiang N**, Ye Q, Han Z. Scale Match for Tiny Person Detection[C]. In Proceedings of IEEE Winter Conference on Applications of Computer Vision, 2020: 1246-1254.

[5] Yu X, Han Z, Gong Y, **Jiang N**, et al. The 1st Tiny Object Detection Challenge: Methods and Results[C]. In Proceedings of IEEE European Conference on Computer Vision Workshops, 2020: 315-323.

[6] **Jiang N**, Wang K, Peng X, Yu X, Wang Q, Xing J, Li G, Zhao J, Guo G, Han Z. Anti-UAV: A Large Multi-Modal Benchmark for UAV Tracking[J]. Submitted to IEEE Transactions on Multimedia.

[7] Yu X, **Jiang N**, Gong Y, et al. CoarsePoint: Coarse Point Supervised Object

Localization with Self-Refinement[C]. Submitted to IEEE International Conference on Computer Vision.

[8] Han X*, Yu X*, **Jiang N**, et al. Group Sampling for Unsupervised Person Re-identification[C]. Submitted to IEEE International Conference on Computer Vision.

攻读硕士学位期间专利与软件著作权：

[1] 中国科学院大学. 一种基于尺度匹配的弱小人体目标检测方法：中国，201910918836.2[P]. 2019-09-26.（已授权）

[2] 中国科学院大学. 一种基于无监督深度孪生网络的视频去重方法：中国，202010214485.X[P]. 2020-03-24.（已授权）

[3] 中国科学院大学. 基于精确尺度匹配的弱小人体目标检测方法：中国，202010746942.X[P]. 2020-07-29.（已授权）

[4] 中国科学院大学. 基于多源信息融合的弱小目标检测方法：中国，202010215165.6[P]. 2020-03-24. （已授权）

[5] 中国科学院大学. 基于 FPN 的融合因子的弱小目标检测方法：中国，202010752490.6[P]. 2020-07-30.（已授权）

[6] 韩振军，**蒋楠**，余学辉等. 基于精细尺度匹配的弱小目标检测软件. 软件著作权. 登记号：2020SR1052818.

[7] 韩振军，余学辉，**蒋楠**等. 基于尺度匹配的弱小航拍人体目标检测软件. 软件著作权. 登记号：2019SR1047577.

[8] 韩振军，余学辉，**蒋楠**等. 基于多源信息融合的弱小目标检测软件. 软件著作权. 登记号：2020SR0129500.

[9] 韩振军，宫宇琦，余学辉，**蒋楠**等. 基于融合因子的弱小目标检测软件. 软件著作权. 登记号：2020SR1052810.

攻读硕士学位期间的获奖情况：

[1] 校级三好学生，2019.

[2] 三好学生标兵，2020.

[3] 研究生国家奖学金，2020.

[4] 优秀毕业生，2021.