中图分类号: 学校代码: 10055
UDC: 密级: 公开

南開大學

博 士 学 位 论 文

序列图像的识别与检测算法研究

Research on Recognition and Detection Algorithm for Sequence-Image

论文作者 梅杰 指导教师 程明明 教授
申请学位 工学博士 培养单位 计算机学院
学科专业 计算机科学与技术 研究方向 计算机视觉
答辩委员会主席 周国栋 教授 评 阅 人 匿名评阅人

南开大学研究生院
二〇二二年十月

南开大学学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师指导下进行研究工作所取得的研究成果。除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人创作的、已公开发表或者没有公开发表的作品的内容。对本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。本学位论文原创性声明的法律责任由本人承担。

学位论文作者签名：_____ 年 月 日

非公开学位论文标注说明

(本页表中填写内容须打印)

根据南开大学有关规定，非公开学位论文须经指导教师同意、作者本人申请和相关部门批准方能标注。未经批准的均为公开学位论文，公开学位论文本说明为空白。

论文题目			
申请密级	☐ 限制 (≤2 年) ☐ 秘密 (≤10 年) ☐ 机密 (≤20 年)		
保密期限	20 年 月 日至 20 年 月 日		
审批表编号		**批准日期**	20 年 月 日

南开大学学位评定委员会办公室盖章 (有效)

注：限制 ★2 年 (可少于 2 年); 秘密 ★10 年 (可少于 10 年); 机密 ★20 年 (可少于 20 年)

南开大学学位论文使用授权书

本人完全了解《南开大学关于研究生学位论文收藏和利用管理办法》关于南开大学 (简称“学校”) 研究生学位论文收藏和利用的管理规定，同意向南开大学提交本人的学位论文电子版及相应的纸质本。

本人了解南开大学拥有在《中华人民共和国著作权法》规定范围内的学位论文使用权，同意在以下几方面向学校授权。即：

1. 学校将学位论文编入《南开大学博硕士学位论文全文数据库》，并作为资料在学校图书馆等场所提供阅览，在校园网上提供论文目录检索、文摘及前 16 页的浏览等信息服务；

2. 学校可以采用影印、缩印或其他复制手段保存学位论文；学校根据规定向教育部指定的收藏和存档单位提交学位论文；

3. 非公开学位论文在解密后的使用权同公开论文。

本人承诺：本人的学位论文是在南开大学学习期间创作完成的作品，并已通过论文答辩；提交的学位论文电子版与纸质本论文的内容一致，如因不同造成不良后果由本人自负。

本人签署本授权书一份（此授权书为论文中一页），交图书馆留存。

学位论文作者暨授权人 (亲笔) 签字：_____

20　　年　　月　　日

南开大学研究生学位论文作者信息

论 文 题 目	序列图像的识别与检测算法研究				
姓　　名	梅杰	学号	1120190168	答辩日期	2022 年 11 月 24 日
论 文 类 别	博士 ☑　学历硕士 ☐　专业学位硕士 ☐　同等学力硕士 ☐　划☑选择				
学院（单位）	计算机学院		学科/专业 (专业学位) 名称		计算机科学与技术
联 系 电 话	18811475184		电子邮箱	meijie@mail.nankai.edu.cn	
通讯地址 (邮编)：天津市津南区海河教育园区同砚路 38 号南开大学津南校区（300350）					
非公开论文编号			备注		

注：本授权书适用我校授予的所有博士、硕士的学位论文。如已批准为非公开学位论文，须向图书馆提供批准通过的《南开大学研究生申请非公开学位论文审批表》复印件和“非公开学位论文标注说明”页原件。

摘要

序列图像是一组按照特定时空域顺序排列的图像集合，与普通图像不同，序列图像不仅蕴含图像内部的语义信息，图像之间也存在着语义依赖关系（即序列图像的“序列性”）。对序列图像的处理是计算机视觉领域重要的研究课题，在医学影像辅助诊断、遥感图像处理以及视频分析等领域都有着广泛的应用。本文以序列图像的识别与检测算法为研究目标，将序列图像展开为空间和时间序列图像，致力于探索如何高效地挖掘图像内部和图像之间的深层次语义关联信息。从基于立体空间序列图像的肺结节检测、基于平面空间序列图像的道路提取和基于时间序列图像的变化检测等三个具体的视觉任务出发，对不同分布的序列图像进行分析，通过改进视觉注意力机制，建立了具有更强的序列关联性表达能力的深度网络模型。本文的主要研究内容和贡献包括以下几个方面：

1. 对于立体空间序列图像中的肺部 CT 图像，针对其立体空间维度较高的难点，提出了切片关联注意力网络进行肺结节的检测。受到医生临床诊断肺结节方式的启发，设计了一种切片分组非局部模块，将切片维度分组的思想引入自注意力机制中，来充分学习 CT 图像中的立体空间序列信息。三维区域候选网络对肺结节的检测通常会带来大量的假阳性样例，设计了基于多尺度特征图的假阳性抑制模块来进一步优化检测的结果。此外，提出了医学影像领域目前为止规模最大的肺结节检测数据集 PN9，与之前的肺结节检测数据集相比，其在数据规模、种类多样性、图像丰富度和检测困难程度上都有了较大的提高。通过在不同数据集上的大量实验，充分验证了所提出的切片关联注意力网络能够有效提高肺结节检测的性能。

2. 对于平面空间序列图像中的道路遥感图像，针对其它地物会对道路进行遮挡的问题，提出了拓扑连通注意力网络从而直接提取出连通性较好的道路。考虑到道路在平面维度上连续分布并呈现出跨度大且细长的形状，设计了条形卷积模块，其利用水平、垂直、左对角线和右对角线等四种不同方向的条形卷积来学习道路的长距离依赖信息，同时抑制不相关区域对特征学习的干扰。此外设计了连通性注意力模块来探索相邻像素之间的连通关系，其能够缓解建筑物或树木等对道路的遮挡问题，提高道路的拓扑正确性。通过在两个公开数据

集上的大量实验，验证了拓扑连通注意力网络在保证道路连通性方面的有效性。

3. 对于时间序列图像中的双时序高分辨率遥感图像，提出了差异感知注意力网络来同时进行建筑物分割和多级别变化检测。为了探索不同时序图像中能够反映出差异性变化模式的通道，设计了双时序聚合模块，其能够同时学习全局变化信息。此外，考虑到图像中存在的不同级别变化，进一步设计了差异注意力模块来探索多级别变化之间的局部联系，并提高对不同等级变化的判别能力。在大规模建筑物变化检测数据集上的大量实验表明，相比于其他的方法，本文提出的差异感知注意力网络具有较大的优越性。

4. 基于时间序列图像中的双时序高分辨率遥感图像，进一步提出了基于全局差异与局部注意力的网络模型进行变化检测。结合卷积神经网络能够更好的提取图像的低阶细节信息和 Transformer 可以对长距离依赖关系进行建模的优势，采用混合两者的架构作为编码器来提取图像特征。设计了全局差异模块，来学习全局变化信息并提高对图像中所有像素的整体理解。此外，设计了局部门控注意力模块，来学习局部变化差异并增强对双时序图像间多级别变化的判别能力，其利用门控自注意力机制来学习相邻变化敏感特征块之间的局部依赖性。通过大量实验，验证了此模型应对变化检测任务是有效的。

