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　　学　号 23021211705
分 类 号 TP391　　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。

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

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

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，以 BEiT[5] 为代表的掩码图像建模预训练技术与 Vision Transformer (ViT)[2] 架构的应用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[1,6]，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[7]。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。

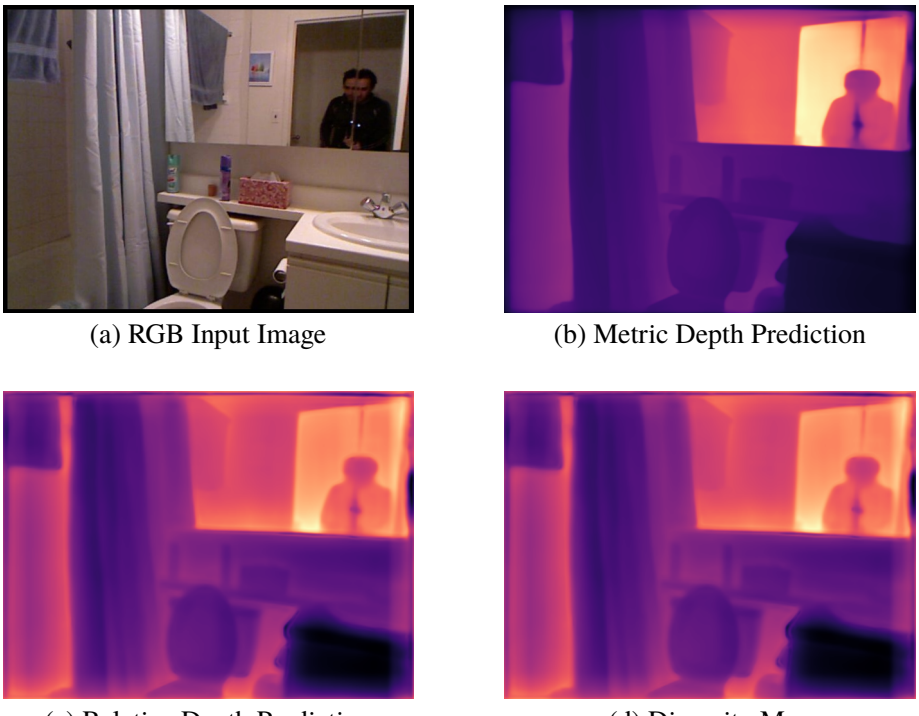
本文认为，融合相对深度的几何先验优势与绝对深度的尺度特性，是实现跨场景稳定感知的关键路径。通过设计一种能够解耦几何结构与物理尺度的预测框架，利用大模型提取的全局几何一致性来约束局部尺度恢复，可显著降低模型对特定相机内参的耦合。这种方法旨在打破单目深度估计在未知场景下的精度瓶颈，为无人系统在复杂、全天候环境下的高精度感知提供新的理论支撑与技术方案。

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。



(a) RGB Input Image　(b) Metric Depth Prediction

(c) Relative Depth Prediction　(d) Disparity Map

图 1.2　Qualitative comparison of depth estimation results. (a) Input RGB image. (b) Absolute metric depth. (c) Relative depth. (d) Disparity map (inverse depth).

相对深度特征的建模最早可追溯至 Saxena 等人[3]的研究，该工作首次利用马尔

可夫随机场（MRF）建模像素块间的空间相关性，奠定了几何建模基础。2016 年，Chen 等人[8]真正将“相对深度估计”确立为一个独立的研究命题，证明了通过人类标注的像素对排序先验（Ranking Perception）即可实现复杂场景下的深度感知。随后，Ranftl 等人[7]正式提出了跨数据集混合训练方案，通过尺度与平移不变损失函数（Scale-and-shift-invariant Loss），极大提升了模型的零样本（Zero-shot）泛化能力。与此同时，Yin 等人[9] 进一步探索了利用多种异构数据集恢复 3D 场景形状的方法，有效增强了模型对复杂几何布局的适应性。

