西安交通大学

博士学位论文

多光谱融合智能光电处理算法与系统设计

学位申请人：陈炜煌

指导教师：孙宏滨教授

学科名称：控制科学与工程

2025 年 12 月

Intelligent Electro-Optical Processing Algorithm and Systems Design based on Multispectral Fusion

A dissertation submitted to

Xi'an Jiaotong University

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

By

Weihuang Chen

Supervisor: Prof. Hongbin Sun

Control Science and Technology

December 2025

博士学位论文答辩委员会

多光谱融合智能光电处理算法与系统设计

答辩人：陈炜煌

答辩委员会委员：

西安交通大学教授：辛景民（注：主席）

西北工业大学教授：王鹏

西安电子科技大学教授：董伟生

西安交通大学教授：魏平

西安交通大学教授：杜少毅

答辩时间：2023 年 05 月 14 日

答辩地点：西安交通大学科学馆 324

摘 要

自动驾驶在节约驾驶成本、提高交通效率、减少环境污染等方面拥有巨大优势，成为了学术界和工业界的热门研究课题。为了实现安全可靠、稳定高效的驾驶行为，自动驾驶车辆需要精准地预测出周围环境中交通参与者的未来行为轨迹，并规划出自身无碰撞且运动学可行的短时运动轨迹。传统轨迹预测方法无法保证长期预测的精度，严重依赖启发式设计的传统运动规划方法也无法保证其泛化性能。近年来，基于数据驱动的深度学习方法得到了快速发展，为完成预测规划任务带来了新思路。从数据输入输出的角度考虑，预测和规划都是对交通参与者的历史特征进行建模后输出未来轨迹。因此，这两项具有共性的任务均可以采用具有强大特征拟合能力的深度学习方法来完成。然而，此类方法仍然存在交通参与者异质性处理能力差，缺少概率性预测结果以及无法保证轨迹平滑性等问题，使得自动驾驶的安全性受到威胁，阻碍了自动驾驶技术进一步的发展。

本文聚焦于利用 Transformer 网络解决上述核心难点问题：1）如何构造更精准、更快速的实用化轨迹预测网络模型；2）如何在保证完成自动驾驶任务的前提下促使运动规划方法尽可能地减少交通违规行为。主要研究工作如下。

1. 提出了一种基于时空 Transformer 网络的单模态轨迹预测网络模型，弥补了之前方法只能有效预测同质交通参与者的缺陷，提高了密集交通环境下时空交互建模能力。针对之前方法对时间序列数据进行串行处理造成的记忆能力弱以及空间邻域范围设置不合理等问题，该方法采用 Transformer 网络并构建了全感知域的时空图模型。整个网络包括时空 Transformer 编码器、时间 Transformer 编码器和时间 Transformer 解码器三个部分。时空 Transformer 编码器能够对时空图特征按照不同维度交替提取，从而充分融合时空信息。经过时间 Transformer 编码器对于时间信息的进一步处理后，时间 Transformer 解码器生成了关于异质交通参与者的单模态轨迹。在自动驾驶轨迹预测公开数据集上的实验结果表明，该方法比当时最好的方法在主要性能指标上提高了至少 7.2%。

2. 提出了一种基于概率性候选轨迹网络的多模态轨迹预测网络模型，在加快模型推理速度的同时，提高了多模态轨迹预测的精度。针对当前多模态轨迹预测方法无法提供概率性预测结果的问题，该方法设计了一种既能生成目标点引导信息，又能提供概率性结果的三阶段轨迹预测过程。首先，该方法利用无监督学习自动获取交通参与者的潜在意图集合，并应用分类网络筛选出符合当前交通参与者运动趋势的概率性目标点集合。然后，通过 Transformer 网络生成中间位置锚点。最后，使用连续曲线光滑连接当前位置、锚点和目标点，形成表达能力更强的概率性候选轨迹集。多个公开轨迹预测数据集的实验结果验证了该方法在提供高性能、高效率的概率性预测结果的同时，能够确保概率较高的预测结果更符合交通参与者的下一步行为。

3. 提出了一种基于安全轨迹树网络的运动规划网络模型，减少了之前基于学习的运动规划方法在完成自动驾驶任务时出现的大量交通违规行为。针对之前方法因不能满足相关运动约束而造成的违规问题，该方法提出了一种具有曲率连续性和运动学可行性的轨迹树。该轨迹树既能够用于运动规划主任务，也能够作用于共性的轨迹预测辅助任务，从而帮助模型通过学习预测规划间的交互提升性能。针对高维栅格化特征输入可解释性差、计算效率低的问题，该方法采用包含交通参与者和局部任务路线的离散化输入表达方式，增加了模型的可解释性。该方法还利用 Transformer 主干网络精准提取不同输入之间的空间交互信息。针对自动驾驶汽车在复杂场景中保持长期静止不动的问题，该方法在训练过程中引入了焦点损失函数，鼓励自动驾驶车辆安全高效地完成导航任务。多个自动驾驶闭环测试基准的实验结果表明，该方法不仅在自动驾驶任务完成度和违规得分方面比之前最好的方法分别提高了 39.2% 和 10.6%，而且推理速度加快了 1.5 倍。

综上所述，本文所提出的单模态轨迹预测、多模态轨迹预测和运动规划方法获得了高性能的表现，具有精度高、速度快和违规驾驶行为少的优势，为保证自动驾驶安全性发挥了重要作用。

关 键 词： 自动驾驶；轨迹预测；运动规划；自注意力模型

论文类型： 应用研究

ABSTRACT

Autonomous driving is an innovative and advanced research field in academia and industry, with potential to reduce road fatalities, improve traffic efficiency, and decrease environmental pollution. To achieve safe, reliable, stable, and efficient driving behavior, autonomous vehicles need to accurately predict the future trajectories of surrounding traffic participants and plan collision-free, kinematically feasible short-term motion trajectories. Traditional trajectory prediction methods often lack accuracy in long-term predictions, while motion planning methods based on heuristic design may lack generalization performance. In recent years, rapid advancements in data-driven deep learning methods have revolutionized prediction and planning tasks. Deep learning methods offer powerful feature fitting capabilities that enable accurate modeling of historical traffic patterns and output of future trajectories. Consequently, both prediction and planning tasks can be achieved using deep learning techniques. Despite these remarkable benefits, these methods still face various challenges, such as poor ability to deal with traffic participant heterogeneity, lack of probabilistic prediction results, and inability to guarantee trajectory smoothness. These issues pose significant safety concerns for autonomous driving and impede the further advancement of this technology.

This dissertation proposes to leverage Transformer network to address these core difficulties. The objectives of this study are twofold: 1) improving the accuracy and inference speed of practical trajectory prediction models, 2) enhancing motion planning methods to minimize traffic violations while guaranteeing the completion of autonomous driving tasks. The main contributions of this research are as follows.

1. This dissertation proposes a new Spatio-Temporal Transformer Network for unimodal trajectory prediction, which addresses the limitations of previous homogeneous prediction methods and improves spatio-temporal interactive modeling capabilities. To address the problems of weak memory ability and unreasonable setting of spatial neighborhood range caused by the serial processing of time series data in the previous method, we adopt Transformer network and constructs a spatio-temporal graph of the whole perceptual domain. The network consists of three parts, i.e. spatio-temporal Transformer encoder, temporal Transformer encoder, and temporal Transformer decoder. The first Transformer encoder extracts spatio-temporal features by alternating between different dimensions to fully integrate spatio-temporal information. The second Transformer encoder further processes temporal information, and the temporal Transformer decoder generates unimodal trajectories for heterogeneous traffic participants. Experimental results demonstrate that the proposed method enhances key performance metrics by at least 7.2% over state-of-the-art methods.

2. This dissertation proposes a new Probabilistic Proposal Network for multimodal trajectory prediction which not only enhances the prediction accuracy of multimodal trajectory prediction, but also accelerates the inference speed. To address the problem that previous multimodal trajectory prediction methods cannot provide probabilistic prediction results, we devise a three-stage trajectory prediction process that generates target point guidance information and provides probabilistic outcomes. Firstly, the proposed method employs unsupervised learning to automatically obtain the potential intention set of traffic participants and applies a classification network to filter out a set of probabilistic target points that comply with the current movement trend of traffic participants. Next, Transformer network generates intermediate position anchors. Finally, a continuous curve is used to smoothly link the current position, anchors, and target point, producing a more expressive set of probabilistic trajectory candidates. Experimental results demonstrate that the proposed method yields high-performance and high-efficiency probabilistic prediction results while ensuring that the prediction results with higher probability align more closely with the next behavior of traffic participants.

3. This dissertation proposes a new safe Trajectory Tree Network for motion planning, which can effectively reduce traffic violations while completing autonomous driving tasks. The key component of TTNet is a predefined trajectory tree that conforms to vehicle dynamics constraints and explicitly reflects different intentions. This tree is used for both the main planning task and an auxiliary trajectory prediction task. To enhance interpretability, we introduce input expressions typically used in traditional planning algorithms into our integrated framework. Additionally, to promote safe and efficient navigation, we incorporate a focal loss during training and employ a Transformer-based backbone network to accurately capture spatial interactions not only among the ego vehicle and its surroundings, but also among dynamic agents and the reference line. Experimental results demonstrate that the proposed method significantly improves task completion and violation scores by 39.2% and 10.6%, respectively, compared to SOTA methods while accelerating the inference speed by 1.5 times.