关键词： 空间序列图像；时间序列图像；注意力机制；肺结节检测；道路提取；变化检测

Abstract

Sequence image is a collection of images arranged in a specific temporal or spatial order. Unlike ordinary images, sequence images not only contain semantic information within the images, but also have semantic dependencies between images (*i.e.,* "sequentiality"). The processing of sequence images is an important research topic in computer vision, which has a wide range of applications in medical image aided diagnosis, satellite imagery processing, and video analysis. This paper is devoted to exploring how to efficiently learn the deep-level continuous sequence information, which takes the recognition and detection algorithm of sequence-image as the research goal, and expands the sequence-image into spatial and temporal sequence-image. Based on three specific tasks, *i.e.,* lung nodule detection with stereo-spatial sequence-image, road extraction with planar-spatial sequence-image, and change detection with temporal sequence-image, models with a stronger ability to express sequence correlation are built by improving visual attention mechanisms. The main contributions of this paper include:

1. This paper presents a Slice-Aware Network (SANet) for lung nodule detection, which is based on CT images in stereo-spatial sequence-image to handle its high spatial dimension. Inspired by the diagnosis way of doctors, a Slice Grouped Non-Local module (SGNL) is designed to introduce the idea of slice grouping into self-attention mechanism, which can learn the stereo-spatial sequence information in CT images. The detection of nodules by 3D Region Proposal Network usually brings many false positive samples, thus a False Positive Reduction module (FPR) based on multi-scale features is designed to optimize the detection results. This paper proposes a new dataset for lung nodule detection, namely PN9, which has been improved in terms of data size, variety, image richness, and detection difficulty compared with previous datasets. Extensive experiments verify that SANet can improve the performance of lung nodule detection.

2. This paper proposes a Connectivity Attention Network (CoANet), which can alleviate the occlusions and extract well-connected roads from satellite imagery in planar-spatial sequence-image. Considering that roads are continuously distributed in the pla-

nar dimension and exhibit a slender shape, a Strip Convolution Module (SCM) is developed to learn the long-range context information of roads using four strip convolutions with different directions, such as horizontal, vertical, left diagonal, and right diagonal. It also suppresses the interference on feature learning from irrelevant regions. A Connectivity Attention module (CoA) is designed to explore the connectivity between neighboring pixels, which alleviates the occlusions of buildings or trees and improves the topological correctness of roads. Extensive experiments on two public datasets demonstrate the effectiveness of the proposed CoANet in ensuring road connectivity.

3. This paper proposes a Difference-Aware Attention Network (D2ANet) for simultaneous building segmentation and multi-level change detection, which is based on the dual-temporal satellite imagery in temporal sequence-image. To explore the channels in different images that can reflect the differential patterns of change, a dual-temporal aggregation module (DTA) is designed to learn the global change information. Considering the change of different levels in the images, a difference-attention module (DA) is developed to exploit the local correlations among the multi-level changes, which improves the performance to identify different change scales. Extensive experiments on a large-scale building change detection dataset show that the proposed D2ANet has greater advantages compared to other change detection methods.

4. This paper presents a model based on Global Difference and Local Attention (GDLA) for change detection with the dual-temporal satellite imagery in temporal sequence-image. Combining the advantages of CNN that can extract low-level detail information of images and Transformer can model long-range dependencies, a hybrid CNN-Transformer architecture is adopted as the encoder to extract image features. A Global Difference module (GD) is designed to capture the global change information and improve the overall awareness of all pixels in images. A Local Gated Attention module (LGA) is proposed to learn the local variation and enhance the discrimination of multi-level changes between dual-temporal images. It uses a gated self-attention mechanism to learn the local dependencies between adjacent change-sensitive feature maps. Extensive experiments demonstrate that our GDLA is effective for change detection.

Key Words: Sequence-image of spatial; sequence-image of time; attention mechanism; lung nodule detection; road extraction; change detection

目录

图目录

表目录

第一章 绪论

第一节 研究背景和意义

随着人类需求的不断提高以及数据采集设备和存储硬件的逐渐优化，各个领域产生了大量的图像数据，例如医学图像、遥感图像、拍照及摄影、监控视频等[1-4]。日常中的图像数据大多都呈现零散分布的状态，彼此之间没有很强的关联关系。如果一组图像按照一定的空间或时间顺序依次排列，并且能够随着空间的移动或时间的推进而有规律的变化，那么这些图像数据可以被称为序列图像。目前，序列图像处理已成为计算机视觉领域一个研究热点和难点，在多个领域都有着广泛的应用，例如医学辅助诊断[5]、遥感图像处理[6]、灾害评估[7]以及视频分析[8]等。

根据图像序列所在域的不同，序列图像可分为空间序列图像和时间序列图像。众所周知，自然图像可以表示为三维数组，第一、二维分别表示图像的高、宽，而第三维则表示图像的通道数。其中通道数一般为 3，即 R、G、B 三个通道。空间序列图像是指在图像某一维度连续排列的一组图像集合，根据其排列维度的不同，可分为立体空间序列图像和平面空间序列图像。立体空间序列图像是指在通道维度（即第三维度）上包含连续序列信息的一组图像，例如医学图像中的电子计算机断层扫描（Computed Tomography，CT）图像以及遥感图像中的高光谱图像等。与传统的仅包含 R、G、B 三个通道维度的自然图像相比，这类图像通常包含更多的、具有相互关联的通道数。平面空间序列图像是指在平面维度（即第一、二维度）上有连续信息的图像，它可以看成多个顺序排列的子图像的集合，例如道路遥感图像。此外，还有一些图像是在一段时间内按照特定时间间隔获取的，它们虽然在图像维度上不存在序列信息，图像之间却蕴含着时间变化信息。本文将这类图像的集合称为时间序列图像，即在时间维度上按顺序排列的一组图像，例如多时序的图像以及视频数据等。无论是空间序列图像，还是时间序列图像，都可以看成一组存在着某种关联关系的图像（或子图像）的集合，本文将这种关联关系称为序列关系，也称为“序列性”。是否具有序列性是序列图像区别于普通图像及其集合的本质特征。

在计算机视觉领域，图像的识别与检测是两个重要且基础的的任务，其旨在从图像中识别出感兴趣的目标，或者检测出图像中物体的具体位置及类别。目前主流的图像识别与检测模型一般都是基于机器学习算法实现的，其中基于传统机器学习算法的模型通常通过建立较为复杂的特征工程，使模型具有识别图像中复杂模式的能力[9-11]。这类算法一般通过分析图像的颜色、形状、纹理等信息来手工提取特征，然而在处理序列图像时，却存在很多局限，例如：

- 序列图像中蕴含丰富且复杂的空间或时间信息，而手工特征提取方式一般基于研究者的先验知识，有一定的主观性和局限性，因此难以捕捉完整的时空域信息。
- 序列图像的数据维度和规模一般都比较大，而手工特征提取过程不仅耗时，而且可迁移性较差，因此难以适用于序列图像。
- 在一组序列图像中，不仅单张图像中蕴含大量的语义信息，不同图像之间还存在广泛的语义依赖关系，而手工特征提取方式一般仅能捕捉到图像的底层信息，因此难以对序列图像内部以及图像之间所存在的大量语义信息进行建模。

总而言之，传统机器学习算法难以挖掘出序列图像的深层次特征，从而导致欠佳的识别与检测结果。随着计算机硬件和深度学习的发展，很多研究利用深度神经网络，尤其是卷积神经网络（Convolutional Neural Network，CNN）来进行序列图像的识别与检测[12-13]。相比于传统机器学习算法，卷积神经网络不再需要人为地构造特征模式，能够自动从图像中学习到更高阶的潜在抽象特征，这使其在处理序列图像时具有天然的优势。然而，目前的研究大多将序列图像视为多张图像的简单组合，并直接针对单张图像进行处理，利用卷积神经网络提取单张图像的特征，来完成识别和检测等任务[7,14-15]。这些方法严重忽略了序列图像的“序列性”，而由于“序列性”表征了图像之间的语义关联信息，因此对“序列性”的忽略是阻碍当前研究无法实现序列图像处理最优性能的一个关键问题。如何将序列图像的序列信息（包括空间序列信息和时间序列信息）结合起来，设计一个新的序列图像特征提取方法，是进一步提升序列图像识别与检测性能的重要手段。

