

基于 BEiT 相对映射网络的单目深度估计

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 BEiT 相对映射网络的单目深度估计

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。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

图 1.1 无人机	2
-----------------	---

表格索引

表 1.1 基础三线表示例	3
---------------------	---

符号对照表

符号	符号名称
σ	Sigmoid 函数
\cap	交集
\cup	并集
\exp	e 指数
\log	e 对数
\int	积分
\sum	累加
\prod	累积
\odot	逐点相乘

缩略语对照表

缩略语	英文全称	中英文对照
MDE	Metric Depth Estimation	绝对深度估计
MRF	Markov Random Field	马尔可夫随机场

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
1.2 国内外研究现状	2
1.2.1 相对深度估计研究现状	2
1.2.2 绝对深度估计研究现状	3
第二章 第二章	5
第三章 第三章	7
第四章 第四章	9
第五章 第五章	11
参考文献	13
致谢.....	15
作者简介	17

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范式变迁^[1-2]。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；在具身智能领域，深度估计则赋予了智能体感知空间结构并进行物理交互的能力。

从技术演进的维度看，单目深度估计的研究历程主要经历了三个阶段。初期阶段主要依赖手工设计的几何特征与先验假设^[3]，通过概率图模型整合图像的底层信息，但在复杂场景下的建模鲁棒性较差。中期阶段随着卷积神经网络（CNN）的兴起，研究重心转向端到端的监督学习^[4]，通过多尺度网络架构显著提升了像素级的预测精度。现阶段则迈入了以大模型和多任务迁移为特征的新时期，Vision Transformer (ViT) 等架构的应用^[2] 极大增强了模型对全局几何上下文的理解。

然而，现有的深度估计方法在非约束场景下的表现仍面临严峻挑战。由于单目深度估计本质上是一个不适定问题，存在固有的比例模糊性^[1,5]，神经网络往往倾向于通过学习训练集中的统计偏见（如物体位置与深度的相关性）来“走捷径”，而非真正理解场景的物理几何。这种策略导致模型过度拟合了特定数据集（如 KITTI 或 NYU Depth V2）的成像特性，使其对相机内参及拍摄视角具有极强的依赖性。一旦应用于光照剧变、极端天气或异质场景，由于领域鸿沟的存在，模型的预测精度往往会出现断崖式下降，限制了其在跨平台部署时的零样本迁移能力。

针对上述挑战，学术界开始探索一种新的研究范式：利用强泛化性的相对深度信息辅助绝对深度的估计^[6]。相对深度虽然不具备物理单位，但其能通过海量异质数据的预训练，捕捉到稳健的几何拓扑关系与遮挡先验，展现出极佳的场景鲁棒性。

本文认为，融合相对深度的几何先验优势与绝对深度的尺度特性，是实现跨场景稳定感知的关键路径。通过设计一种能够解耦几何结构与物理尺度的预测框架，利用大模型提取的全局几何一致性来约束局部尺度恢复，可显著降低模型对特定相机内参的耦合。这种方法旨在打破单目深度估计在未知场景下的精度瓶颈，为无人系统在复杂、全天候环境下的高精度感知提供新的理论支撑与技术方案。



图 1.1 无人机

1.2 国内外研究现状

1.2.1 相对深度估计研究现状

单目深度估计根据输出表征的不同，通常可分为绝对深度估计与相对深度估计。绝对深度估计旨在恢复具有明确物理尺度的度量距离，但在非约束场景下常面临严峻的尺度模糊性挑战；而相对深度估计则侧重于建模场景内物体的几何序关系，在复杂环境及跨领域迁移中展现出更强的鲁棒性。相对深度特征的建模最早可追溯至 Saxena 等人^[3]的研究，虽然其核心目标是恢复场景的三维结构，但该工作首次大规模利用马尔可夫随机场（Markov Random Field, MRF）来建模像素块间的空间相关性与相对位置关系，为后续相对深度概念的提出奠定了几何建模基础。2016 年，Chen 等人^[7]在《Single-Image Depth Perception in the Wild》中真正将“相对深度估计”确立为一个独立的研究命题。该工作证明了深度信息可以脱离昂贵的传感器真值，通过人类标注的像素对排序先验（Ranking Perception）进行端到端训练，实现了在自然界复杂场景下的深度感知。随后，Ranftl 等人^[6]进一步完善了该范式，正式提出了跨数据集混合训练方案。针对电影、激光雷达、虚拟游戏等异质数据集绝对单位不统一的问题，该研究通过尺度与平移不变的损失函数（Scale-and-shift-invariant Loss），将多元数据统一在相对深度框架下进行联合训练，极大地提升了模型的零样本（Zero-shot）泛化能力。尽管上述工作在几何结构的稳健性方面取得了显著突破，但由于相对深度输出缺乏真实的物理尺度信息（Metric Scale），导致其在自动驾驶、无人机避障等对绝对距离高度敏感的实际任务中应用受限。