In summary, our proposed methods achieve outstanding performance for unimodal trajectory prediction, multimodal trajectory prediction and motion planning, with the advantages of high precision, high speed and less driving violations, thus playing crucial roles in ensuring the safety of autonomous driving.

KEY WORDS: Autonomous Driving; Trajectory Prediction; Motion Planning; Transformer

TYPE OF DISSERTATION: Application Research

目 录

CONTENTS

1 基于多特征聚焦与跨阶段 Transformer 的红外小目标检测网络

在机载光电系统中，红外成像传感器是实现全天时、全天候环境感知的核心组件。机载平台下的红外小目标检测面临三重严峻挑战：首先，红外图像本身存在分辨率低、缺乏纹理细节、对比度差的固有限制，其次，从空中俯视的复杂动态背景对小目标检测造成干扰，导致算法虚警率高，最后，无人机载荷对功耗、重量和计算资源的严格限制，要求算法必须在极高的实时性与足够的检测精度之间取得平衡。当前多数基于像素级分割的检测网络，虽在静态基准测试中表现良好，但其庞大的计算量和对高分辨率特征图的需求，使其难以直接部署于机载嵌入式平台进行实时处理。为应对这些挑战，本章提出一种专为无人机平台设计的高效红外小目标检测网络 MFF-DCNet。该网络通过协同优化特征提取与融合机制，实现精度与速度的兼顾，核心创新在于：基于深度分离卷积的跨阶段 Transformer（Depth-wise Cross-stage Transformer，DCFormer）和多特征聚焦（Multi-Feature Focus，MFF）颈部结构。DCFormer 模块通过深度可分离卷积与跨阶段特征融合的结合，在显著降低计算开销的同时，有效增强了主干网络对多尺度上下文信息的建模能力。优化特征提取过程，提升了模型在复杂场景下对小目标的特征判别能力，并且为在边缘设备上的实时部署提供了可能。重新设计的 MFF 颈部结构通过构建新颖的特征聚合机制，增强了跨尺度特征的整合能力，使模型能够更好地融合不同层级的语义信息和空间细节。这种设计显著提升了模型对多尺度目标的检测性能，特别是在复杂背景下对微小红外目标的精准识别能力。

1.1 引言

红外探测系统凭借其被动成像，抗干扰能力强及全天时工作的独特优势，显著提升了无人机的自主感知能力。然而，红外图像普遍存在空间分辨率低、缺乏色彩与丰富纹理细节的局限，导致为高分辨率可见光图像设计的通用深度网络难以直接迁移并提取有效的判别性特征，在典型的无人机应用场景中，地面目标在红外图像中仅占据极少的像素，表现为信噪比极低的弱小目标。同时，复杂的地物背景会引入大量与目标热辐射特征相似的杂波，使得在低信噪比条件下实现精准的“目标-背景”分类变得困难。传统的神经网络，尤其是计算密集的 Transformer 架构，其高复杂度在算力受限的边缘设备上难以实现实时性能，构成了实际部署的显著瓶颈。因此，在嵌入式设备上实现高精度、高实时性的红外小目标检测是一个亟待解决且具有重要实际价值的研究课题。

针对小目标检测问题，研究者们提出了多种解决方案，如数据增强、背景建模以及聚焦检测等。然而，当这些方法直接应用于红外图像时，其性能往往会因红外成像的物理特性而显著衰减。数据增强策略在可见光图像中能有效增加小目标样本，但在红外图像中可能破坏目标与背景之间的热对比度关系，导致性能提升不稳定且泛化能力有限。背景建模利用目标周围区域提供辅助信息，然而在杂波丰富的红外背景中，过度依

赖上下文极易引入干扰噪声，反而模糊了目标本身的特征。聚焦检测类的方法计算成本高，且将大量计算浪费在对广阔背景区域的处理上。

现有的红外小目标检测方法可分为两大技术路线：基于分割的方法与基于检测的方法。基于分割的方法在天空、海面等纯净背景下表现良好。然而，当应用于具有复杂背景的无人机航拍图像时，这类方法的误报率显著升高。红外传感器的远距离成像导致小目标信噪比较低，且目标越小，其像素表现越模糊、不确定性越高，在复杂环境中获取精确的像素级标注也变得更加困难。此外，分割网络通常采用的编码器解码器结构会引入巨大的计算开销，无法在资源受限的嵌入式平台上达到实时处理的要求。基于检测的方法专注于直接识别与定位小目标。其中两阶段模型虽能达到较高精度，但常受计算复杂度高的困扰，而单阶段模型则因其高效性在嵌入式系统中日益流行。为提升上下文建模能力，Transformer 架构被引入小目标检测网络，以解决捕获长程依赖关系的挑战。其自注意力机制通过计算特征的相关性，使网络能够聚焦于目标区域并捕获更广泛的上下文信息。后续工作中对编码器做出了进一步优化，例如在编码器中引入局部感知块的 Local Perception Swin Transformer (LPPSW)，以及融合了全局-局部特征交错模块的双网络结构（Dual network structure with Interweaved Global-Local，DIAG）。这些方法均致力于优化特征提取，以应对复杂航拍图像中的小目标检测难题。虽然这些改进提升了检测精度，但通过复杂自注意力机制带来的性能提升，往往以计算开销的大幅增加为代价。因此，基于 Transformer 的方法在部署于红外小目标检测系统时，特别是在计算资源有限的边缘计算场景中，仍面临显著局限。

针对上述问题，本章提出 MFF-DCNet，一种基于 YOLOv11 的高效红外小目标检测网络。通过两项关键创新 MFF-DCNet 实现了业界领先的效率精度平衡：深度可分离跨阶段 Transformer 模块（Depth-wise Cross-stage Former，DCFormer）与多特征聚焦颈部结构（Multi-Feature Focus，MFF）。DCFormer 通过深度可分离卷积对标准 Transformer 编码器进行改进，在降低计算复杂度的同时提升了特征提取能力。MFF 颈部结构重新定义了原有框架的颈部结构，特征聚合机制跨尺度选择并融合差异化特征，有效抑制了冗余信息。

本文的主要贡献总结如下：

- 重新设计了整个颈部结构，提出了多特征聚焦模块 MFF。这是一种新颖的特征聚合结构，能有效增强不同尺度间特征信息的融合，从而提升模型对多尺度目标（尤其是复杂环境中的红外小目标）的检测能力。
- 通过引入 DCFormer 模块增强了主干网络的特征提取能力。该先进的增强模块集成了深度可分离的空间特征提取与跨阶段特征融合，在降低计算成本的同时优化了多尺度上下文建模，提升了复杂场景下的红外小目标检测精度。
- 在 HIT-UAV 和 DroneVehicle 数据集上进行了充分的实验。MFF-DCNet 在检测精度（AP_{5095} 达到 57.4%，较同类先进方法提升 5.8%）与处理效率（帧率提升 10%）上均取得显著进步。同时，该网络在 NVIDIA Jetson Orin NX 边缘计算模块上达

到了 39.6 FPS 的稳定实时处理能力，验证了其满足实际机载任务严苛的实时性、可靠性与低功耗要求，具备直接的工程应用价值。

1.2 相关工作

1.2.1 面向嵌入式平台的红外小目标检测

传统基于建模的方法，如稀疏分解与背景建模，均建立在一个先验假设之上，即小目标可以从结构化背景中被有效分离。与基于深度学习的方法相比，这类方法在面对典型的城市无人机复杂运行环境时，性能严重下降。基于深度学习的红外小目标检测方法主要分为分割方法和检测方法两条技术路线。分割方法将该任务建模为一个正负样本极不平衡的二值语义分割问题，代表性方法可归类为超分辨率、多尺度表征、上下文信息和尺度感知训练等。近期工作 LRRNet 试图将深度学习与传统的稀疏分解和背景建模相结合。然而，这些分割网络的推理速度通常很慢，即便在标准的消费级 GPU 上平均帧率也低于 10FPS，嵌入式平台的算力通常不足普通桌面级 GPU 的十分之一，难以支持高复杂度的分割网络实时运行。