在学习图像内部的语义依赖关系方面，注意力机制（Attention Mechanism）是一种常用且有效的方式。人类的视觉系统在接收到复杂场景的输入信息时，能够快速地将注意力放于感兴趣的区域，同时忽略掉其他无关区域，从而能够

高效地对场景进行分析理解并获取重要信息[16-17]。模仿人类视觉系统的这一机制，研究者们将注意力机制引入到深度学习中[18-21]，使得神经网络可以关注图像特征中的关键信息，同时抑制次要信息。对于序列图像而言，数据的维度和规模一般要大于自然图像，其中也会存在较多的冗余信息。直接利用卷积神经网络提取特征，会将冗余信息也考虑进来，增大模型的训练时间，并导致模型的泛化能力降低。注意力机制可以有效地缓解这一问题，其能够对序列图像中的重要信息进行筛选，增强神经网络所提取特征的表征力。此外，注意力机制中的自注意力能够建模输入特征图中不同部分之间的相关性，进而编码整张图像的语义依赖信息。然而，当前的注意力机制大多仅能对单张图像进行语义建模，无法充分地探索序列图像中不同图像之间的语义关联，即序列性。因此，如何利用注意力机制挖掘序列图像的时空域依赖信息，并提升序列图像识别和检测的性能，是当前亟待解决的一个重要研究方向。

本文以序列图像的识别与检测算法为研究目标，将序列图像展开为空间和时间序列图像，致力于探索如何高效地挖掘图像内部及图像之间的深层次语义关联信息。从基于 CT 图像的肺结节检测、基于道路遥感图像的道路提取和基于双序列图像的变化检测三个具体的视觉任务出发，对不同分布的序列图像进行分析，并通过改进视觉注意力机制，建立了具有更强的序列关联性表达能力的深度网络模型，实现了更好的识别与检测性能。其中，肺部 CT 图像在通道维度包含多张连续的断层扫描切片，不同切片间关联性很大，共同组成了肺部的组织，属于立体空间序列图像。肺结节检测是指将 CT 图像中肺结节的三维位置、类别等检测出来，其能够协助医生对 CT 图像的解析，进而有效地防治肺癌。道路遥感图像在平面维度包含道路连续不断的信息，可以看做多张顺序排列的道路子图像（不同子图像共同形成了连通的道路）的集合，属于平面空间序列图像。道路提取旨在将道路遥感图像中的道路识别出来并保证其连通性，它是快速更新道路地图网络的重要手段。双时序图像包含相同区域两个不同时间的图像，属于时间序列图像。变化检测是一项用于识别不同时序图像间差异的技术，其在异常检测、环境监测和灾害评估等领域有着广泛的应用。

第二节　国内外发展现状

本节将对近年来国内外与本文研究内容相关的代表性研究工作进行回顾总结，主要介绍了三个方面：空间序列图像的识别与检测、时间序列图像的检测

以及注意力机制。

1.2.1 空间序列图像的识别与检测

空间序列图像是指图像在空间上根据一定的规则依次排列，其中包含了物体在连续空间中的关联序列信息。面向不同的空间维度，空间序列图像可以分为立体空间序列图像和平面空间序列图像。相比于传统的自然图像，立体空间序列图像通常包含更多的通道序列信息，而平面空间序列图像会包含平面维度上的连续关联序列信息。本文选取立体空间序列图像中的 CT 图像进行肺结节检测的任务，同时选取平面空间序列图像中的道路遥感图像进行道路提取的任务。本小节将对这两个任务的相关研究工作进行介绍。

1.2.1.1 基于 CT 图像的肺结节检测

根据人体各个器官部位对 x 射线的透过率不同的特点，CT 利用 X 射线对人体部位进行断层扫描，经过数据处理产生检查部位的横断面图像[22]。肺结节是肺部的一种病灶[23]，区别于肺部其他连续管状结构的血管、支气管等组织，肺结节通常呈现出孤立的球状结构。由于肺结节可能是肺癌的早期表现形态[24]，及早地基于 CT 图像对肺结节进行诊断可以有效的预防肺癌。不同于一般的二维目标检测，肺结节检测是一个利用三维 CT 图像进行三维目标检测的问题，由于其临床价值较大，近年来吸引了研究者们越来越多的重视。

（1）基于手工设计特征或形态学操作的肺结节检测方法。早期的肺结节检测方法大多依赖于手工设计的特征或形态学操作。Messay 等人[25]介绍了一种全自动肺部分割算法，其结合了形态学操作和强度阈值来分割出候选肺结节。Jacobs 等人[26]采用了形状、纹理和强度等属性，并设计了一系列上下文特征来检测半实性肺结节。Duggan 等人[27]提出了一种基于全局分割的候选肺结节检测方法，其将基于简单规则的滤波和平均曲率最小化相结合。Lopez 等人[28]利用人工设计的滤波器来筛选 CT 图像中可能存在的肺结节，但是这些滤波器的设计需要丰富的医学知识。Gupta 等人[29]首先利用泛洪填充算法和形态学闭运算来进行肺部分割，之后通过多级阈值处理操作来进行肺结节的检测。Wang 等人[30]结合灰度增强与球形增强滤波器来识别候选肺结节，并利用形状约束的 Chan-Vese 模型[31]来去除其中的假阳性结节。一些研究[32-34]首先设计一系列特征来表示肺结节，例如形状、纹理和光谱等特征，再基于这些特征利用 SVM 分类器分类出结节和非结节。然而，这些方法很难适用于检测复杂部位的结节，特别是呈高

度血管附着性的结节。

（2）基于二维卷积神经网络的肺结节检测方法。随着深度学习的不断发展，利用卷积神经网络（CNN）的模型，例如 Faster R-CNN[35]、SSD[36]、YOLO[37]等陆续被提出用于目标检测任务并取得了较好的性能。基于 CNN 的方法也被应用于肺结节检测中，这些方法可以被分为两类：基于二维 CNN 的方法和基于三维 CNN 的检测方法。对于二维 CNN，Setio 等人[38]提出了用于肺结节检测的多视图卷积神经网络，其输入是一组来自不同截面的二维切片，并利用设计的特征融合方法实现对 CNN 输出的特征组合。Fu 等人[39]利用肺部 CT 图像、结节增强图像和血管增强图像等三种类型的图像进行训练，并从每种图像中提取出 9 个不同方向的二维图像块用于肺结节的检测。Jiang 等人[40]利用 Frangi 滤波器对 CT 图像的图像块进行增强，之后基于两组图像，利用四通道的 CNN 网络进行肺结节的检测。George 等人[14]基于 YOLO[37]来检测肺部 CT 图像中的结节，但是其整体的检测性能较差。Ding 等人[41]将反卷积结构加入到 Faster R-CNN 中[35]，并进一步利用轴向的切片进行肺部结节的检测。Xie 等人[15]将两个区域生成网络和一个反卷积层添加到 Faster R-CNN[35]中来检测候选肺结节，同时利用基于 2D CNN 的分类器从候选结节中辨别出真阳性结节。这些方法通常需要经过后处理来将二维的检测候选框整合为三维的候选框，使得其效率低下并且会影响肺结节检测的准确性。