进入 2023 年后，相对深度估计在“视觉基础模型”的推动下取得了突破性进展。以 DINOv2[10] 为代表的大规模自监督预训练模型证明，强力的语义特征与场景的几何拓扑存在高度一致性。在此基础上，Depth Anything[11] 系列工作通过海量无标签数据的判别式训练，实现了在任意“野外”场景下极具鲁棒性的深度结构恢复。这些工作不仅提供了高精度的深度图输出，更重要的是构建了一个包含丰富几何纹理与空间上下文的特征表示空间（Feature Representation Space）。此外，针对深度图细节丢失的问题，Boosting 架构[12] 通过多尺度切片合并策略实现了高分辨率深度细节的恢复；而近期出现的 Marigold[13] 则开创性地利用预训练扩散模型（Diffusion Models）的生成先验，将深度估计转化为图像生成任务，在零样本场景下展现了极高的几何一致性与纹理精度。

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、无人机避障等绝对距离敏感任务中应用受限。因此，如何有效提取并利用相对深度模型中蕴含的强泛化特征块（Feature Blocks），将其作为先验引导来实现高精度的尺度恢复，已成为当前打通“几何结构”与“物理度量”逻辑鸿沟的关键研究方向。

1.2.2 绝对深度估计研究现状

绝对深度估计（Metric Depth Estimation）旨在建立图像特征与真实物理距离之间的回归映射。早期研究如 Eigen 等人[14]利用多尺度 CNN 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），通过离散化策略将回归任务转化为有序分类，显著提升了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的挑战，Godard 等人[15]提出了基于光度一致性约束的自监督学习范式，极大拓宽了应用边界。为克服传统卷积神经网络（CNN）在感受野局限性与长距离依赖建模上的不足，基于 Transformer[16]的架构一度成为单目深度估计领域的主流骨干。Ranftl 等人提出的 DPT (Dense Prediction Transformers)[2] 率先将 Vision Transformer 引入密集预测任务，利用自注意力机制捕捉全局上下文信息，显著提升了深度图的整体结构

一致性。针对深度值的离散化回归难题，AdaBins[17] 创新性地设计了基于 Transformer 的区间划分模块，通过自适应预测深度中心来实现更精细的深度估计。

进入“基础模型”时代以来，单目深度估计正经历从特定场景拟合向零样本泛化（Zero-shot Generalization）的范式转变。以 Depth Anything[11] 和 Metric3D[18] 为代表的工作，通过大规模弱监督预训练与数据蒸馏技术，证明了利用海量无标签数据学习稳健几何表示的可能性。随后，Metric3D-v2[19] 进一步通过万能相机模型（Canonical Camera Space）解决了跨数据集训练中的标签歧义问题。然而，即便在基础模型的支撑下，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深层的“尺度-语义耦合”误区。[20]的研究深刻指出，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，导致模型倾向于通过过拟合训练集中特定相机的成像特性（如焦距、安装高度）以及常见物体的先验尺寸（如车辆、人高）来推断距离，而非基于纯粹的投影几何关系。这种统计过拟合（Statistical Overfitting）导致模型极易受到单目歧义性的干扰：一旦测试环境的焦距或传感器尺寸发生变化，预测结果便会产生剧烈的尺度漂移，无法实现真正的物理一致性。

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。从数学本质上看，绝对深度预测是一个受限于相机内参的“病态问题”，而相对深度由于具备尺度不变性（Scale-invariant），在捕捉物体边缘、空间拓扑及遮挡关系方面表现出极强的泛化潜能。遗憾的是，多数绝对深度研究仍倾向于构建从图像到绝对数值的直接映射（Direct Mapping）。这种“一步到位”的回归范式往往试图让模型同时学习“复杂的几何感知”与“脆弱的绝对回归”，导致模型在面对异质场景时，容易为了拟合数值精度而丧失基本的几何结构约束，产生结构畸变或物体边缘模糊。

目前，如何在高精度度量需求下实现几何结构与绝对尺度的深度解耦与耦合，已成为学术界关注的焦点。尽管 ZoeDepth[21] 等工作尝试引入相对深度特征，但其复杂的端到端训练策略仍难以完全摆脱尺度耦合的影响。鉴于此，探讨如何将具有强泛化力的相对深度作为“几何基座”，并通过轻量化的映射机制恢复物理尺度，对于构建全天候、高可靠的无人系统具有重要的学术意义与应用价值。

第二章　基于相对深度引导的映射模块

2.1 引言

随着具身智能技术的普及，单目深度估计（MDE）已成为自动驾驶与机器人感知的核心环节。按照表征形式，MDE 可分为相对深度估计与绝对深度（Metric Depth）估计，后者因能恢复具备真实物理尺度的像素级距离，成为系统从“视觉感知”迈向“物理交互”的关键桥梁。在避障规划、目标抓取及视觉 SLAM 尺度恢复等任务中，绝对尺度信息提供了不可或缺的几何度量约束。相比激光雷达等昂贵设备，单目绝对深度估计凭借低成本与高灵活性，已成为构建全场景无人系统感知的核心基石。

