

基于深度融合与注意力机制的 RGBT 目标跟踪

作者姓名 徐程升

指导教师姓名、职称 高新波 教授

申请学位类别 工学硕士

学校代码 10701
分类号 TP391

学号 23021211705
密级 公开

西安电子科技大学

硕士学位论文

基于深度融合与注意力机制的 RGBT 目标跟踪

作者姓名：徐程升

一级学科：控制科学与工程

二级学科（研究方向）：模式识别与智能系统

学位类别：工学硕士

指导教师姓名、职称：高新波 教授

学 院：电子工程学院

提交日期：2026 年 4 月

RGBT Object Tracking Based on Deep Fusion and Attention Mechanism

A thesis submitted to
XIDIAN UNIVERSITY
in partial fulfillment of the requirements
for the degree of Master
in Control Science and Engineering

By
Chengsheng XU
Supervisor: Xinbo GAO Title: Professor
April 2026

西安电子科技大学 学位论文独创性（或创新性）声明

秉承学校严谨的学风和优良的科学道德，本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果；也不包含为获得西安电子科技大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同事对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文若有不实之处，本人承担一切法律责任。

本人签名：_____ 日 期：_____

西安电子科技大学 关于论文使用授权的说明

本人完全了解西安电子科技大学有关保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权属于西安电子科技大学。学校有权保留并向国家有关部门或机构送交学位论文的复印件和电子版，以学术交流为目的赠送和交换学位论文，允许学位论文被查阅、借阅和复印，将学位论文的全部或部分内容编入有关数据库进行检索和提供相应阅览服务；允许采用影印、缩印或其它复制手段保存学位论文。同时本人保证，结合学位论文研究成果完成的论文、发明专利等成果，署名单位为西安电子科技大学。

本人保证遵守上述规定。

（保密的论文在解密后遵守此规定）

本人签名：_____ 导师签名：_____

日 期：_____ 日 期：_____

摘 要

这是中午摘要。这是中午摘要。这是中午摘要。这是中午摘要。这是中午摘要。
这是中午摘要。这是中午摘要。这是中午摘要。这是中午摘要。这是中午摘要。这是
中午摘要。这是中午摘要。这是中午摘要。这是中午摘要。这是中午摘要。这是中午
摘要。这是中午摘要。这是中午摘要。这是中午摘要。这是中午摘要。这是中午摘要。
这是中午摘要。

这是中午摘要。这是中午摘要。这是中午摘要。这是中午摘要。这是中午摘要。
这是中午摘要。这是中午摘要。这是中午摘要。这是中午摘要。这是中午摘要。这是
中午摘要。这是中午摘要。这是中午摘要。这是中午摘要。这是中午摘要。这是中午
摘要。这是中午摘要。这是中午摘要。这是中午摘要。这是中午摘要。这是中午摘要。
这是中午摘要。这是中午摘要。这是中午摘要。这是中午摘要。这是中午摘要。这是
中午摘要。这是中午摘要。这是中午摘要。这是中午摘要。这是中午摘要。这是中午
摘要。这是中午摘要。这是中午摘要。这是中午摘要。这是中午摘要。这是中午摘要。
这是中午摘要。这是中午摘要。这是中午摘要。这是中午摘要。这是中午摘要。这是
中午摘要。这是中午摘要。

关键词： 三维计算机视觉，单目深度估计，绝对深度估计

ABSTRACT

Visual object tracking is a widely studied topic in computer vision and pattern recognition due to its significant theoretical research value and diverse applications in both civilian and military domains, including video surveillance, autonomous driving, and battlefield situation awareness. Object tracking based solely on visible light often face challenges such as smoke interference and varying levels of illumination, leading to frequent failures. Data from both visible and infrared spectra (RGB/Thermal, RGBT) share consistency and provide complementary information about the target, enabling a dual-mode tracker to enhance the robustness and accuracy of visual object tracking. Nonetheless, current RGBT object tracking algorithms suffer from a deficiency in effective feature selection mechanisms for dual-mode feature fusion and the absence of a decision-level fusion algorithm for both modes. This study leverages two frameworks: convolutional neural network and large-scale vision model, incorporating a deep fusion algorithm and attention mechanism to address the identified issues. The main contributions are summarized as follows.