（3）基于三维卷积神经网络的肺结节检测方法。考虑到 CT 图像的三维属性，近年来越来越多的研究采用了基于三维 CNN 的方法。Dou 等人[42]提出了一种基于三维 CNN 的肺结节检测方法，其引入多级纹理信息编码策略来处理结节差异大和其他组织噪声干扰的问题。Li 等人[43]提出了一种带有编码器-解码器结构的三维 CNN 进行肺结节的检测，并采用动态尺度交叉熵来减小假阳性率，利用挤压-激励模块[44]来充分地获取通道间的依赖性。Zhu 等人[45]提出了一种基于 3D 双路结构的 3D Faster R-CNN 用于肺结节的检测，其中采用了类似于 U-Net[46]的结构来有效地学习结节特征。为了推动深度学习在肺结节检测任务中的发展，Liao 等人[47] 采用了 3D 区域候选网络（Region Proposal Network，RPN）结构，并引入 Leaky Noisy-OR Gate 模块选择置信度较高的前五个结节来评估癌症概率。Kim 等人[48]提出了一种多尺度渐进融合 CNN，对多尺度的输入采用循序渐进的特征提取策略，这种方式能够有效地减少假阳性结节。Ozdemir 等人[49]介绍了一种端对端的概率诊断系统，其利用计算机辅助检

测模块来实现对可疑肺结节的检测，应用计算机辅助诊断模块实现对患者恶性肿瘤的分类。Harsono 等人[50]提出了一种肺结节检测及分类模型 I3DR-Net，其结合了膨胀三维 CNN I3D[51]和 RetinaNet[52]，并应用了特征金字塔网络（Feature pyramid network，FPN）结构。Song 等人[53]设计了一种 3D 中心点匹配检测网络（Center-Points Matching Detection Network，CPM-Net）来进行肺结节的检测，此方法可以自主预测肺结节的位置和类别，同时不需要人工设计的锚参数。Zhu 等人[45]将 3D 双路径网络（Dual-Path Network，DPN）[54]和类似于 U-Net[46]的编码器-解码器架构用于 3D Faster R-CNN 网络，并采用梯度提升机（Gradient Boosting Machine，GBM）进行肺结节的分类，这些策略有效提高了肺结节检测的性能。Khosravan 等人[55]利用密集连接的三维 CNN 进行肺结节的检测，其不需要任何后处理或者人工指导来改进检测的结果。相比于二维 CNN，三维 CNN 的参数量更大，导致后者需要更长的时间和更多的 GPU 显存来进行训练，但是其也在针对 CT 图像的肺结节检测中获得了比二维 CNN 更好的结果[56]。

1.2.1.2 基于道路遥感图像的道路提取

道路是一种很重要的地物目标，其外观表现为细长且连续的条带。道路网络的创建是多个应用领域基础但不可或缺的前提步骤，包括：自动驾驶、城市规划、车辆导航和地理信息的更新等。近年来遥感影像获得了快速地发展，其不仅能够呈现出道路的几何纹理，也可以提供多个时期甚至实时的影像，这为道路的快速更新提供了必要的支持。从遥感影像中提取道路[57-59]已经成为目前主流的更新道路的方式。

（1）基于人工设计特征的道路提取方法。传统的道路提取方法通常利用手工设计的特征并定义一些特定的标准来匹配[60-62]。He 等人[63]提出了一种基于颜色的道路检测算法，其结合了灰度图像的边界估计和彩色图像的道路区域提取结果。Zhang 等人[64]引入了角度纹理的描述符，并使用模糊逻辑分类器来识别道路分割块。Laptev 等人[65]基于多尺度道路图像进行道路提取，其结合使用了蛇形几何约束对道路边缘进行检测。Chai 等人[66]利用连接点处理来恢复航空图像和视网膜图像中的线网络。Wegner 等人[67]提出了一种用于道路网络提取的高阶条件随机场（Conditional Random Field，CRF）模型。一些研究利用边缘检测算法来进行道路的提取，这些方法主要利用了灰度值的阶跃性变化特点。具体而言，曾等人[68]采用了基于 Canny 边缘检测的算法，Gaetano 等人[69]结合 Canny 算子和图割理论来进行道路的提取。Stoica 等人[70]使用 Gibbs 点处理框架

从遥感图像中提取道路，滕等人[71]进一步将 Gibbs 局部抽样和 Canny 算子、高斯滤波等结合起来进行道路提取。这些方法通常利用像素的灰度或者光谱特征来提取道路，比较适用于道路类型单一的场景，但不能很好的应对建筑、树木以及阴影对道路的遮挡问题。为了缓解建筑和树木的遮挡对道路提取的影响，Li 等人[72] 利用二叉划分树（Binary Partition Tree）来分层次的表示道路区域，基于这些道路兴趣区，结合几何特征和结构特征来提取道路。Alshehhi 等人[57]首先基于形态学滤波提取特征以增强道路和非道路像素之间的对比度，再利用基于图的分割算法提取道路并通过后处理来去除不规则的道路块。这些方法通常需要复杂的处理步骤，效率低下，不适合用于大面积区域道路的提取和更新。

（2）基于卷积神经网络的道路提取方法。随着深度学习的发展，具有编码器-解码器架构的卷积神经网络[73-77] 被提出并被证明在图像语义分割任务中是有效的。一些研究[78-80] 采用基于 CNN 的模型将道路提取任务当做分割问题来处理。Mnih 等人[81]通过使用在图形处理上实现的神经网络来检测道路。Cheng 等人[82]提出了一个级联的端到端卷积神经网络，利用超高分辨率遥感图像来同时完成道路分割和道路中心线提取两个任务。Panboonyuen 等人[83]提出了面向道路分割的 CNN 框架，同时设计了景观度量指标来减少错误分类的道路像素，并利用条件随机场来锐化提取的道路。Mendes 等人[84]基于全卷积神经网络（Fully Convolutional Network，FCN）利用上下文语义信息来提取道路。U-Net[46]和 LinkNet[85] 是被广泛使用的针对语义分割任务的编码器-解码器结构网络，它们的变体模型也被提出并用于学习细长的道路特征。Zhang 等人[79]结合残差学习和 U-Net 进行道路区域的提取。在 CVPR DeepGlobe 2018 道路提取挑战赛中，Zhou 等人[86]提出的 D-LinkNet 获得了第一名，其基于 LinkNet[85]结构并在中心部位加入了空洞卷积来扩大网络的感受野。这些方法通过使用跃层连接或多尺度特征图，能够融合低阶细节和高阶语义信息，从而获得较好的道路分割结果，但是由于在空间维度上损失了信息的连续性而不能保证道路的连通性。

道路的连通性是最重要的道路属性之一，也是车辆导航、自动驾驶和路线规划等应用所必需的条件。近年来针对道路提取的研究越来越关注道路的连通性，目前所采用的方法可分为三类：结合道路分割和后处理步骤、迭代优化道路和多模态数据融合。Wegner 等人[87]首先将航拍图像分割成超像素，并使用最短路径算法将似然性高的候选道路相连接。Máttyus 等人[88]首先使用具有编码

器-解码器结构的网络模型获得航拍图像的道路分割图，由于分割结果不能保证道路的连通性，因此他们引入了后处理步骤，通过使用最短路径算法来推理缺失的道路。在这些方法中，道路连通性是利用后处理来实现的，这种方式不适用于遮挡严重、道路外观模糊和道路密度高的区域。为了直接获得连通性更好的道路提取结果，Mosinska 等人[89] 利用 U-Net 网络结构，同时结合像素级损失和拓扑感知损失对道路轮廓进行迭代细化。Bastani 等人[58]提出了 RoadTracer，这种方法基于卷积神经网络，构建了决策函数来引导迭代搜索过程，其能够从航拍图像中自动提取道路网络。Batra 等人[59]提出了一个堆叠的多分支模块来有效地利用道路分割和方向学习任务之间的关联信息。为了提高道路的连通性，他们还设计了一种连通性细化方法，能够迭代优化预测道路网络的拓扑结构。然而，由于迭代步骤通常比较耗时，这些方法需要较长的时间来进行训练。