如前文所述，从单幅图像推算绝对深度本质上是一个典型的“不适定问题”（Ill-posed Problem）。目前主流的单目绝对深度估计多采用端到端的直接回归范式，这种“一步到位”的映射机制使得模型不可避免地陷入尺度信息与几何结构的强耦合困境。在这种机制下，模型往往过度依赖训练集中的语义先验（如物体的平均尺寸）来推断距离，而非理解真正的投影几何关系。这导致模型在面对异质场景或相机内参发生波动时，极易产生剧烈的尺度漂移（Scale Drift），并导致物体边缘与空间布局的一致性受损。

值得注意的是，现有的绝对深度估计工作大多倾向于构建从图像特征到物理数值的直接线性或非线性映射，这种粗放的回归方式忽略了相对深度中蕴含的稳健几何拓扑约束。由于绝对深度真值（Ground Truth）通常由稀疏的激光雷达获取，且覆盖场景有限，模型在拟合有限的数值分布时，容易丧失对物体间细粒度几何序关系的把握。这导致模型在处理复杂场景时，虽然能在统计数值上逼近真值，但其三维空间结构常发生畸变，无法满足高精度导航对环境结构化的严苛要求。

然而，现有的绝对深度估计工作大多倾向于构建从图像到物理数值的直接映射，这种“一步到位”的粗放回归方式往往忽略了相对深度中蕴含的稳健几何拓扑约束。这导致模型在处理复杂场景时，虽然能在数值上逼近真值，但其空间结构常发生畸变，且极易受相机参数波动的干扰。

针对上述挑战，本章提出了一种融合相对深度特征先验的二阶段解构式绝对深度估计方法。该方法的核心逻辑在于将绝对深度估计任务解构为“结构感知”与“尺度适配”两个独立且互补的环节。第一阶段利用具有强泛化力的预训练基础模型提取尺度不变的相对几何特征；第二阶段则通过一个轻量化映射模块，利用多尺度特征块（Feature Blocks）中丰富的上下文信息引导模型完成从相对空间到物理尺度的精准转换。这种解耦设计不仅能最大限度保留基础模型在“野外”环境下的结构泛化力，更

通过特征驱动的映射机制实现了物理尺度的精确对齐，为构建高鲁棒性的绝对深度感知系统提供了新思路。

2.2 模型总体结构设计

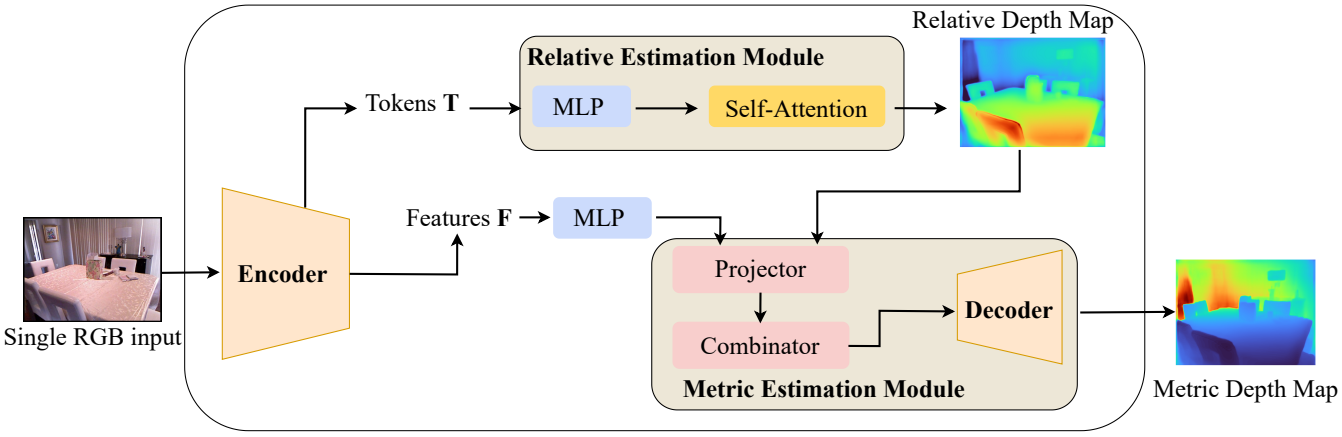


图 2.1　单目深度估计任务解构与特征提取框架

本章所设计的绝对深度估计模型采用了任务解构的二阶段架构，旨在充分利用预训练基础模型的几何结构泛化能力，并通过轻量化的度量映射分支实现物理尺度的精准恢复。模型总体架构如图2.1所示，主要由特征提取编码器（Encoder）、相对深度估计模块（Relative Estimation Module）以及度量深度估计模块（Metric Estimation Module）三部分组成。