嵌入式系统在计算能力、内存和能耗方面面临严格限制。尽管存在众多适用于消费级 GPU 的高性能算法，但它们往往难以直接部署于无人机等边缘设备上。在边缘设备上部署高性能红外检测算法的主要挑战，在于高计算需求与硬件约束之间的冲突，这推动着研究向轻量化和专用化模型发展。单阶段检测模型因其高效性在嵌入式系统中日益流行，例如 MobileNet、ShuffleNet 和 GhostNet。MobileNet 通过使用深度可分离卷积减少了参数量，ShuffleNet 使用分组卷积将输入通道划分为更小的组，降低了计算复杂度和参数量，GhostNet 引入 Ghost 模块，通过先使用较少卷积核生成主要特征图，再生成额外的特征图，实现了参数量和计算量的大幅降低。为提高小目标检测精度，这些轻量级主干网络常与多尺度特征学习或超分辨率技术结合使用。例如，SuperYOLO 采用对称紧凑的多模态融合技术整合多种数据模态（RGB 与红外），并融入超分辨率学习以获取高分辨率特征表示，YOLC 利用针对聚类区域的局部尺度模块，但在处理稀疏分布的目标时效果欠佳。

YOLO 系列模型在速度与准确性之间取得了出色平衡，是嵌入式平台的理想选择。近期，基于 DETR 框架的实时检测器（如 RT-DETR）被提出。然而，在没有对应领域成熟预训练模型的情况下，DETR 极难应用于新领域，因此目前应用最广泛的实时目标检测器仍是 YOLO 系列。本章提出的 MFF-DCNet 基于 YOLOv11 构建，以架构效率为核心设计原则，DCFormer 模块通过深度可分离设计减少参数，MFF 模块在不依赖昂贵计算成本的编解码器结构的前提下增强了特征判别力。与基于注意力的 Transformer 方法相比，MFF-DCNet 实现了更少的参数量和计算负载，使其更适用于资源受限的嵌入式设备。

1.2.2 多尺度特征学习方法

深度卷积神经网络会生成具有不同空间分辨率的层级化特征图。其中，低层特征蕴含更丰富的细节和定位信息，而高层特征包含更强的语义信息。对于红外小目标检测而言，随着网络深度的增加，小目标的特征表示在最终的特征图中会逐渐减弱。由于红外图像固有的特性（如分辨率低、纹理信息弱），这一问题在红外小目标检测中被进一步放大。为此，一种有效的解决方案是多尺度特征学习，通过整合不同深度的特征来增强对小目标的表征能力。

特征金字塔网络（Feature pyramid Network，FPN）通过构建自顶向下的路径与横向连接，在不同尺寸的特征图上进行预测，并根据目标尺寸将不同尺度的目标分配到对应的金字塔层级。这一多尺度预测的方式被广泛集成于各类目标检测网络中。受到 FPN 的启发，PANet 通过引入双向路径来丰富特征层次，利用精确的定位信号强化深层特征，以更直观方式实现多尺度特征融合的优化。AugFPN 则提出残差特征增强与一致性监督，以缩小不同尺度特征间的语义差距。双向特征金字塔网络（Bi-directional Feature Pyramid Network，BiFPN）引入了双向连接，允许信息在网络中同时进行自顶向下和自底向上的流动，确保了对小目标差异化特征的提取，并提升了网络效率。EfficientDet 提出了加权的双向 FPN，通过引入可学习的权重来评估不同输入特征的重要性，从而实现融合过程中多尺度特征图均衡贡献。空间金字塔池化网络（Spatial Pyramid Pooling，SPP）通过引入空间金字塔池化层，实现从任意尺寸的图像中生成固定长度的特征表示，提升了图像分类与目标检测任务的精度与效率。

传统 CNN 在处理多尺度目标时面临挑战，视觉 Transformer 利用层级化的自注意力机制来构建跨尺度的特征表示，而不过度依赖空间降采样，这为红外小目标检测提供了另一种思路。多尺度 ViT（Multiscale ViT）将 CNN 结构中的多尺度特征提取思想与 Transformer 结合，实现多尺度特征提取。金字塔 ViT（Pyramid ViT）使用渐进缩小的金字塔结构来减少大尺寸特征图的计算量，可作为 CNN 主干网络的替代方案，在目标检测中展现出优越性能。CrossViT 则采用双分支 Transformer 处理不同尺寸的图像块，生成多尺度图像特征，并通过交叉注意力机制进行特征交互学习。

尽管上述多尺度方法整合了不同层级的特征，但它们通常不加区分地融合所有层次的特征，这会带来计算成本的增加，并可能放大背景噪声，从而导致性能下降。这些局限性凸显了当前方法需要一种更精细、更具选择性的融合策略，以专注于跨尺度中最具信息量的特征。

1.3 基于多特征聚焦与跨阶段 Transformer 的检测网络

1.3.1 模型架构

本节将详细介绍基于多特征聚焦与跨阶段 Transformer 的目标检测网络 MFF-DCNet，其整体框架如图1-1所示。该网络专为红外航拍图像中的小目标检测任务设计，核心包

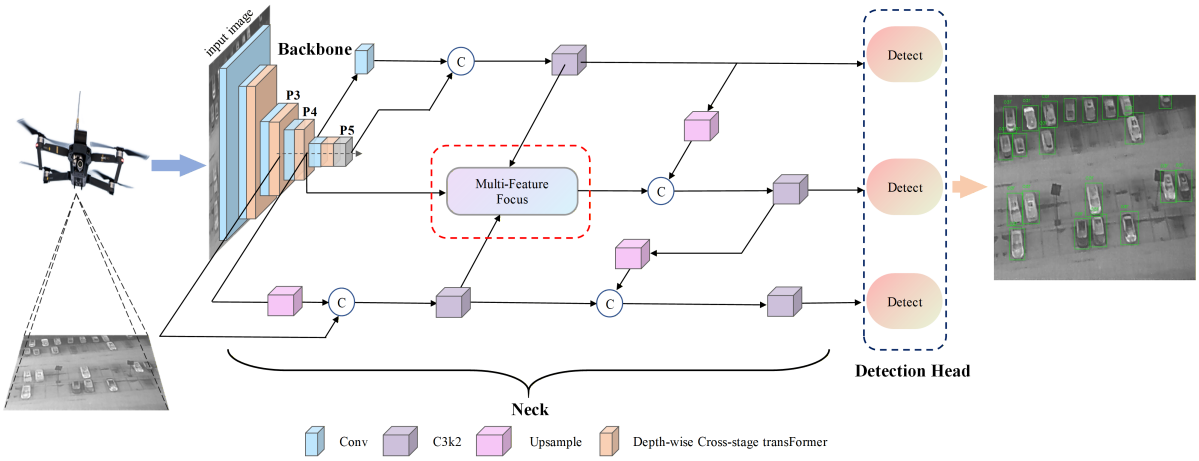

图 1-1 MFF-DCNet 网络架构示意图

含一个集成了新型 DCFormer 模块的增强型主干网络和一个设计有 MFF 模块的创新颈部结构。针对 YOLOv11 主干网络中 C3k2 模块存在的计算效率低下以及跨阶段特征融合能力不足的问题，本章设计了 DCFormer 模块作为其替代。该模块融合了 Transformer 的设计思想，利用深度可分离卷积作为轻量化的特征混合器，并结合跨阶段残差连接。这一设计通过空间解耦的操作显著降低计算成本，同时通过跨阶段特征重组与高效的长程依赖捕获，增强了模型的上下文建模能力。与此同时，为从根本上改善小目标检测性能，我们设计了一个包含多特征聚焦（Multi-Feature Focus，MFF）模块的全新颈部结构。当前应对多尺度挑战的主流方法是构建特征金字塔来整合不同尺度的特征。然而，对于小目标而言，其特征信息在经过连续的卷积层后会逐渐衰减，导致在最终特征图中保留的有效像素极少。因此，在检测头之前有策略地保留并增强高分辨率特征信息，无疑是提升小目标检测能力的关键。MFF 模块正是通过优化特征聚合过程来实现这一目标，它显著增强了网络对于微小目标的检测能力。

MFF 模块通过一个结构化的融合过程来聚合多尺度特征。具体而言，在主干网络完成特征提取后，我们获得特征图 P3、P4 和 P5。这些特征图的分辨率分别为输入图像尺寸的 1/8、1/16 和 1/32。其中，P3 层捕获空间细节，P4 层在空间细节与语义丰富性之间取得平衡，而 P5 层则专注于高层语义信息。P4 层因其居中的位置，成为双向特征传播的核心枢纽。我们为 P4 层创建了两个处理分支。一个分支对 P4 特征图进行 3×3 卷积以实现下采样，将其与经过 SPPF 模块和 C2PSA 模块处理的 P5 层特征融合，随后通过 DCFormer 模块处理，输出一个尺寸为 20×20×512 的特征图。SPPF 模块是传统空间金字塔池化（SPP）的优化版本，在保留 SPP 核心功能的同时提升了计算效率，它通过级联结构串行执行多个相同核尺寸的最大池化操作，C2PSA 模块则整合了跨阶段局部连接（Cross Stage Partial）结构与注意力机制，以增强特征表征能力。P4 层的另一个分支通过最近邻插值进行上采样，与 P3 层特征合并，并经由 DCFormer 模块处理，生成一个 80×80×128 的特征图。上述操作得到的特征，连同原始的 P4 层特征，一并送入 MFF 模块进行进一步的处理。