一些研究通过将多级道路特征或其它地理数据相整合，以获得连通性更好的道路提取结果。Liu 等人[90]提出了 RoadNet 来同时预测道路表面、边缘和中心线，其集成了多级特征以便于处理各种场景中的道路并提高道路的连通性。Li 等人[91]设计了一种新颖的框架来有效地整合多种道路形状特征，包括点、边缘和区域特征，同时引入方向感知注意力模块以进一步提高道路连通性和道路识别精度。雷达数据（Light Detection and Ranging，LiDAR）[92-94]和 GPS（Global Positioning System）轨迹[95-98] 也被应用于推断道路网络。Sun 等人[99]将行人的 GPS 数据与遥感图像相结合来提取道路，实验结果表明其优于单独使用 GPS 数据或图像的模型。LiDAR 和 GPS 数据可以帮助改善道路的连通性，特别是在有建筑或树木遮挡的区域。然而，收集足够覆盖大区域的 LiDAR 和 GPS 数据具有比较大的挑战性，同时这些数据的预处理通常也比较复杂。

1.2.2 时间序列图像的检测

时间序列图像是指图像按时间顺序依次排列，并且能够随着时间的发展而不断变化。针对时间的长短，时间序列图像可以分为短时间序列和长时间序列。短时间序列图像通常包含几帧不同时序的图像，而长时间序列图像包含的帧数更多且时间跨度更大。本文选取双时序遥感图像，进行建筑物的变化检测。遥感图像能够直观反映出地物的信息表征和状态，随着遥感成像技术的发展，遥感影像逐渐呈现出高空间和高时间分辨率的特点，从而为人类了解地球环境以及监测地表动态变化提供了必要的信息支持。基于遥感图像进行的变化检测，是为了比较分析同一区域内不同时序的两幅图像，进而获得不同时序地物的变

化信息。这一技术在多个领域得到了应用，包括：灾害评估[100]、环境监测[101]、城市化评价[102] 和资源管理[103]等。目前的人工目视解译方法需要依靠专业人员分析双时序遥感图像来检测变化，其效率较低且容易受到主观条件的影响，不能满足对大区域环境进行动态监测的需求。为了加快这一过程，基于双时序遥感图像的自动变化检测方法近年来受到了越来越多的关注[100,102,104]。

（1）传统的变化检测方法。早期的变化检测研究通常基于像素，对不同时序的图像进行直接比较，包括阈值比较[105]、图像差值[106]、图像比值[107]和回归分析[108]等方法。由于没有考虑像素间的关联性，这些方法通常会产生较多的噪声。为了更加充分地利用遥感图像中的光谱信息，采用图像变换的方法被用于变化检测，例如：变化向量分析[109]、主成分分析（Principal Component Analysis，PCA）[110]、独立成分分析（Independent Component Analysis，ICA）[111]和多元变化检测[112]等。为了提高变化检测方法的鲁棒性，一些研究将人工设计的特征和传统机器学习分类算法相结合，包括马尔科夫随机场（Markov Random Field）[113]、支持向量机（Support Vector Machine，SVM）[114] 和条件随机场（Conditional Random Field）[115]等算法。Nemmour 等人[116]提出了一个结合模糊积分支持向量机和吸引子动量的框架，用于地表覆盖的变化检测。Im 和 Jensen[117]利用多时序高空间分辨率图像，提出了一种三通道邻域相关性图像（Neighborhood Correlation Image，NVI）方法。Zhang 等人[118]引入了多尺度不确定性分析的方法，利用支持向量机对不确定的区域进行迭代分析。在此基础上，Tan 等人[119]利用多尺度不确定性区域分析结合多个分类器，介绍了一种基于对象的方法，进一步提升了变化检测的性能。然而，这些方法的步骤繁琐且效率较低，一般只适用于处理小范围的区域。

（2）基于卷积神经网络的变化检测方法。随着深度学习的发展，基于 CNN 的模型被逐渐应用到变化检测任务中。Caye 等人[120]设计了一种基于迭代训练方案的全卷积神经网络，用于从噪声图像数据中检测变化。Papadomanolaki 等人[121]基于多时态的高分辨率遥感图像，采用循环神经网络（Recurrent Neural Network，RNN）和 FCN 相结合的方式，进行像素级的变化检测。Wu 等人[122]提出了一种用于多时序高分辨率遥感图像的核主成分分析变化检测方法。孪生神经网络（Siamese Neural Network）是基于两个神经网络构建的耦合结构。其以两个图像样本作为输入，并输出它们的特征表征以比较两个样本的相似性，这一特点使得孪生神经网络在变化检测任务中十分有效。Daudt 等人[123]基于高分辨

率遥感图像和多光谱遥感图像，提出了两个孪生模块，并将其加入到全卷积神经网络中进行变化检测。Liu 等人[124]提出了一个深度孪生网络，基于两幅图像进行变化检测。Wang 等人[125]提出了一个具有混合特征提取模块的深度孪生神经网络，其使用决策模型根据特征差异来检测变化。Chen 等人[126]结合 CNN 和 RNN 的优势，提出了一种孪生卷积循环神经网络进行变化检测。Du 等人[127]采用两个对称的卷积网络来学习双时序图像的特征，之后采用慢特征分析（Slow Feature Analysis，SFA）模块来突出转换特征中的变化部分。然而，由于孪生神经网络需要使用两个网络来对不同时序的图像进行处理，从而会引入更多的参数量。为了更加充分的利用双时序图像中的差异信息，一些研究引入了注意力机制。Liu 等人[128]在孪生神经网络的基础上，加入了双重注意力模块以同时完成变化检测和建筑物分割。Zhang 等人[129]利用空间注意力和通道注意力模块对双时序遥感图像的特征从空间和通道维度上进行优化，但是其缺乏对长程依赖关系进行建模的能力。Chen 等人[130]引入自注意力机制来获取更具有判别性的特征，但是其中的矩阵运算会带来较大的计算量。Chen 等人[131]引入 Transformer 来更好的建模双时序图像的长程依赖信息，然而这种将图像特征划分成特征块再输入 Transformer 的方式会造成全局结构信息的损失。

当自然灾害发生时，快速地检测灾害造成的变化对实施有效的灾害响应来说至关重要，因此近期的变化检测研究更加关注由灾害引起的变化。Xu 等人[132]将灾害前后的遥感图像联系起来，用于检测在地震中受损的建筑物。Zhu 等人[133]提出了一种多级实例分割网络，从航拍视频中检测损毁的建筑物。Ji 等人[134]采用 CNN 从地震后的遥感图像中检测倒塌的建筑物。Duarte 等人[135]提出了一个结合残差连接和空洞卷积的 CNN 框架，基于无人机图像对建筑物的损毁进行分类。Rudner 等人[136]利用全卷积神经网络，通过融合多分辨率、多传感器和多时相遥感图像，对洪水淹没的建筑进行分割。这些方法通常只用于检测由单一灾害类型造成的变化，不能很好的应对现实中复杂的自然灾害。Gupta 等人[7]发布了一个包含 19 种不同自然灾害和 4 种损毁等级的大规模数据集 xBD，用于建筑物的分割和变化检测。他们进一步提出了一个基准模型，其采用 U-Net 结构[46] 进行建筑物的分割，同时利用 ResNet-50[137]来对损毁等级进行分类。Weber 等人[138]在 Mask R-CNN[139]的基础上添加了特征金字塔网络模块[140]和语义分割模块来进行建筑损毁评估。Shen 等人[141]提出了一个基于 CNN 的两阶段框架，其采用 U-Net 进行建筑物分割，之后利用两分支 U-Net 进