2.2.1 任务解构与特征提取

为了从根本上解决单目绝对深度估计中由于投影歧义性导致的尺度与结构强耦合问题，本文将深度恢复过程解构为“结构感知”与“尺度适配”两个解耦的子任务。该设计旨在利用大规模预训练模型（如 DPT-BEiT 或 MiDaS 系列）的几何泛化能力，为绝对尺度的恢复提供稳健的拓扑约束。

如图2.1左侧所示，编码器码器模块是整个深度估计框架的特征提取基石。给定输入 RGB 图像 $I \in \mathbb{R}^{H\times W\times 3}$，编码器的主要任务是将高维像素信息映射为包含语义的深层特征表示。为了同时满足相对深度估计对全局结构的需求，以及度量深度恢复对局部细节的需求，本文采用基于 Vision Transformer (ViT)[22]的架构（如 DPT/BEiT）作为主干网络。与传统的卷积神经网络不同，Transformer 架构天然具备长距离建模能力，能够有效捕捉场景中的上下文依赖关系。编码器在处理过程中生成两类关键的特征表示，分别服务于后续的两个并行分支。

1. 全局语义标记 (Tokens **T**)：如图中上方路径所示，输入图像首先被切分为固定

大小的图块（Patches），并被展平为序列向量。具体而言，对于分辨率为 $H \times W$ 的输入图像 I，我们将切分为 $N = HW/P^2$ 个尺寸为 $P \times P$ 的图块。通过线性投影映射与位置编码相加，初始化输入序列 $\mathbf{z}_0$：

$$\mathbf{z}_0 = [\mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}; \cdots ; \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{pos} \tag{2-1}$$

其中，$\mathbf{x}_p^i$ 表示第 i 个图像块，$\mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D}$ 为线性投影矩阵，$\mathbf{E}_{pos} \in \mathbb{R}^{N \times D}$ 为可学习的位置嵌入，D 为特征维度。经过 Transformer 层的多头自注意力（MSA）机制处理后，输入序列通过 L 层 Transformer 编码器进行处理，每一层的特征更新遵循标准的自注意力（MSA）与多层感知机（MLP）机制：

$$\begin{aligned} \mathbf{z}'_l &= \text{MSA}(\text{LN}(\mathbf{z}_{l-1})) + \mathbf{z}_{l-1} \\ \mathbf{z}_l &= \text{MLP}(\text{LN}(\mathbf{z}'_l)) + \mathbf{z}'_l, \quad l = 1 \dots L \end{aligned} \tag{2-2}$$

其中 $\text{LN}(\cdot)$ 表示层归一化（Layer Normalization）。最终，编码器的最后一层输出即为全局语义标记 $\mathbf{T} = \mathbf{z}_L$。输出序列化的标记集合 $\mathbf{T} = \{t_1, t_2, \dots, t_N\}$。这些 Token 保留了序列形式而未被重组为特征图，它们通过自注意力机制充分交互，编码了图像的全局上下文信息。$\mathbf{T}$ 将直接被送入相对估计模块，用于推断物体间的遮挡关系和整体场景结构。

2. 多尺度空间特征 (Features $\mathbf{F}$)：如图中下方路径所示，除了序列化的 Token，编码器还通过特征重组操作，从不同深度的 Transformer 层中提取出具有空间分辨率的特征图。为了恢复空间结构，我们从编码器的不同层级索引 $\mathcal{I} = \{l_1, l_2, l_3, l_4\}$ 中提取特征序列，并通过重组操作（Reassemble）将其映射回空间域：

$$f_i = \text{Conv}_{1\times1}(\text{Reshape}(\mathbf{z}_{l_i})), \quad i \in \{1, 2, 3, 4\} \tag{2-3}$$