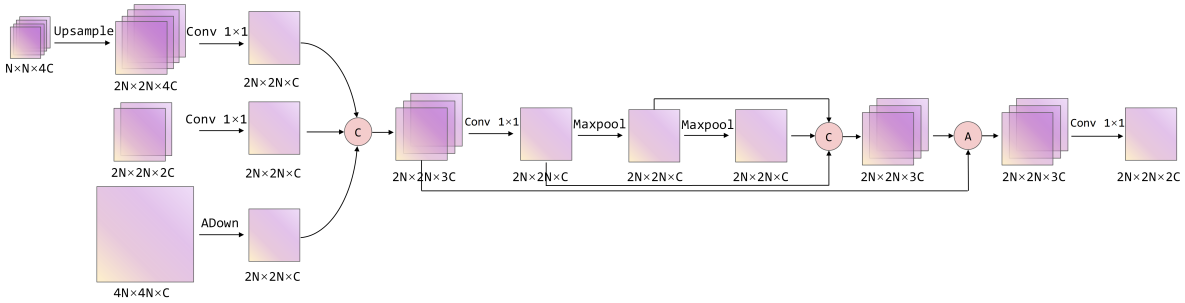

图 1-2 多特征聚焦模块示意图

1.3.2 多特征聚焦模块

多特征聚焦模块是特征聚焦颈部结构中的核心组件，负责对来自三个不同尺度的特征进行集成与融合。其详细结构如图1-2所示，完整的算法流程总结于算法 1。具体而言，对于尺寸为 $20 \times 20 \times 512$ 的 P5 层输出特征，首先通过最近邻上采样将其空间尺度扩大一倍至 $40 \times 40 \times 512$，随后经过一个 1×1 卷积进行通道调整，得到尺寸为 $40 \times 40 \times \mathcal{C}$ 的特征。对于尺寸为 $40 \times 40 \times 256$ 的 P4 层输出特征，同样使用 1×1 卷积调整通道数，得到尺寸为 $40 \times 40 \times \mathcal{C}$ 的特征。对于尺寸为 $80 \times 80 \times 128$ 的 P3 层输出特征，则先通过一个 Adown 模块进行下采样，再进行通道调整，最终得到尺寸为 $40 \times 40 \times \mathcal{C}$ 的特征，其中，$\mathcal{C}$ 被设定为 P4 层特征通道数的一半，在本文中取值为 128。此步骤将所有输入特征在空间维度和通道维度上对齐，为后续的特征融合做好准备。

经过上述处理，所有三个尺度的特征均被对齐到统一的尺度 $40 \times 40 \times \mathcal{C}$。随后，这三个尺度相同的特征图被拼接起来，形成初始的融合特征图，其尺寸为 $40 \times 40 \times 384$。该拼接后的特征接着通过一个 1×1 卷积进行处理，以将通道数调整至 128，从而得到尺寸为 $40 \times 40 \times 128$ 的特征。此后，使用一个 5×5 的池化窗口、步长为 1 并进行边缘填充（以保持输出特征图尺寸）执行两次最大池化操作。将前述三个操作（包括一次 1×1 卷积和两次最大池化）产生的特征进行拼接。具体而言，这次拼接融合了：经过 1×1 卷积后的特征（$40 \times 40 \times 128$）、第一次最大池化后的特征以及第二次最大池化后的特征，最终得到一个尺寸为 $40 \times 40 \times 384$ 的特征。优化后的特征首先与初始拼接阶段产生的融合特征进行相加，随后通过 1×1 卷积对合并后的特征进行进一步优化，同时保持特征尺度不变，输出最终的融合特征图，其尺寸为 $40 \times 40 \times 256$。

1.3.3 基于深度分离卷积的跨阶段 Transformer

YOLOv11 的主干网络主要由连续的标准卷积和 C3k2 模块构成。C3k2 模块通过跨阶段局部连接和可变核卷积进行特征整合，其瓶颈结构是负责通道变换和局部上下文聚合的核心。然而，传统的 C3k2 模块在红外小目标检测中存在关键局限：其串行的卷积结构不可避免地会削弱小目标的判别性特征表示，这一问题在目标缺乏显著纹理、且与复杂背景对比度低的红外图像中尤为突出。

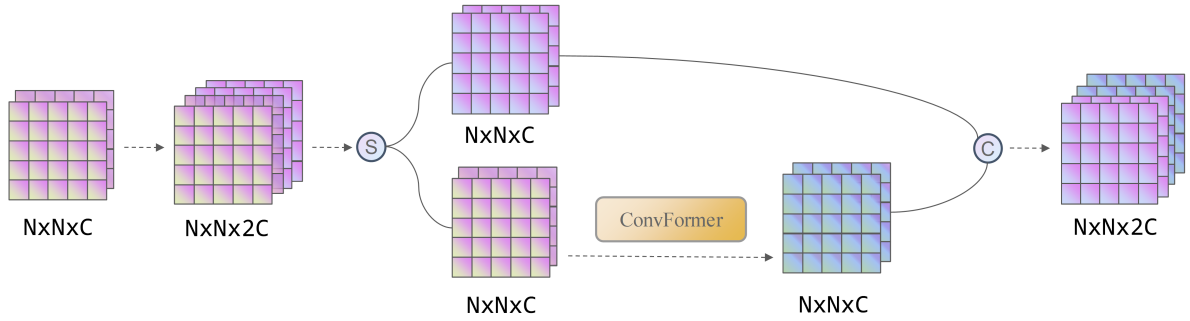

图 1-3 DCFormer 架构示意图

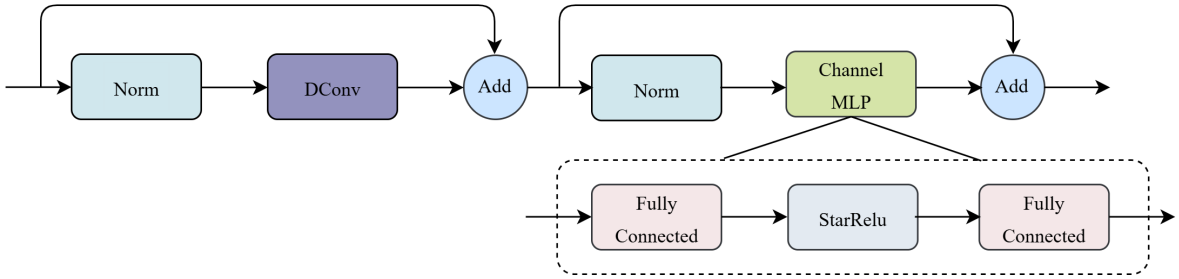

图 1-4 Convformer 架构示意图

DCFormer 模块通过重新设计特征传播路径，利用深度可分离卷积和受 Transformer 启发的跨阶段特征提取机制，有效应对了上述局限。DCFormer 在降低计算成本的同时，增强了对上下文信息的捕获和跨阶段特征提取能力。如图1-3所示，其处理流程始于一个卷积-批归一化-SiLU 激活模块（CBS），所得特征图沿通道维度被均匀分割为两部分，其中一部分通过 ConvFormer 块（其详细结构见图1-4），ConvFormer 采用深度可分离卷积作为高效的特征混合器，以替代标准的自注意力机制。另一部分则与 ConvFormer 的输出沿通道维度进行拼接，从而实现有效的跨阶段信息传播。最后，拼接后的特征再经由另一个 CBS 模块处理。DCFormer 的架构结合了卷积与 Transformer 思想，优化了特征提取过程，特别是增强对红外小目标的检测能力。值得注意的是，该设计相比原始的 C3k2 模块，所需的计算成本更低。

DCFormer 架构的一个核心组件是 ConvFormer，其设计灵感来源于 MetaFormer 框架，作为主要的特征提取单元。与依赖计算密集型自注意力机制的传统 Transformer 不同，ConvFormer 通过深度可分离卷积实现了一种高效的特征混合策略。该架构用具有线性复杂度的深度可分离卷积替代了具有二次复杂度的注意力操作，使其能够在保持实时处理能力的同时，捕获细微的红外目标特征。ConvFormer 的详细结构如图1-4所示，它保持了标准 Transformer 编码器的结构，包含两个主要部分：一是用于空间信息提取的特征混合器（token mixer），二是带有残差连接的多层感知机（Multi-Layer Perceptron，MLP）。在 ConvFormer 中，特征混合器由深度可分离卷积实现，可以表示为：

$$X = DConv(Norm(X)) + X. \tag{1-1}$$

其中，X 表示输入特征图，$Norm(\cdot)$ 表示归一化操作，$DConv(\cdot)$ 表示深度可分离卷积。通过这种设计，ConvFormer 能够高效地捕获空间上下文信息，同时保持较低的计算复杂度。随后，通过一个带有 StarReLU 激活函数的双层 MLP 进行特征变换：

$$X = \sigma(Norm(X)W_1)W_2 + X. \tag{1-2}$$

其中，W_1 和 W_2 分别表示 MLP 的权重矩阵，$\sigma(\cdot)$ 表示 StarReLU 激活函数。

在 ConvFormer 中采用深度可分离卷积作为特征混合器，主要基于其在计算效率方面的显著优势，这对于实时红外检测系统至关重要。与标准卷积对所有输入通道执行卷积运算不同，深度可分离卷积将此过程分解为两步：深度卷积和逐点卷积。深度卷积独立处理每个输入通道，而逐点卷积则负责跨通道集成输出。这种分解在保持表征能力的同时，显著减少了参数量和计算成本。假设输入特征图尺寸为 $W_i \times H_i \times C_i$，卷积核尺寸为 $K_w \times K_h \times C_i$, 输出特征图尺寸为 $W_o \times H_o \times C_o$, 对于标准卷积，单个卷积核包含 $K_w \times K_h \times C_i$ 个参数和一个偏置项，共有 C_o 个卷积核，其参数量和计算量如下：

$$Params_{std\_conv} = (K_w \times K_h \times C_i + 1) \times C_o. \tag{1-3}$$

$$FLOPs_{std\_conv} = K_w \times K_h \times C_i \times W_o \times H_o \times C_o. \tag{1-4}$$