行变化检测。然而，这些方法将建筑分割和变化检测分为两个独立的阶段，使得模型通常需要复杂的步骤和更多的时间来训练。

1.2.3 注意力机制

人类的视觉系统可以同时接收大范围区域的输入信息，但是其能够很快地将注意力放于感兴趣的区域，同时忽略掉其他无关区域，从而高效地获取重要信息。模仿人类视觉系统的这一机制，研究者们将注意力机制引入到深度学习中，使得神经网络可以自动学习图像特征中的主要信息，同时抑制次要信息。对于序列图像而言，由于数据的维度比自然图像要大，直接利用卷积神经网络难以挖掘出序列图像的内部关系，而采用注意力机制能够有效地学习序列关联性，并进一步提升模型处理序列图像的性能。本节对注意力机制的研究现状进行了总结，从注意力机制的工作机理出发，将其分为选择性注意力机制和自注意力机制两个类别。

1.2.3.1 选择性注意力机制

选择性注意力通过预测特征各个部分对于具体任务优化目标的重要程度，得到注意力权重，再根据注意力权重对特征中的不同部分进行增强或抑制。为了突出图像中的重要区域，Jaderberg 等人[142] 设计了空间转换器，使得神经网络能够对特征图进行空间变换。对于卷积神经网络提取的特征，不同的通道通常会呈现出不同的信息。Yu 等人[143]设计了一种通道注意力模块，其将相邻阶段的特征图拼接并计算权重向量，利用得到的权重向量对浅层的特征图通道进行增强或抑制。Hu 等人[44]提出了一种挤压-激励网络（Squeeze-and-Excitation Network，SENet），其中的挤压-激励模块是一种通道注意力机制，能够根据特征图通道对任务目标的重要程度对其进行重校准。具体而言，SE 模块对特征图通道按顺序进行挤压（Squeeze）和激励（Excitation）操作来获得通道注意力权重，再利用此注意力权重对特征图的不同通道进行重校准以便获取关键的通道信息。SE 模块仅仅激励了特征图的通道，而忽略了空间维度，而对很多任务来说，像素级的空间信息也非常重要。因此在 SE 模块的基础上，Roy 等人[144]进一步提出了空间和通道挤压-激励模块（Spatial and Channel Squeeze & Excitation，scSE），其对特征图沿着通道和空间维度进行重校准，并将两者的输出进行融合。此外，Woo 等人[145]提出了一种卷积块注意力（Convolutional Block Attention Module，CBAM）来同时学习通道维度和空间维度的重要信息，其沿着通道和

空间维度计算注意力图，然后将注意力图乘到输入的特征图上来对特征图进行优化。CBAM 是一个通用的模块，其可以轻松整合到卷积神经网络的架构中。

1.2.3.2 自注意力机制

自注意力机制是为了增强输入特征图中不同部分之间相关性的注意力方法。其最早是在自然语言处理（Natural Language Processing，NLP）领域被提出的[146]，由于性能卓越，迅速被应用于计算机视觉领域。Wang 等人[18]针对视频分类和目标检测任务，提出了一种基于自注意力机制的非局部（Non-Local）模块，其通过计算输入特征中所有像素的特征加权和来得到一个位置上像素的响应。由于需要将特征中的每个像素与其他像素进行矩阵乘法操作，使得其计算量较大，带来了较高的时间复杂度和 GPU 显存需求。为了改善这个计算量较大的问题，Huang 等人[147]提出了 CCNet（Criss-Cross Network）来学习整张图像的上下文信息。不同于非局部模块对图像中的每个像素生成密集注意力图的方式，CCNet 中的交叉注意力模块将每个像素与其所在行和列的像素生成稀疏注意力图。为了进一步获取密集的上下文依赖，他们在交叉注意力模块的基础上进一步设计了递归交叉注意力（Recurrent Criss-Cross Attention），通过递归使用多次交叉注意力，来生成每个像素的密集注意力图。Zhu 等人[148]提出了不对称非局部操作（Asymmetric Non-Local），其在非局部模块中加入了金字塔采样模块，对键（Key）和值（Value）进行采样，能够在不损失性能的情况下有效的减少计算量和内存消耗。Yin 等人[149]将非局部模块解耦成两项，一个表示两个像素间关系的二元项来学习区域内相似性，一个表示每个像素显著性的一元项来学习显著边界。此外，Mei 等人[5]将分组的思想引入到自注意力机制中，可以有效的减少计算量。Wang 等人[150]在二维自注意力机制的基础上，分解出两个一维自注意力，这种做法能够减少计算复杂度并允许模型在更大甚至全局区域内执行注意力操作。Zhao 等人[151]考虑了两种自注意力的变体，分别是成对自注意力和成块自注意力。其中前者拓展了标准的点乘注意力，而后者能够区分出特定的位置。Fu 等人[152]提出了双注意力网络来将全局依赖关系整合到局部特征中。其设计了两种注意力，分别是位置注意力和通道注意力，前者用于获取图像中任意两个像素的依赖关系，而后者用来学习任意两个通道的联系。

针对机器翻译任务而提出的 Transformer[146]，在很多自然语言处理任务中都获得了较好的性能[153]。Transformer 中的关键结构是自注意力模块，并将其拓展为多头自注意力使得模型能够联合不同表征子空间的信息。考虑到 Transformer

可以学习到输入项之间长程依赖关系的能力，最近有一些研究将 Transformer 结构应用于计算机视觉领域的多个任务中[154-156]。Vision Transformer（ViT）[157]直接使用标准的 Transformer 代替卷积神经网络来进行图像的分类任务。具体来说，他们首先把图像划分为多个图像块，利用线性层将这些图像块转换为特征嵌入序列后，再输入到 Transformer 中。在不使用卷积神经网络的情况下，ViT 在图像分类任务上实现了较好的性能。Touvron 等人[158]提出了数据高效的图像 Transformer （Data-efficient image Transformer，DeiT），其利用基于输入项的蒸馏进一步扩展了 ViT。在目标检测任务中，Carion 等人[159]提出了 DETR（DEtection TRansformer），其利用 Transformer 来推理全局图像上下文信息和目标的关系，并直接输出最终的预测结果。此外，该方法抛弃了多个需要人工设计的部分，例如空间锚点和非极大值抑制操作。由于 Transformer 中的注意力模块在处理图像特征图时的限制，DETR 的训练收敛速度较慢，同时特征图的空间分辨率较低，导致其在检测小尺寸物体时性能较差。为了缓解这些问题，Zhu 等人[160]进一步提出了一种可变形的 DETR，其利用可变形的注意力模块来关注一小组采样位置，从而实现了更快的收敛速度和更好的目标检测性能。由于将 Transformer 中的注意力模块替换为可变形的注意力来处理特征图，因此这种方法能够结合可变性卷积的稀疏空间采样和 Transformer 的长程依赖关系建模能力。至于语义分割任务，Zheng 等人[161]设计了一种纯 Transformer 框架（没有使用卷积和池化操作）来将图像编码为图像块的序列，同时结合一个简单的解码器来完成语义分割。

第三节　研究内容和创新点

序列图像是一组按照特定时空域顺序排列的图像集合，与普通图像不同，序列图像不仅蕴含图像内部的语义信息，图像之间也存在着语义依赖关系（即序列图像的“序列性”）。然而近年来的研究大多是把序列图像当做多张零散图像来处理，这种方式忽略了序列图像中图像之间的语义关联信息，也不能准确地描述序列图像的复杂模式。本文针对序列图像的特点及其研究现状，致力于研究如何高效地挖掘图像内部及图像之间的深层次语义关联信息，进而提高序列图像识别与检测的性能。具体而言，如图 1.1所示，本文以序列图像的识别与检测算法为研究目标，将序列图像展开为空间和时间序列图像进行研究。面向不同分布的序列图像，从肺结节检测、道路提取和变化检测等三个具体的计算