1.2.2 绝对深度估计研究现状

绝对深度估计（Metric Depth Estimation）旨在直接建立图像特征与物理距离之间的回归映射。早期研究主要聚焦于如何在受限场景下提升度量精度。Eigen 等人^[8]首次利用多尺度卷积神经网络实现了端到端的深度回归，奠定了深度学习在该领域的基础。随后，为了解决连续值回归收敛困难的问题，Fu 等人^[4]提出了深度序数回归网络（DORN），通过间隔递增离散化策略将回归任务转化为有序分类任务，显著提升了模型在特定数据集上的绝对数值精度。

在真值获取方面，针对激光雷达数据稀疏且昂贵的挑战，Godard 等人^[9]提出了自监督学习范式，利用视频序列间的光度一致性作为约束，极大拓宽了绝对深度估计的应用边界。然而，无论是全监督还是自监督方法，现有的绝对深度估计工作仍面临以下严峻挑战：

首先，模型对相机内参及场景分布存在严重的“尺度耦合”。现有的绝对深度预测框架往往将场景几何结构与物理尺度混合建模，导致模型极易过拟合于特定相机的成像特性。一旦测试环境的焦距、拍摄高度与训练集存在差异，预测结果便会产生剧烈的尺度偏移。

其次，现有工作对相对深度所蕴含的稳健几何先验利用不足。虽然相对深度在建模物体遮挡与空间拓扑方面具有天然优势，但多数绝对深度估计方法仍尝试从头学习像素级的数值映射，而忽视了相对深度作为强力几何约束的潜力。这导致现有模型在面对光照剧变或异质场景时，由于缺乏稳健的几何结构感知，其泛化能力始终无法满足全天候、全场景无人系统的部署需求。

表 1.1 基础三线表示例

方法	参数	准确率	耗时 (s)
Baseline	128	85.2%	12.5
Ours	256	92.4%	18.2
Improved	512	94.1%	25.0

第二章 第二章

第三章 第三章

第四章 第四章

第五章 第五章

参考文献

- [1] EIGEN D, PUHRSCH C, FERGUS R. Depth map prediction from a single image using a multi-scale deep network[J]. Advances in neural information processing systems (NeurIPS), 2014, 27.
- [2] RANFTL R, BOCHKOVSKIY A, KOLTUN V. Vision transformers for dense prediction[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2021: 12173-12183.
- [3] SAXENA A, SUN M, NG A Y. Make3d: learning 3d scene structure from a single still image[J]. IEEE transactions on pattern analysis and machine intelligence, 2008, 31(5): 824-840.
- [4] FU H, GONG M, WANG C, et al. Deep ordinal regression network for monocular depth estimation [C]//Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 2018: 2002-2011.
- [5] KENDALL A, GAL Y. What uncertainties do we need in bayesian deep learning for computer vision? [C]//Advances in neural information processing systems (NeurIPS). 2017.
- [6] RANFTL R, LASINGER K, HAFNER D, et al. Towards robust monocular depth estimation: mixing datasets for zero-shot cross-dataset transfer[J]. IEEE transactions on pattern analysis and machine intelligence, 2020.
- [7] CHEN W, FU Z, YANG D, et al. Single-image depth perception in the wild[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS). 2016: 730-738.
- [8] EIGEN D, PUHRSCH C, FERGUS R. Depth map prediction from a single image using a multi-scale deep network[C]//Advances in neural information processing systems (NeurIPS). 2014.
- [9] GODARD C, MAC AODHA O, FIRMAN M, et al. Digging into self-supervised monocular depth estimation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2019.

致 谢

我他妈写都写不完我还致谢个 der

作者简介

1. 基本情况

徐程升，男，河南郑州人，1999年2月出生，西安电子科技大学电子工程学院控制科学与工程专业2023级硕士研究生。

2. 教育背景

2018.09~2022.06，中原工学院，本科，专业：自动化

2023.09~，西安电子科技大学，硕士研究生，专业：控制科学与工程

3. 攻读硕士学位期间的研究成果

3.1 发表学术论文

[1]Zhong A, Chen S, Wang T, et al. A Mutual-Attention Guided Feature Extraction and Adaptative Decision Fusion Framework for Fine-Grained Dual-Band Radar Target Classification[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2024.

3.2 参与科研项目及获奖

[1] 国家自然科学基金面上项目，复杂环境下小样本高分辨雷达目标识别方法（62173265），2021.01-2025.12

[2] 中电科38所，开放环境下注意力选择技术，2021.06-2022.05

[3] 国防相关项目，红外可见光****技术研究，2021.12-2022.12

[4] 中电科38所，基于元学习****技术，2023.02-2024.03