对于深度卷积，单个卷积核的维度为 $K_w \times K_h \times 1$，并带有一个偏置项。使用 C_i 个卷积核，输出特征图维度为 $W_o \times H_o \times C_i$，参数量和计算量如下：

$$Params_{depth\_conv} = (K_w \times K_h \times 1 + 1) \times C_i. \tag{1-5}$$

$$FLOPs_{depth\_conv} = K_w \times K_h \times W_o \times H_o \times C_i. \tag{1-6}$$

对于逐点卷积，单个卷积核的维度为 $1 \times 1 \times C_i$，并带有一个偏置项，使用 C_o 个卷积核，参数量和计算量如下：

$$Params_{point\_conv} = (1 \times 1 \times C_i + 1) \times C_o. \tag{1-7}$$

$$FLOPs_{point\_conv} = C_i \times W_o \times H_o \times C_o. \tag{1-8}$$

对于深度可分离卷积，总参数量和总计算量为：

$$Params_{Dconv} = (K_w \times K_h + 1) \times C_i + (C_i + 1) \times C_o. \tag{1-9}$$

$$FLOPs_{Dconv} = W_o \times H_o \times C_i \times (K_w \times K_h + C_o). \tag{1-10}$$

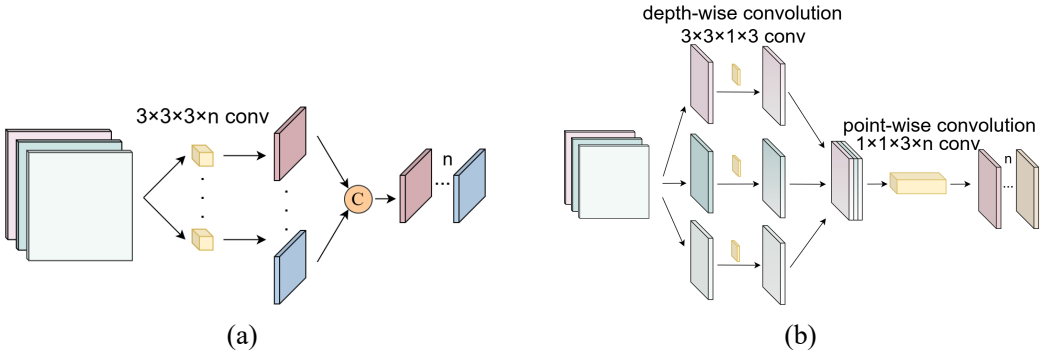

图 1-5 (a) Standard convolution. (b) Depth-wise separable convolution.

假设 $K_w = K_h = K$，通过将传统卷积替换为深度可分离卷积，理论计算量可降低至标准卷积的 $\frac{1}{C_o} + \frac{1}{K^2}$，图1-5直观对比了标准卷积与深度卷积的处理过程。

致　谢

漫漫求学路，最让人回味的，莫属于读博这几年。回首初入交大时情不自禁的喜悦，经历了硕博八年洗礼后，依旧幸福感满存。谨以此文聊表感激之心。

衷心感谢我的导师孙宏滨教授在博士期间对我的悉心指导与关怀。在刚进组时，孙老师就将新发现号的搭建工作交与我全权负责，帮助我从系统的角度对无人驾驶整体研究有了深刻认识；在博二时，孙老师就让我担任了发现号车队队长并认真细致地指导我们准备每年的未来挑战赛，极大地提高了我的组织管理能力；在科研工作中，孙老师指点迷津，引领我做好科研探索。孙老师严谨务实的科研态度，一丝不苟的治学精神，高屋建瓴的学术见地，勤奋谦虚的个人品质都深深感染着我，激励着我，使我受益终生。

感谢我们敬爱的郑南宁院士。郑老师对于无人驾驶车队的关心和指导使我们整个车队的技术水平得到不断提高。感谢我博士前两年的合作导师辛景民教授的关怀，感谢魏平教授在智能车未来挑战赛备赛和比赛过程中的悉心教导，感谢王乐教授在轨迹预测方面的支持，感谢薛建儒教授、兰旭光教授、任鹏举教授、杜少毅教授、徐林海高级工程师、陈仕韬助理教授、王芳芳工程师以及其他所有人工智能学院老师在我读博期间给予的帮助和支持。

感谢课题组张旭翀师兄和汪航师兄对我科研工作一直以来的帮助，两位师兄扎实的理论功底和极强的解决问题能力都给我留下了很深印象。感谢沈源、张婧、刘丹为我们的学习生活提供的便利。

感谢王潇、史菊旺、李庚欣、陶中幸、张璞等师兄师姐在科研上的关照。感谢冯洋、杨帅、吴金强、冯超、向钊宏、陈达、张志浩、王玉学、韩伟光、权柄章、钱成龙、葛冲、陈科、李诚、罗鑫凯、陈煜炜、王申奥、李天航等师弟在发现号无人驾驶平台开发和无人车比赛中的付出。感谢戴赫、孙长峰、郑方、段景海、石刘帅等师弟在小论文上的帮助。感谢同届张剑、杨少飞、李宝婷的帮助。感谢唐浩雯师妹在科研生活中的交流与帮助。感谢好友冯立琛、丁兆伦、雷洁、马晨、荣韧闲暇时度过的快乐时光。感谢和我一起在创新港并肩战斗的赵博然，在科研和为人处世方面都对我产生了很大影响。

最后，感谢我的父母和家人多年来对我学习和生活上的关心和支持，是你们的坚强后盾让我能够全身心地投入到科研探索中。感恩一路有你们相伴，你们永远是我内心最温暖的港湾。

参考文献

[1] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal Speed and Accuracy of Object Detection[J/OL]. ArXiv, 2020, abs/2004.10934. https://api.semanticscholar.org/CorpusID:216080778.

[2] Zhang H, Cisse M, Dauphin Y N, et al. mixup: Beyond Empirical Risk Minimization[C/OL]. 2018. https://openreview.net/forum?id=r1Ddp1-Rb.

[3] Yun S, Han D, Oh S J, et al. Cutmix: Regularization strategy to train strong classifiers with localizable features[C]. Proceedings of the IEEE/CVF international conference on computer vision. 2019: 6023-6032.

[4] Kisantal M, Wojna Z, Murawski J, et al. Augmentation for small object detection[J]. arXiv preprint arXiv:1902.07296, 2019.

[5] Chen C, Zhang Y, Lv Q, et al. RRNet: A Hybrid Detector for Object Detection in Drone-Captured Images[C/OL]. 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). 2019: 100-108. DOI: 10.1109/ICCVW.2019.00018.

[6] Xiao J, Guo H, Zhou J, et al. Tiny object detection with context enhancement and feature purification [J]. Expert Systems with Applications, 2023, 211: 118665.

[7] Ünel F Ö, Özkalayci B O, Çiğla C. The Power of Tiling for Small Object Detection[C/OL]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2019: 582-591. DOI: 10.1109/CVPRW.2019.00084.

[8] Yu X, Gong Y, Jiang N, et al. Scale Match for Tiny Person Detection[C/OL]. 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). 2020: 1246-1254. DOI: 10.1109/WACV45572.2020.9093394.

[9] Lin J, Jing W, Song H. SAN: Scale-aware network for semantic segmentation of high-resolution aerial images[J]. arXiv preprint arXiv:1907.03089, 2019.

[10] Zoph B, Cubuk E D, Ghiasi G, et al. Learning data augmentation strategies for object detection[C]. European conference on computer vision. 2020: 566-583.

[11] Yang F, Choi W, Lin Y. Exploit All the Layers: Fast and Accurate CNN Object Detector with Scale Dependent Pooling and Cascaded Rejection Classifiers[C/OL]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 2129-2137. DOI: 10.1109/CVPR.2016.234.

[12] Cai Z, Fan Q, Feris R S, et al. A Unified Multi-scale Deep Convolutional Neural Network for Fast Object Detection[C]. Leibe B, Matas J, Sebe N, et al. Computer Vision – ECCV 2016. Cham: Springer International Publishing, 2016: 354-370.