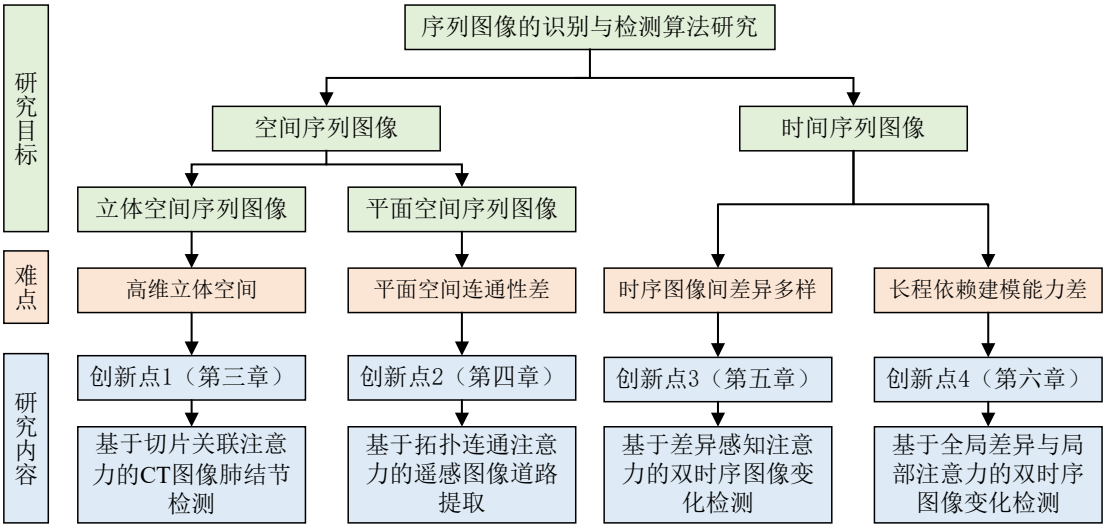


图 1.1 本文的主要研究内容和创新点。

机视觉任务出发，针对各个任务存在的难点，通过改进视觉注意力机制，建立了具有更强序列关联性表达能力的深度网络模型。本文的主要研究内容和创新点包括以下几个方面：

（1）第三章提出了基于切片关联注意力的 CT 图像肺结节检测模型。肺部 CT 图像是由多张连续的断层扫描切片组成的，立体空间维度较高，针对这一特点，本章提出了一种切片关联注意力网络（Slice-Aware Network，SANet）来进行肺结节的检测。医生在诊断肺结节时，会查看多个连续的 CT 图像切片，以识别出区别于其它连续管状结构组织的孤立球状结构的肺结节。受到医生临床诊断方式的启发，本章设计了一种切片分组非局部模块（Slice Grouped Non-Local，SGNL），将切片维度分组的思想引入到自注意力机制中，来学习一组切片内特征图中任意位置和任意通道之间的长程依赖信息。其能够充分学习 CT 图像的立体空间序列信息，同时也可以有效地降低非局部模块的计算复杂度。三维区域候选网络对肺结节的检测具有较高的灵敏度，但其通常会带来大量的假阳性样例，本章设计了基于多尺度特征图的假阳性抑制模块（False Positive Reduction，FPR），以进一步优化肺结节检测的结果。此外，提出了医学影像领域目前为止规模最大、种类最丰富的肺结节检测数据集 PN9（Pulmonary Nodule Dataset with 9 Classes）。该数据集包含了 8798 个肺结节病例、40439 个标注的肺结节以及 9 种不同的类别，同时该数据集覆盖了广泛的收集场景、不同的 CT 设备以及丰富的 CT 图像属性。与之前的肺结节检测数据集相比，本章提出的数据集在数据规

模、种类多样性、图像丰富度和检测困难程度上都有了较大的提高。在 PN9 数据集和公开数据集上，将本章所提出的方法与几种基于二维卷积神经网络和三维卷积神经网络的肺结节检测模型进行了比较，实验结果充分验证了所提出方法在肺结节检测任务中具有较好的性能表现。

（2）第四章提出了基于拓扑连通注意力的遥感图像道路提取模型。道路在平面维度上连续分布，呈现出跨度大、细长且持续连通的形状。在识别道路的研究中，如何缓解其他地物遮挡或复杂交通场景的影响，并利用道路的拓扑序列信息来保证所识别道路的连通性，是一个难点。目前的研究通常使用道路分割和后处理的方式来连通道路残缺块，但是其难以适用于复杂的道路环境且操作步骤繁琐。本章提出了一种拓扑连通注意力网络（Connectivity Attention Network，CoANet），能够直接从高分辨率遥感影像中提取出连通性较好的道路。考虑到道路细长的形状，传统的方形卷积核不能很好的获取道路的线性特征且会引入更多来自相邻像素的噪声，本章设计了一种条形卷积模块（Strip Convolution Module，SCM），其利用水平、垂直、左对角线和右对角线等四个条形卷积从四个不同的方向来学习道路的长距离上下文信息，同时能够抑制不相关区域对特征学习的干扰。此外，为了缓解由建筑物或树木等对道路区域的遮挡问题，本章提出了一种连通性注意力模块（Connectivity Attention Module，CoA）来探索图像中相邻像素之间的关系。其能够预测给定像素与周围八个相邻像素的连通性，使模型能够整合通常在图形网络中学到的信息，进而提高道路拓扑连接的正确性。在两个公开数据集上的大量实验，验证了本章提出的方法与其他道路提取方法相比具有较大的优越性。

（3）第五章提出了基于差异感知注意力的双时序图像变化检测模型。本章基于双时序高分辨率遥感图像，通过探索时序信息来检测不同图像间发生的多级别变化。目前的研究通常采用孪生网络对双时序图像进行特征提取并进一步融合，但是其会带来较大的计算复杂度而且不能很好的检测图像间的多样差异。本章提出了一个差异感知注意力网络（Difference-Aware Attention Network，D2ANet），基于双时序遥感图像同时进行建筑物分割和多级别变化检测。灾前和灾后图像特征之间不同通道可能表达出不同的信息，具体来说，一些通道主要反映了差异性的变化模式，而另外一些通道可能倾向于描述与背景相关的信息。本章提出了一个双时序聚合（Dual-Temporal Aggregation Module，DTA）模块来探索变化敏感的通道，同时学习全局变化信息。此外，考虑到不同的自然灾害

通常会对建筑物造成不同程度的损毁，而充分利用多级别变化之间的相关性可以帮助识别不同级别的建筑物损毁变化。本章进一步提出了一个差异注意力模块（Difference-Attention Module，DA），通过将特征图划分重组，再利用自注意力来获取一个特征立方体组内任意位置和任意通道之间的依赖关系，其中每个组中的小特征立方体都有表示多级别变化的潜力。在大规模建筑物变化检测数据集上的大量实验表明，相比于其他的变化检测方法，本章所提出的差异感知注意力网络具有较大的优越性。

（4）第六章提出了基于全局差异与局部注意力的双时序图像变化检测模型。由于卷积运算的内在局部性，基于卷积神经网络的结构在建模时间序列图像的长程依赖关系上存在不足。因此，本章引入了 Transformer，结合卷积神经网络能够更好的提取图像低阶细节信息的特点和 Transformer 可以对长程依赖关系进行建模的能力，采用混合卷积神经网络和 Transformer 的架构作为编码器来提取图像特征。利用图像块来训练 Transformer 能够有效地减少运算量并加快训练速度，然而只利用图像块对于遥感图像的多级别变化检测是不充分的。一张高分辨率遥感图像会覆盖比较大的区域，其中包含了灾害发生后不同损毁程度的建筑。图像块的尺寸要小于整幅图像，这会限制模型学习不同变化之间长程依赖关系的能力。为了学习全局变化信息并提高对图像中所有像素的整体理解，本章设计了一种全局差异模块（Global Difference，GD）。此外，本章设计了一个局部门控注意力模块（Local Gated Attention，LGA）来学习局部变化差异并增强对双时序图像间多级别变化的判别能力，其利用门控机制来探索双时序特征块中变化敏感的特征序列，进而利用自注意力机制来学习相邻变化敏感特征块之间的局部依赖性。在公开数据集上的大量实验表明，本章所提出的方法应对变化检测任务是有效的。