其中，$\text{Reshape}(\cdot)$ 操作将序列 $\mathbb{R}^{N \times D}$ 还原为 $\mathbb{R}^{\frac{H}{P_i} \times \frac{W}{P_i} \times D}$ 的特征张量，$\text{Conv}_{1\times1}$ 用于调整通道维度以适配后续模块。本文将这些分层特征记为 $\mathbf{F} = \{f_1, f_2, f_3, f_4\}$，构成了包含从边缘纹理到抽象语义的特征块组合。$\mathbf{F}$ 是绝对深度估计模块的主要输入之一。在本文提出的方法中，这些多尺度特征将被送入 Xcs-SeedRegressor 和 Xcs-Attractor 模块，利用其中的坐标注意力机制（CoordAtt[23]）进一步挖掘空间位置线索，从而实现对绝对尺度的精确恢复。

2.2.2 相对几何结构感知分支

在单目深度估计任务中，建立稳健的几何拓扑关系是恢复绝对度量维度的前提。本章所设计的相对几何结构感知分支旨在不考虑绝对物理尺寸的情况下，利用大规模预训练模型捕获的深层语义特征，重构场景内物体的相对位置关系。该分支的输入

为共享编码器输出的全局标记向量（Tokens）$\mathbf{T}$，如公式 2-4 所示：

$$D_{rel} = \Psi_{\text{rel}}(\mathbf{T}) \tag{2-4}$$

其中，Ψ_{rel} 表示相对深度估计映射函数。

标记向量 $\mathbf{T}$ 进入相对估计模块后，首先通过多层感知机（MLP）进行维度对齐与特征压缩，随后进入自注意力机制（Self-Attention）层进行长程依赖建模。自注意力机制的引入使得模型能够摆脱局部感受野的限制，通过计算全图范围内的特征相关性，实现对场景宏观布局的深度理解。这种全局建模能力对于识别复杂的物体间拓扑关系、处理大面积平坦区域以及跨越遮挡边界获取一致的几何先验至关重要。

该分支最终输出一张尺度不变的相对深度图 D_{rel}。该图能够精确捕获物体的轮廓细节、场景平面的斜率以及物体间的相互遮挡序关系。尽管其数值不具备物理含义，但 D_{rel} 提供的几何稳定性为下一阶段的尺度适配任务奠定了坚实基础。通过将深度估计任务的第一阶段聚焦于几何拓扑的构建，模型能够显著增强在非约束场景下的泛化能力，为后续实现精确的“结构-度量”映射提供可靠的几何度量锚点。

2.2.3 绝对深度对齐分支

度量深度估计模块（第二阶段）是整个感知框架中实现物理尺度恢复的核心组件。本章并未采用传统的直接数值回归范式，而是引入了一种特征驱动的离散化期望回归机制。其核心逻辑在于将第一阶段捕获的稳健几何结构作为结构锚点，结合编码器提取的多尺度空间特征，实现从相对几何序关系向物理度量空间的精准映射。

在处理流程上，该分支首先接收来自共享编码器的多尺度空间特征 F。通过级联的投影器（Projector）对特征进行非线性变换与空间维度对齐，随后利用吸引器（Attractor）机制，根据特征块中蕴含的语义上下文对深度分桶（Bins）的质心分布进行自适应调整。这种机制允许模型根据场景复杂度动态优化深度的搜索空间，从而在非约束环境下获得更细粒度的尺度表现。

为了最大化保留物体的边缘细节并抑制尺度漂移，本章引入了结构注入与融合机制。第一阶段生成的相对深度图 D_{rel} 经过双线性插值上采样后，与编码器末端的高维空间特征进行维度拼接（Concatenation），构建出增强型的几何-尺度联合表征。如公式 2-5 所示：

$$X_{fusion} = [F_{last} \oplus \text{up}(D_{rel})] \tag{2-5}$$

其中，$\oplus$ 表示通道维度的拼接操作。

在最终的预测环节，模型利用条件对数二项式分布（Conditional Log Binomial）计算每个像素在各深度分桶上的概率响应。最终的绝对度量深度 D_{met} 是通过对所有