[13] Redmon J, Farhadi A. YOLOv3: An Incremental Improvement[EB/OL]. 2018. https://arxiv.org/abs/1804.02767. arXiv: 1804.02767 [cs.CV].

[14] Lin T Y, Dollár P, Girshick R, et al. Feature Pyramid Networks for Object Detection[C/OL]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 936-944. DOI: 10.1109/CVPR.2017.106.

[15] Ghiasi G, Lin T Y, Le Q V. NAS-FPN: Learning Scalable Feature Pyramid Architecture for Object Detection[C/OL]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 7029-7038. DOI: 10.1109/CVPR.2019.00720.

[16] Qiao S, Chen L C, Yuille A. DetectoRS: Detecting Objects with Recursive Feature Pyramid and Switchable Atrous Convolution[C/OL]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 10208-10219. DOI: 10.1109/CVPR46437.2021.01008.

[17] Li J, Liang X, Shen S, et al. Scale-Aware Fast R-CNN for Pedestrian Detection[J/OL]. IEEE Transactions on Multimedia, 2018, 20(4): 985-996. DOI: 10.1109/TMM.2017.2759508.

[18] Yang C, Huang Z, Wang N. QueryDet: Cascaded Sparse Query for Accelerating High-Resolution Small Object Detection[C/OL]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022: 13658-13667. DOI: 10.1109/CVPR52688.2022.01330.

[19] Singh B, Davis L S. An Analysis of Scale Invariance in Object Detection - SNIP[J/OL]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017: 3578-3587. https://api.semanticscholar.org/CorpusID:4615054.

[20] Singh B, Najibi M, Davis L S. SNIPER: Efficient Multi-Scale Training[J]. NeurIPS, 2018.

[21] Najibi M, Singh B, Davis L S. AutoFocus: Efficient Multi-Scale Inference[J]. ICCV, 2019.

[22] Chen Y, Zhang P, Li Z, et al. Dynamic Scale Training for Object Detection[EB/OL]. 2021. https://arxiv.org/abs/2004.12432. arXiv: 2004.12432 [cs.CV].

[23] Li Y, Chen Y, Wang N, et al. Scale-Aware Trident Networks for Object Detection[J]. ICCV 2019, 2019.

[24] Liu S, Qi L, Qin H, et al. Path Aggregation Network for Instance Segmentation[C/OL]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 8759-8768. DOI: 10.1109/CVPR.2018.00913.

[25] Tan M, Pang R, Le Q V. EfficientDet: Scalable and Efficient Object Detection[C/OL]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 10778-10787. DOI: 10.1109/CVPR42600.2020.01079.

[26] Zhang H, Wang K, Tian Y, et al. MFR-CNN: Incorporating Multi-Scale Features and Global Information for Traffic Object Detection[J/OL]. IEEE Transactions on Vehicular Technology, 2018, 67(9): 8019-8030. DOI: 10.1109/TVT.2018.2843394.

[27] Woo S, Hwang S, Kweon I S. StairNet: Top-Down Semantic Aggregation for Accurate One Shot Detection[J/OL]. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2017: 1093-1102. https://api.semanticscholar.org/CorpusID:13681687.

[28] Zhao Q, Sheng T, Wang Y, et al. M2Det: A Single-Shot Object Detector based on Multi-Level Feature Pyramid Network[C]. The Thirty-Third AAAI Conference on Artificial Intelligence,AAAI. 2019.

[29] Liu Z, Gao G, Sun L, et al. IPG-Net: Image Pyramid Guidance Network for Small Object Detection[C/OL]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2020: 4422-4430. DOI: 10.1109/CVPRW50498.2020.00521.

[30] Gong Y, Yu X, Ding Y, et al. Effective Fusion Factor in FPN for Tiny Object Detection[C/OL]. 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). 2021: 1159-1167. DOI: 10.1109/WACV48630.2021.00120.

[31] Hong M, Li S, Yang Y, et al. SSPNet: Scale Selection Pyramid Network for Tiny Person Detection From UAV Images[J/OL]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 1-5. DOI: 10.1109/LGRS.2021.3103069.

[32] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]. NIPS'14: Pro-

ceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2. Montreal, Canada: MIT Press, 2014: 2672-2680.

[33] Li J, Liang X, Wei Y, et al. Perceptual Generative Adversarial Networks for Small Object Detection [J/OL]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 1951-1959. https://api.semanticscholar.org/CorpusID:6704804.

[34] Bai Y, Zhang Y, Ding M, et al. SOD-MTGAN: Small Object Detection via Multi-Task Generative Adversarial Network[C]. Computer Vision – ECCV 2018. Cham: Springer International Publishing, 2018: 210-226.

[35] Pang Y, Cao J, Wang J, et al. JCS-Net: Joint Classification and Super-Resolution Network for Small-Scale Pedestrian Detection in Surveillance Images[J/OL]. IEEE Transactions on Information Forensics and Security, 2019, 14(12): 3322-3331. DOI: 10.1109/TIFS.2019.2916592.

[36] Kim J, Lee J K, Lee K M. Accurate Image Super-Resolution Using Very Deep Convolutional Networks[C/OL]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 1646-1654. DOI: 10.1109/CVPR.2016.182.

[37] Cao J, Pang Y, Li X. Learning Multilayer Channel Features for Pedestrian Detection[J/OL]. IEEE Transactions on Image Processing, 2017, 26(7): 3210-3220. DOI: 10.1109/TIP.2017.2694224.

[38] Fu C Y, Liu W, Ranga A, et al. DSSD : Deconvolutional Single Shot Detector[EB/OL]. 2017. https://arxiv.org/abs/1701.06659. arXiv: 1701.06659 [cs.CV].

[39] Corsel C W, van Lier M, Kampmeijer L, et al. Exploiting Temporal Context for Tiny Object Detection[C/OL]. 2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW). 2023: 1-11. DOI: 10.1109/WACVW58289.2023.00013.

[40] Zhang S, Wen L, Bian X, et al. Single-Shot Refinement Neural Network for Object Detection [C/OL]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 4203-4212. DOI: 10.1109/CVPR.2018.00442.

[41] Yi K, Jian Z, Chen S, et al. Feature Selective Small Object Detection via Knowledge-based Recurrent Attentive Neural Network[EB/OL]. 2019. https://arxiv.org/abs/1803.05263. arXiv: 1803.05263 [cs.CV].

[42] Yang X, Yang J, Yan J, et al. SCRDet: Towards More Robust Detection for Small, Cluttered and Rotated Objects[C/OL]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2019: 8231-8240. DOI: 10.1109/ICCV.2019.00832.

[43] Fu J, Sun X, Wang Z, et al. An Anchor-Free Method Based on Feature Balancing and Refinement Network for Multiscale Ship Detection in SAR Images[J/OL]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 59(2): 1331-1344. DOI: 10.1109/TGRS.2020.3005151.

[44] Lu X, Ji J, Xing Z, et al. Attention and Feature Fusion SSD for Remote Sensing Object Detection [J/OL]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-9. DOI: 10.1109/TIM.2021.3052575.

[45] Ran Q, Wang Q, Zhao B, et al. Lightweight Oriented Object Detection Using Multiscale Context and Enhanced Channel Attention in Remote Sensing Images[J/OL]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 5786-5795. DOI: 10.1109/JSTARS.2021.3079968.

[46] Li Y, Huang Q, Pei X, et al. Cross-Layer Attention Network for Small Object Detection in Remote Sensing Imagery[J/OL]. IEEE Journal of Selected Topics in Applied Earth Observations and

Remote Sensing, 2021, 14: 2148-2161. DOI: 10.1109/JSTARS.2020.3046482.

[47] Tian Z, Shen C, Chen H, et al. FCOS: A Simple and Strong Anchor-Free Object Detector[J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(4): 1922-1933. DOI: 10.1109/TPAMI.2020.3032166.

[48] Yang F, Fan H, Chu P, et al. Clustered Object Detection in Aerial Images[C/OL]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2019: 8310-8319. DOI: 10.1109/ICCV.2019.00840.

[49] Duan C, Wei Z, Zhang C, et al. Coarse-grained Density Map Guided Object Detection in Aerial Images[C/OL]. 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). 2021: 2789-2798. DOI: 10.1109/ICCVW54120.2021.00313.

[50] Li C, Yang T, Zhu S, et al. Density Map Guided Object Detection in Aerial Images[C/OL]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2020: 737-746. DOI: 10.1109/CVPRW50498.2020.00103.

[51] Wang Y, Yang Y, Zhao X. Object Detection Using Clustering Algorithm Adaptive Searching Regions in Aerial Images[C]. Computer Vision – ECCV 2020 Workshops. Springer International Publishing, 2020: 651-664.

[52] Deng S, Li S, Xie K, et al. A Global-Local Self-Adaptive Network for Drone-View Object Detection [J/OL]. IEEE Transactions on Image Processing, 2021, 30: 1556-1569. DOI: 10.1109/TIP.2020.3045636.