A novel RGBT target tracking algorithm is introduced in this study to address the issues of inadequate network feature representation and variations in the reliability of visible light and thermal infrared for decision fusion, utilizing adaptive attention feature selection and decision fusion techniques. The study employs an adaptive hybrid attention mechanism that integrates channel, spatial, and positional information to improve the network's feature representation, thereby offering more precise evidence for decision-level fusion. The reliability of two modes is modeled using the Dirichlet distribution, the D-S criterion is employed for decision-level evidence fusion, and an online updated multi-mode branch loss adaptive fusion framework is utilized to reinforce the network's robustness in tracking. Extensive experiments conducted on the open datasets GTOT and RGBT234 demonstrate an accuracy and success rate of 90.9%/75.3% and 77.4%/55.6% correspondingly, providing strong evidence for the efficacy of the developed algorithm.

An RGBT target tracking algorithm is introduced in this study to address the issue of limited interaction between template search images and the dynamic changes of the target, utilizing channel space self-attention and template online updating techniques. Leveraging the tracking methods of template and search images in paired networks, a robust benchmark ex-

perimental algorithm is developed by integrating a Transformer-based large-scale vision. The backbone network equipped with dual-branch embedding layers and weights to enhance the feature interaction between visible light and thermal infrared modes. This study introduces a channel space self-attention mechanism based on the correlation between template and search images to improve the interaction and extract diverse complementary features across the modes. Lastly, the study introduces the template online update module to address the issue of model drift due to target time changes, incorporating online template updating and fractional head design mitigate the drift. Extensive experiments conducted on public datasets GTOT and RGBT234 reveal accuracy and success rates of 93.3%/75.6% and 87.2%/63.8%, validating the effectiveness of the proposed algorithm.

Keywords: 3D Computer Vision, Monocular Depth Estimation, Absolute Depth Estimation

插图索引

表格索引

符号对照表

缩略语对照表

目 录

| | |
|-------------------|-----|
| 摘要..... | I |
| ABSTRACT..... | III |
| 插图索引..... | V |
| 表格索引..... | VII |
| 符号对照表 | IX |
| 缩略语对照表..... | XI |
| 第一章 绪论..... | 1 |
| 1.1 研究意义与背景 | 1 |
| 参考文献..... | 3 |
| 致谢..... | 5 |
| 作者简介 | 7 |

第一章 绪论

1.1 研究意义与背景

深度估计作为三维计算机视觉领域的核心基础问题，旨在从二维图像中重建场景的几何特征，在过去十余年中经历了从手工特征建模到深度学习驱动的模式变迁。随着传感器技术与计算能力的提升，深度估计已成为智能无人系统实现环境感知与定位导航的关键技术：在自动驾驶中，它为障碍物检测与路径规划提供必要的距离信息；在无人机技术中，它是实现自主避障与三维测绘的基础；在工业机器人领域，深度估计则赋予了机械臂精准的抓取与协作能力。鉴于三维空间信息的不可替代性，深度估计在未来空间计算与具身智能（Embodied AI）的发展浪潮中，其战略性地位将得到进一步巩固与提升。

然而，现有的深度估计方法在复杂多变的实际场景下仍表现出一定的局限性，其中泛化性能不足的问题尤为突出。具体而言，模型往往在特定的训练数据集（如 KITTI 或 NYU Depth V2）上表现优异，但在面对未见过的光照变化、极端天气或异质场景时，预测精度通常会出现显著下降。更深层次的挑战在于，深度估计模型往往对相机内参（如焦距、光心位置）具有极强的依赖性。由于单目深度估计本质上是一个病态或不适用（Ill-posed）问题，即单幅二维图像中的每一个像素点都对应着三维空间中无数种可能的深度解释。目前的深度神经网络往往通过“走捷径”的方式，过度拟合训练数据中特定相机的成像几何特性，而非真正理解场景的物理尺度。这种对特定成像设备的过度耦合，使得模型在跨相机、跨平台部署时，难以实现稳健的零样本（Zero-shot）迁移。这种局限性在实际应用中引发了诸多挑战。以无人机自主导航为例，理想的系统应具备跨场景的感知能力，然而现有的深度估计模型往往难以在光室内与室外环境间实现无缝切换。当无人机从开阔的室外环境飞入结构复杂的室内空间时，由于场景分布与光照特性的剧烈波动，泛化性能较差的模型往往会产生严重的深度偏差，导致定位失效或碰撞风险。此外，模型对训练数据的过度拟合进一步削弱了其在极端环境下的鲁棒性。在强光、重雾或暴雨等恶劣天气条件下，图像的对比度和信噪比大幅下降，现有的深度模型由于缺乏对场景本质几何特征的提取能力，容易受到环境噪声的干扰，导致深度预测图出现严重的空洞或畸变。这种在非理想成像条件下的性能塌缩，已成为制约深度估计技术走向全天候、全场景应用的核心障碍。

参考文献

致 谢

作者简介