分桶质心 B 进行概率加权求和所得的离散期望值：

$$D_{met} = \sum_{i=1}^{N} P_i \cdot B_i \tag{2-6}$$

相较于单一数值回归，这种基于概率分布的建模方式能够有效缓解单目估计中的尺度歧义性，并在优化全局数值精度的同时，确保物体边缘与空间拓扑的一致性。

2.3 实验结果与分析

为全面验证本章所提方法的有效性与稳健性，本节在单目绝对深度估计领域两个极具代表性且富有挑战性的基准数据集上开展了详尽的实验。首先，对室内场景数据集 NYU-Depth v2[24] 与室外自动驾驶数据集 KITTI[25] 的基本概况进行简要说明。其次，明确绝对深度估计领域通用的性能评价指标，并详细介绍实验的软硬件环境与超参数设置。最后，通过定性与定量的对比实验验证本方法相较于当前主流算法的领先性，并结合消融实验深入探讨各核心模块对模型性能的贡献，旨在阐明本方法在复杂环境下恢复物理度量深度的优越性。

2.3.1 数据集

NYU-Depth v2 数据集是由纽约大学 Silberman 等研究人员通过 Microsoft Kinect 设备采集的大规模室内场景数据集，主要针对的是室内场景理解、语义分割以及深度估计相关问题。该数据集中共包含 1,449 张经过人工标注并对齐的 RGB 与深度图像（RGB-D）对，场景涵盖了卧室、厨房、办公室、客厅等多种复杂的室内环境。数据集类别包括家具、电器、墙壁等共计 894 类物体。由于其采用真实场景数据，因此包含了室内环境中常见的光照变化、物体遮挡以及复杂的空间布局，为单目绝对深度估计的研究提供了重要的数据支撑。

NYU-Depth v2 数据集除了提供高精度的深度标注信息之外，还根据语义一致性提供了密集的像素级分割标签，为计算机视觉系统在室内环境下的开发提供了全方位的数据支持。该数据集主要适合用于单目深度估计的训练与评估，也可用于室内语义分割、目标检测等任务。同时，由于数据中自然包含了不同室内布局下的空间几何信息，其已成为度量绝对深度估计领域最常用的基准测试集之一。

本文关注模型在室内复杂布局下的尺度恢复能力，因此采用的是 NYU-Depth v2 的官方标准划分方式。实验选取了由原始视频序列预处理得到的约 50,000 张图像对作为训练数据，以确保模型能够充分学习室内场景的特征分布；并严格采用官方预定义的 654 张测试图像作为最终的性能评估集。为了准确测试模型的跨场景泛化性能与尺度恢复精度，实验过程中对深度真值进行了有效的范围截断与对齐处理，确保模

型能够在多样化的室内闭环环境下保持较高的估计效果。

（1）KITTI 数据集

KITTI（Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago）数据集是目前国际上最受认可的室外自动驾驶场景计算机视觉算法评测基准之一。该数据集由德国卡尔斯鲁厄理工学院和芝加哥丰田技术研究院联合创建，利用配备了彩色/灰度摄像机、高精度旋转式激光扫描仪（LiDAR）以及 GPS/IMU 导航系统的车载平台，在城市街区、乡村道路和高速公路等多样化的真实道路环境下采集而成。在单目深度估计任务中，KITTI 数据集提供的激光雷达点云数据经投影对齐后，可作为度量深度估计的稀疏真值参考，其覆盖范围通常可达 80 米。

在实验分析中，本文遵循单目深度估计领域的通行规范，采用了广泛认可的 Eigen 划分方式（Eigen Split）[14]。该划分逻辑由 Eigen 等人提出，通过对原始序列进行筛选，有效去除了冗余或传感器失效的帧，最终形成了包含 23,488 对原始图像序列的训练集，以及 697 张具有高质量深度标注的测试图像。由于激光雷达采集的原始点云具有稀疏性，本文在预处理阶段采用了反距离加权插值或官方提供的深度填充算法进行优化，以获得更为致密的监督信号。

针对室外自动驾驶环境下相机内参多样且光照变化剧烈的特点，KITTI 数据集为验证本方法在高速移动场景下的尺度对齐能力提供了绝佳的平台。本文在实验中特别关注了模型对远景（如建筑、植被）与近景（如车辆、行人）的结构刻画能力。通过在该数据集上的测试，能够有效评估本章提出的二阶段解构映射方法在应对室外大尺度空间变化时的鲁棒性与泛化性能。

2.3.2 评价指标

为了客观、全面地评估本文所提模型在单目绝对深度估计任务上的性能，本文在实验部分采用了深度估计领域公认的通用评价指标体系进行定量分析。评价指标主要分为误差指标（Error Metrics）与准确度指标（Accuracy Metrics）两类。令 d_i 表示第 i 个像素的深度真实值（Ground Truth），$\hat{d}_i$ 表示对应的模型预测值，N 为参与评估的像素总数。