[53] Xu J, Li Y, Wang S. AdaZoom: Adaptive Zoom Network for Multi-Scale Object Detection in Large Scenes[EB/OL]. 2021. https://arxiv.org/abs/2106.10409. arXiv: 2106.10409 [cs.CV].

[54] Leng J, Mo M, Zhou Y, et al. Pareto Refocusing for Drone-View Object Detection[J/OL]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(3): 1320-1334. DOI: 10.1109/TCSVT.2022.3210207.

[55] Koyun O C, Keser R K, Akkaya İ B, et al. Focus-and-Detect: A small object detection framework for aerial images[J/OL]. Signal Processing: Image Communication, 2022, 104: 116675. DOI: https://doi.org/10.1016/j.image.2022.116675.

[56] Cui L, Lv P, Jiang X, et al. Context-Aware Block Net for Small Object Detection[J/OL]. IEEE Transactions on Cybernetics, 2022, 52(4): 2300-2313. DOI: 10.1109/TCYB.2020.3004636.

[57] Sun J, Gao H, Wang X, et al. Scale Enhancement Pyramid Network for Small Object Detection from UAV Images[J/OL]. Entropy, 2022, 24(11). https://www.mdpi.com/1099-4300/24/11/1699. DOI: 10.3390/e24111699.

[58] Bolme D S, Beveridge J R, Draper B A, et al. Visual object tracking using adaptive correlation filters [C/OL]. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2010: 2544-2550. DOI: 10.1109/CVPR.2010.5539960.

[59] Henriques J F, Caseiro R, Martins P, et al. High-Speed Tracking with Kernelized Correlation Filters [J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(3): 583-596. DOI: 10.1109/TPAMI.2014.2345390.

[60] Danelljan M, Häger G, Khan F S, et al. Learning Spatially Regularized Correlation Filters for Visual Tracking[C/OL]. 2015 IEEE International Conference on Computer Vision (ICCV). 2015: 4310-4318. DOI: 10.1109/ICCV.2015.490.

[61] Valmadre J, Bertinetto L, Henriques J, et al. End-to-End Representation Learning for Correlation

Filter Based Tracking[C/OL]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 5000-5008. DOI: 10.1109/CVPR.2017.531.

[62] Bhat G, Danelljan M, Van Gool L, et al. Learning Discriminative Model Prediction for Tracking [C/OL]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2019: 6181-6190. DOI: 10.1109/ICCV.2019.00628.

[63] Danelljan M, Van Gool L, Timofte R. Probabilistic Regression for Visual Tracking[C/OL]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 7181-7190. DOI: 10.1109/CVPR42600.2020.00721.

[64] Bertinetto L, Valmadre J, Henriques J F, et al. Fully-Convolutional Siamese Networks for Object Tracking[C]. Computer Vision – ECCV 2016 Workshops. Cham: Springer International Publishing, 2016: 850-865.

[65] He A, Luo C, Tian X, et al. A Twofold Siamese Network for Real-Time Object Tracking[C/OL]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 4834-4843. DOI: 10.1109/CVPR.2018.00508.

[66] Li B, Yan J, Wu W, et al. High Performance Visual Tracking with Siamese Region Proposal Network [J/OL]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018: 8971-8980. https://api.semanticscholar.org/CorpusID:52255840.

[67] Zhu Z, Wang Q, Bo L, et al. Distractor-aware Siamese Networks for Visual Object Tracking[C]. European Conference on Computer Vision. 2018.

[68] Li B, Wu W, Wang Q, et al. SiamRPN++: Evolution of Siamese Visual Tracking With Very Deep Networks[C/OL]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 4277-4286. DOI: 10.1109/CVPR.2019.00441.

[69] Xu Y, Wang Z, Li Z, et al. SiamFC++: Towards Robust and Accurate Visual Tracking with Target Estimation Guidelines[J/OL]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(07): 12549-12556. https://ojs.aaai.org/index.php/AAAI/article/view/6944. DOI: 10.1609/aaai.v34i07.6944.

[70] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C/OL]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 770-778. DOI: 10.1109/CVPR.2016.90.

[71] Voigtlaender P, Luiten J, Torr P H, et al. Siam R-CNN: Visual Tracking by Re-Detection[C/OL]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 6577-6587. DOI: 10.1109/CVPR42600.2020.00661.

[72] Yu Y, Xiong Y, Huang W, et al. Deformable Siamese Attention Networks for Visual Object Tracking [C/OL]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 6727-6736. DOI: 10.1109/CVPR42600.2020.00676.

[73] Liu J, Wang H, Ma C, et al. SiamDMU: Siamese Dual Mask Update Network for Visual Object Tracking[J/OL]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2024, 8(2): 1656-1669. DOI: 10.1109/TETCI.2024.3353674.

[74] Chen X, Yan B, Zhu J, et al. Transformer Tracking[C]. CVPR. 2021.

[75] Wang N, Zhou W, Wang J, et al. Transformer Meets Tracker: Exploiting Temporal Context for Robust Visual Tracking[C/OL]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 1571-1580. DOI: 10.1109/CVPR46437.2021.00162.

[76] Yan B, Peng H, Fu J, et al. Learning Spatio-Temporal Transformer for Visual Tracking[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2021: 10448-10457.

[77] Song Z, Yu J, Chen Y P P, et al. Transformer Tracking with Cyclic Shifting Window Attention [C/OL]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022: 8781-8790. DOI: 10.1109/CVPR52688.2022.00859.

[78] Liu Z, Lin Y, Cao Y, et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows[J/OL]. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021: 9992-10002. https://api.semanticscholar.org/CorpusID:232352874.

[79] Cui Y, Jiang C, Wang L, et al. MixFormer: End-to-End Tracking with Iterative Mixed Attention [C/OL]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022: 13598-13608. DOI: 10.1109/CVPR52688.2022.01324.

[80] Wu H, Xiao B, Codella N, et al. CvT: Introducing Convolutions to Vision Transformers[C/OL]. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2021: 22-31. DOI: 10.1109/ICCV48922.2021.00009.

[81] Lin L, Fan H, Zhang Z, et al. SwinTrack: A Simple and Strong Baseline for Transformer Tracking [C/OL]. Advances in Neural Information Processing Systems: vol. 35. 2022: 16743-16754. https://proceedings.neurips.cc/paper_files/paper/2022/file/6a5c23219f401f3efd322579002dbb80-Paper-Conference.pdf.

[82] Ye B, Chang H, Ma B, et al. Joint Feature Learning and Relation Modeling for Tracking: A One-Stream Framework[C]. ECCV. 2022.

[83] Chen X, Peng H, Wang D, et al. SeqTrack: Sequence to Sequence Learning for Visual Object Tracking[C/OL]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023: 14572-14581. DOI: 10.1109/CVPR52729.2023.01400.

[84] Hong L, Yan S, Zhang R, et al. OneTracker: Unifying Visual Object Tracking with Foundation Models and Efficient Tuning[C/OL]. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024: 19079-19091. DOI: 10.1109/CVPR52733.2024.01805.

[85] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]. Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13. 2014: 740-755.

[86] Jocher G, Qiu J, Chaurasia A. Ultralytics YOLO[CP/OL]. 8.0.0. 2023. https://github.com/ultralytics/ultralytics.

[87] Cao Y, He Z, Wang L, et al. VisDrone-DET2021: The vision meets drone object detection challenge results[C]. Proceedings of the IEEE/CVF International conference on computer vision. 2021: 2847-2854.

[88] Lv W, Zhao Y, Chang Q, et al. Rt-detrv2: Improved baseline with bag-of-freebies for real-time detection transformer[J]. arXiv preprint arXiv:2407.17140, 2024.

[89] Li Y, Wu P, Zhang M. Rethinking the sparse mask learning mechanism in sparse convolution for object detection on drone images[J]. Computer Vision and Image Understanding, 2025: 104432.

[90] Leng J, Ye Y, Mo M, et al. Recent Advances for Aerial Object Detection: A Survey[J]. ACM Computing Surveys, 2024, 56(12): 1-36.

[91] Tan L, Liu Z, Liu H, et al. A Real-Time Unmanned Aerial Vehicle (UAV) Aerial Image Object Detection Model[C]. 2024 International Joint Conference on Neural Networks (IJCNN). 2024: 1-7.

[92] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]. European conference on computer vision. 2020: 213-229.

[93] Huang Y X, Liu H I, Shuai H H, et al. Dq-detr: Detr with dynamic query for tiny object detection [C]. European Conference on Computer Vision. 2024: 290-305.

[94] Du D, Qi Y, Yu H, et al. The unmanned aerial vehicle benchmark: Object detection and tracking [C]. Proceedings of the European conference on computer vision (ECCV). 2018: 370-386.

[95] Wang J, Yang W, Guo H, et al. Tiny Object Detection in Aerial Images[C/OL]. 2020 25th International Conference on Pattern Recognition (ICPR). 2021: 3791-3798. DOI: 10.1109/ICPR48806.2021.9413340.

[96] Xu X, Mao Z, Wang X, et al. Dynamic Anchor: Density Map Guided Small Object Detector for Tiny Persons[J]. Computer Vision and Image Understanding, 2025, 255: 104325.

[97] Li C, Yang T, Zhu S, et al. Density map guided object detection in aerial images[C]. proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020: 190-191.

[98] Du B, Huang Y, Chen J, et al. Adaptive sparse convolutional networks with global context enhancement for faster object detection on drone images[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 13435-13444.

[99] Akyon F C, Altinuc S O, Temizel A. Slicing aided hyper inference and fine-tuning for small object detection[C]. 2022 IEEE international conference on image processing (ICIP). 2022: 966-970.

[100] Zhu X, Su W, Lu L, et al. Deformable detr: Deformable transformers for end-to-end object detection [J]. arXiv preprint arXiv:2010.04159, 2020.

[101] Li F, Zhang H, Liu S, et al. Dn-detr: Accelerate detr training by introducing query denoising[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 13619-13627.

[102] Yao Z, Ai J, Li B, et al. Efficient detr: improving end-to-end object detector with dense prior[J]. arXiv preprint arXiv:2104.01318, 2021.

[103] Roh B, Shin J, Shin W, et al. Sparse detr: Efficient end-to-end object detection with learnable sparsity[J]. arXiv preprint arXiv:2111.14330, 2021.

[104] Zhao Y, Lv W, Xu S, et al. Detrs beat yolos on real-time object detection[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 16965-16974.

[105] Zhang H, Liu K, Gan Z, et al. UAV-DETR: Efficient End-to-End Object Detection for Unmanned Aerial Vehicle Imagery[J]. arXiv preprint arXiv:2501.01855, 2025.

[106] Xue H, Tang Z, Xia Y, et al. HCTD: A CNN-transformer hybrid for precise object detection in UAV aerial imagery[J]. Computer Vision and Image Understanding, 2025: 104409.

[107] Chen L, Fu Y, Gu L, et al. Frequency-aware feature fusion for dense image prediction[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.

[108] Wang J, Chen K, Xu R, et al. CARAFE: Content-Aware ReAssembly of FEatures[C/OL]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2019: 3007-3016. DOI: 10.1109/ICCV.2019.00310.

[109] Wang J, Xu C, Yang W, et al. A normalized Gaussian Wasserstein distance for tiny object detection [J]. arXiv preprint arXiv:2110.13389, 2021.

[110] Wang C Y, Yeh I H, Mark Liao H Y. Yolov9: Learning what you want to learn using programmable gradient information[C]. European conference on computer vision. 2024: 1-21.

[111] Wang A, Chen H, Liu L, et al. Yolov10: Real-time end-to-end object detection[J]. Advances in Neural Information Processing Systems, 2024, 37: 107984-108011.

[112] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]. Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.

[113] Zhu C, He Y, Savvides M. Feature selective anchor-free module for single-shot object detection [C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 840-849.

[114] Liu Z, Gao G, Sun L, et al. HRDNet: High-resolution detection network for small objects[C]. 2021 IEEE international conference on multimedia and expo (ICME). 2021: 1-6.

[115] Xu C, Wang J, Yang W, et al. Detecting tiny objects in aerial images: A normalized Wasserstein distance and a new benchmark[J/OL]. ISPRS Journal of Photogrammetry and Remote Sensing, 2022, 190: 79-93. DOI: https://doi.org/10.1016/j.isprsjprs.2022.06.002.

[116] Guo G, Chen P, Yu X, et al. Save the Tiny, Save the All: Hierarchical Activation Network for Tiny Object Detection[J/OL]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34: 221-234. DOI: 10.1109/TCSVT.2023.3284161.

攻读学位期间取得的研究成果

I. 学术论文

[1] **Weihuang Chen**, Zhigang Yang, Lingyang Xue, Jinghai Duan, Hongbin Sun, Nanning Zheng. Multimodal pedestrian trajectory prediction using probabilistic proposal network[J]. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), 2022. (SCI 1 区，IF：5.859，DOI：10.1109/TCSVT.2022.3229694)

[2] **Weihuang Chen**, Fang Zheng, Liushuai Shi, Yongdong Zhu, Hongbin Sun, Nanning Zheng. Multiple goals network for pedestrian trajectory prediction in autonomous driving[C]. IEEE International Conference on Intelligent Transportation Systems (ITSC), 2022:717–722.

[3] **Weihuang Chen**, Fangfang Wang, Hongbin Sun. S2tnet: Spatio-temporal transformer networks for trajectory prediction in autonomous driving[C]. Asian Conference on Machine Learning (ACML), 2021:454–469. (引用：10）

[4] **Weihuang Chen**, Yuwei Chen, Shen'ao Wang, Tianhang Li, Xuchong Zhang, Hongbin Sun. Motion planning using trajectory tree network for autonomous driving[J]. IEEE Transactions on Vehicular Technology (TVT), 2023, Under review. （投稿号：VT-2023-00733）

[5] Cheng Li, **Weihuang Chen**, Xinkai Luo, Fangfang Wang, Jingmin Zhang, Yanlong Yang, Hongbin Sun. Optimal preview distance control using model prediction for autonomous vehicle[C]. CAA International Conference on Vehicular Control and Intelligence (CVCI). 2021:1–8.

II. 专利

[6] 孙宏滨、**陈炜煌**、王玉学、章浩飞、李煊、吴彝丹，一种面向多场景的自动驾驶规划方法及系统 [P]，专利授权号：ZL202110276175.5

III. 科研获奖

[7] 第一届全国研究生智能挑战赛，三等奖，2019 年。（队长）

[8] 第六届中国研究生智慧城市技术与创意设计大赛，二等奖，2019 年。

[9] 第十二届中国智能车未来挑战赛，全国第 5 名，发现号自动驾驶平台，2020 年。（队长）

IV. 参与项目

[10] 国家重点研发计划项目（2018.05-2023.04）：“下一代深度学习理论、方法与关键技术”（项目编号：2017YFA0700800）

[11] 国家自然科学基金重大项目（2018.01-2022.12）：“极限工况下的人机协同机理及切换控制”（项目编号：61790563）

[12] 横向项目（2021.03-2021.09）：“基于深度学习的传感器数据融合”（项目编号：202103136）

答辩委员会会议决议

轨迹预测与规划是自动驾驶领域的重要研究问题。论文开展了基于深度神经网络的轨迹预测和运动规划方法研究，选题具有重要的研究与应用价值。主要创新点如下：

1. 提出了一种基于时空 Transformer 网络的单模态轨迹预测模型，提升了密集交通环境下不同类别交通参与者的轨迹预测能力。

2. 提出了一种基于概率性候选轨迹网络的多模态轨迹预测模型，提高了交通参与者的多模态轨迹预测速度和精度。

3. 提出了一种基于安全轨迹树网络的运动规划模型，提高了自动驾驶车辆的运动规划性能。

论文写作认真，结构清晰，论述清楚，工作量饱满，表明作者已掌握本学科宽广坚实的基础理论和系统深入的专业知识，独立从事科研工作的能力强，是一篇高质量的博士学位论文。

答辩中讲述清晰，回答问题正确，经答辩委员会讨论和无记名投票表决，一致同意通过学位论文答辩，并一致建议授予陈炜煌同学工学博士学位。

常规评阅人名单

本学位论文共接受 3 位专家评阅，其中常规评阅人 2 名，名单如下：

魏平　教授　　西安交通大学

邓成　教授　西安电子科技大学

学位论文独创性声明（1）

本人声明：所呈交的学位论文系在导师指导下本人独立完成的研究成果。文中依法引用他人的成果，均已做出明确标注或得到许可。论文内容未包含法律意义上已属于他人的任何形式的研究成果，也不包含本人已用于其他学位申请的论文或成果。

本人如违反上述声明，愿意承担以下责任和后果：

1. 交回学校授予的学位证书；
2. 学校可在相关媒体上对作者本人的行为进行通报；
3. 本人按照学校规定的方式，对因不当取得学位给学校造成的名誉损害，进行公开道歉。
4. 本人负责因论文成果不实产生的法律纠纷。

论文作者（签名）：　　日期：　　年　　月　　日

学位论文独创性声明（2）

本人声明：研究生　　所提交的本篇学位论文已经本人审阅，确系在本人指导下由该生独立完成的研究成果。

本人如违反上述声明，愿意承担以下责任和后果：

1. 学校可在相关媒体上对本人的失察行为进行通报；
2. 本人按照学校规定的方式，对因失察给学校造成的名誉损害，进行公开道歉。
3. 本人接受学校按照有关规定做出的任何处理。

指导教师（签名）：　　日期：　　年　　月　　日

学位论文知识产权权属声明

我们声明，我们提交的学位论文及相关的职务作品，知识产权归属学校。学校享有以任何方式发表、复制、公开阅览、借阅以及申请专利等权利。学位论文作者离校后，或学位论文导师因故离校后，发表或使用学位论文或与该论文直接相关的学术论文或成果时，署名单位仍然为西安交通大学。

论文作者（签名）：　　日期：　　年　　月　　日

指导教师（签名）：　　日期：　　年　　月　　日