（1）误差指标分析

误差指标旨在衡量模型预测结果与真实值之间的数值偏差。首先，平均相对误差（Abs Rel）反映了预测深度相对于真实深度的平均偏移比例，其计算公式如式（2-7）所示：

$$\text{Abs Rel} = \frac{1}{N}\sum_{i=1}^{N}\frac{|d_i - \hat{d}_i|}{d_i} \tag{2-7}$$

该指标对全量程范围内的误差具有较好的均衡评价能力。其次，为了评估模型预测的稳定性，本文引入了均方根误差（RMSE）。由于 RMSE 采用了平方运算，其对预测结果中的异常值和较大偏差极度敏感，计算公式如式（2-8）所示：

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(d_i - \hat{d}_i)^2} \tag{2-8}$$

此外，对数平均误差（$\log_{10}$）通过将深度值映射到对数空间进行计算，能够有效缩小远近景数值量级的差异，从而更客观地评价模型在不同距离下的线性一致性：

$$\log_{10} = \frac{1}{N}\sum_{i=1}^{N}|\log_{10} d_i - \log_{10} \hat{d}_i| \tag{2-9}$$

对于上述三项误差指标而言，其数值越小，通常代表模型的绝对深度还原精度越高。

（2）准确度指标分析

准确度指标通过统计满足特定阈值条件的像素占比来衡量预测图的贴合程度。本文采用阈值准确度 δ_n 作为评价标准，其定义如式（2-10）所示：

$$\delta_n = \text{percentage of } d_i \text{ s.t. } \max\left(\frac{d_i}{\hat{d}_i}, \frac{\hat{d}_i}{d_i}\right) < 1.25^n \tag{2-10}$$

其中，$n \in \{1, 2, 3\}$ 分别对应三个不同严苛程度的阈值。在实验结果分析中，δ_1（即阈值小于 1.25）是衡量深度估计精度的最关键指标，反映了模型实现高精度像素级还原的能力；而 δ_2 和 δ_3 则代表了在中等及较宽容许误差下的准确率。对于此类准确度指标，数值越大代表模型的性能表现越优异。

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] BAO H, DONG L, PIAO S, et al. BEiT: BERT pre-training of image transformers[C]//International Conference on Learning Representations (ICLR). 2022.

[6] 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.

[7] 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.

[8] 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.

[9] YIN W, LIU Y, SHEN C, et al. Learning to recover 3D scene shape from a single image[J]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[10] OQUAB M, DARCET T, MOUTAKANNI T, et al. DINOv2: learning robust visual features without supervision[A]. 2023.

[11] YANG L, KANG B, HUANG Z, et al. Depth anything: unleashing the power of large-scale unlabeled data[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024.

[12] MIANGOLEH S M, SZELISKI R, AKBARI H, et al. Boosting monocular depth estimation models to high resolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021.

[13] KE B, OBUKHOV A, HUANG S, et al. Repurposing diffusion-based image generators for monocular depth estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern

Recognition (CVPR). 2024.

[14] 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.

[15] 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.

[16] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems (NeurIPS): Vol. 30. 2017.

[17] BHAT S F, ALHASHIM I, WONKA P. AdaBins: depth estimation using adaptive bins[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 4009-4018.

[18] YIN W, ZHANG C, CHEN H, et al. Metric3D: towards zero-shot metric depth prediction via large-scale multi-dataset training[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2023.

[19] YIN W, ZHANG C, CHEN H, et al. Metric3D v2: a versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation[A]. 2024.

[20] PICCINELLI L, YANG Y H, SAKARIDIS C, et al. UniDepth: isolating depth from camera parameters in one-stage anticipation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024.

[21] BHAT S F, BIRKL R, BIENECK D, et al. ZoeDepth: zero-shot transfer by combining relative and metric depth[A]. 2023.

[22] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[C]//International Conference on Learning Representations (ICLR). 2021.

[23] HOU Q, ZHOU D, FENG J. Coordinate attention for efficient mobile network design[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 13713-13722.

[24] SILBERMAN N, HOIEM D, KOHLI P, et al. Indoor segmentation and support estimation from rgbd images[C]//European Conference on Computer Vision (ECCV). Springer, 2012: 607-620.

[25] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? the kitti vision benchmark suite[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2012: 3354-3361.

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