第四节　本文组织结构

本文围绕序列图像的识别与检测算法研究这一主题，分为七个章节来阐述和讨论相应的研究内容和成果，论文组织结构及各个章节的主要内容如下：

第一章是绪论。本章介绍了序列图像识别与检测的研究背景和意义，阐述了该领域存在的问题和挑战以及国内外发展现状，同时提出了本文的主要研究内容、技术路线和创新点。

第二章是图像识别与检测的相关理论介绍。本章主要概述了本文研究内容

所涉及的计算机视觉两大任务的经典模型，同时也对注意力机制的技术和理论进行了介绍。

第三章是基于切片关联注意力的 CT 图像肺结节检测。本章首先介绍了肺结节检测的研究背景和意义。受启发于医生临床诊断肺结节的方式，本章设计了一种切片分组非局部模块，来充分学习 CT 图像的立体空间序列信息，提高对不同尺寸肺结节的识别性能。最后通过在本章提出的大规模肺结节检测数据集和公开数据集上的实验，对该模型进行了验证。

第四章是基于拓扑连通注意力的遥感图像道路提取。本章首先阐述了道路网络识别与提取的研究背景和意义，并分析了道路提取目前面临的挑战。为了保证所提取道路的连通性，本章设计了一种拓扑连通注意力网络，其能够基于道路形状来获取长程上下文信息，同时探索一定范围邻域内像素间的连接关系，从而提高道路的拓扑正确性。最后在两个公开的道路数据集上对该模型进行了评估。

第五章是基于差异感知注意力的双时序图像变化检测。本章首先介绍了变化检测的研究意义，基于双时序高分辨率遥感图像，提出了一种差异感知注意力网络来同时进行建筑物的分割以及多级别变化检测。利用双时序图像间的联系和差异，来提高对不同级别变化的检测能力。通过与其他方法的对比实验和消融实验，验证了此模型在多任务学习中的有效性和高效性。

第六章是基于全局差异与局部注意力的双时序图像变化检测。本章首先介绍了 Transformer 在图像检测研究中的发展，并分析了卷积神经网络和 Transformer 在学习图像特征时的优劣。结合两者的优势，设计了一种基于全局差异与局部注意力的变化检测方法，来获取双时序图像之间的全局变化模式和多级别变化之间的局部依赖性。实验结果表明，此模型可以增强对图像间不同变化的判别能力。

第七章是总结与展望。本章对本文的序列图像识别与检测研究工作进行了总结，并对未来进一步的研究工作进行了讨论和展望。

第二章 图像识别与检测的相关理论介绍

为了更加充分的理解本文的研究内容以及其中的核心技术，本章介绍了与本文相关的一些理论知识。第一节概述了本文研究内容所涉及的计算机视觉两大任务的经典模型，分别是图像目标检测和图像语义分割；第二节对注意力机制的经典模型及其理论基础进行了介绍；第三节是对本章内容进行的总结。

第一节 相关任务介绍

本文的研究主要涉及到图像目标检测和图像分割两个大的视觉任务，本小节将对这两个任务的研究现状、相关经典模型进行介绍。

2.1.1 图像目标检测

图像目标检测的任务是识别出图像中所有目标的类别，同时预测出目标的位置框。目标检测是计算机视觉领域的一个基础任务，也是解决其他更复杂视觉任务的关键，例如目标跟踪、图像实例分割、自动驾驶以及行人检测等。

2.1.1.1 传统的图像目标检测方法

传统的图像目标检测方法首先在图像上生成大量的目标候选区域（或在图像上滑动窗口），然后提取手工设计的特征，再利用分类器对候选区域进行分类，预测其中是否含有目标以及目标的类别。经典的人工设计特征算子包括：尺度不变特征变换（Scale-Invariant Feature Transform，SIFT）[9]、Haar-like[162]、局部二值模式（Local Binary Pattern，LBP）[10] 和方向梯度直方图（Histograms of Oriented Gradient，HOG）[163]等。

为了增强所提取特征对尺度和图像旋转的鲁棒性，Lowe 等人[9]提出了尺度不变特征变换 SIFT，其利用高斯微分算子在空间中寻找极值点，并计算极值点邻域像素的方向和梯度信息。由于特征是基于局部兴趣点提取的，因此对图像的大小和旋转变换不敏感，同时对光照变换和噪声等也有较高的鲁棒性。SIFT 特征包含的信息量较大，有助于对大数据样本的处理，但同时也带来了较多的时间消耗。在 SIFT 特征的基础上，Bay 等人[164] 提出了 SURF 特征，其引入了盒子滤波器进行运算。在提取的特征点与 SIFT 几乎相同的情况下，计算量更

小，运算速度也更快。Viola 和 Jones[162]提出了 Haar-like 特征来表示图像中不同区域像素和的差值，并利用积分图来获取图像的 Haar-like 特征。此外，他们使用 AdaBoost 算法和其他决策树分类器组成级联分类器，从而快速过滤背景图片并加快对人脸的检测速度。Ojala 等人[10]提出了局部二值模式 LBP，其通过计算中心像素的灰度值与邻域内像素的关系，获取表示局部纹理特征的二进制向量，具有旋转不变性和对灰度变化较强的鲁棒性。Dalal 等人[163]提出了方向梯度直方图 HOG 来对图像中的行人进行检测，HOG 通过计算图像中部分区域的梯度信息（梯度的方向和大小等）来获取特征。其首先将图像划分成方格单元，对方格单元进行梯度信息的计算，因此对于图像的亮度差异和形变都有较好的鲁棒性。Wang 等人[165]将 LBP 特征和 HOG 特征相结合，从而能够充分利用图像中的纹理和梯度信息，有效提高了行人检测的性能。

2.1.1.2 基于深度学习的图像目标检测方法

基于深度学习的图像目标检测算法主要可以分为两大类：单阶段目标检测方法和两阶段目标检测方法。本小节将对这两类方法中的经典模型进行介绍。

（1）单阶段目标检测。

单阶段目标检测网络直接利用图像特征预测目标的类别和位置，其不需要利用候选框提取特征，检测速度更快。YOLO[37]将目标检测当成回归问题来处理，其网络结构如图 2.1（a）所示。对于输入的图像，YOLO 首先将其分割成 $S \times S$ 个网格区域，其中每个单元预测 B 个边界框以及相应的置信度得分。如果网格单元中存在目标，则置信度等于预测框与真值的交并比（Intersection Over Union，IOU），同时也会对其预测类别概率。YOLO 能够直接输出目标的类别和位置，在检测速度上能够达到实时的标准，但是其定位准确度较低，尤其是对小尺寸的目标。YOLO v2[166]用卷积层将 YOLO 中的全连接层替换，并在每个卷积层后加入了批归一化层（Batch Normalization，BN）来提高模型的收敛速度。此外，YOLO v2 使用了锚框来预测边界框。YOLO v3[167]采用了更深的网络结构和多尺度特征融合来进一步提升目标检测的性能。

Liu 等人[36]针对 YOLO[37]的缺陷提出了新的单阶段目标检测模型 SSD，其结构如图 2.1（b）所示。SSD 采用 VGG-16[168]作为主干网络，去掉全连接层并添加了一些额外的卷积层，利用这些层生成的不同尺度的特征图来进行目标检测。低层特征图富含细节特征，有助于检测小目标，而包含语义信息的高层特征图被用于检测大尺寸目标。此外，SSD 采用了锚框，利用多个尺度的特征图

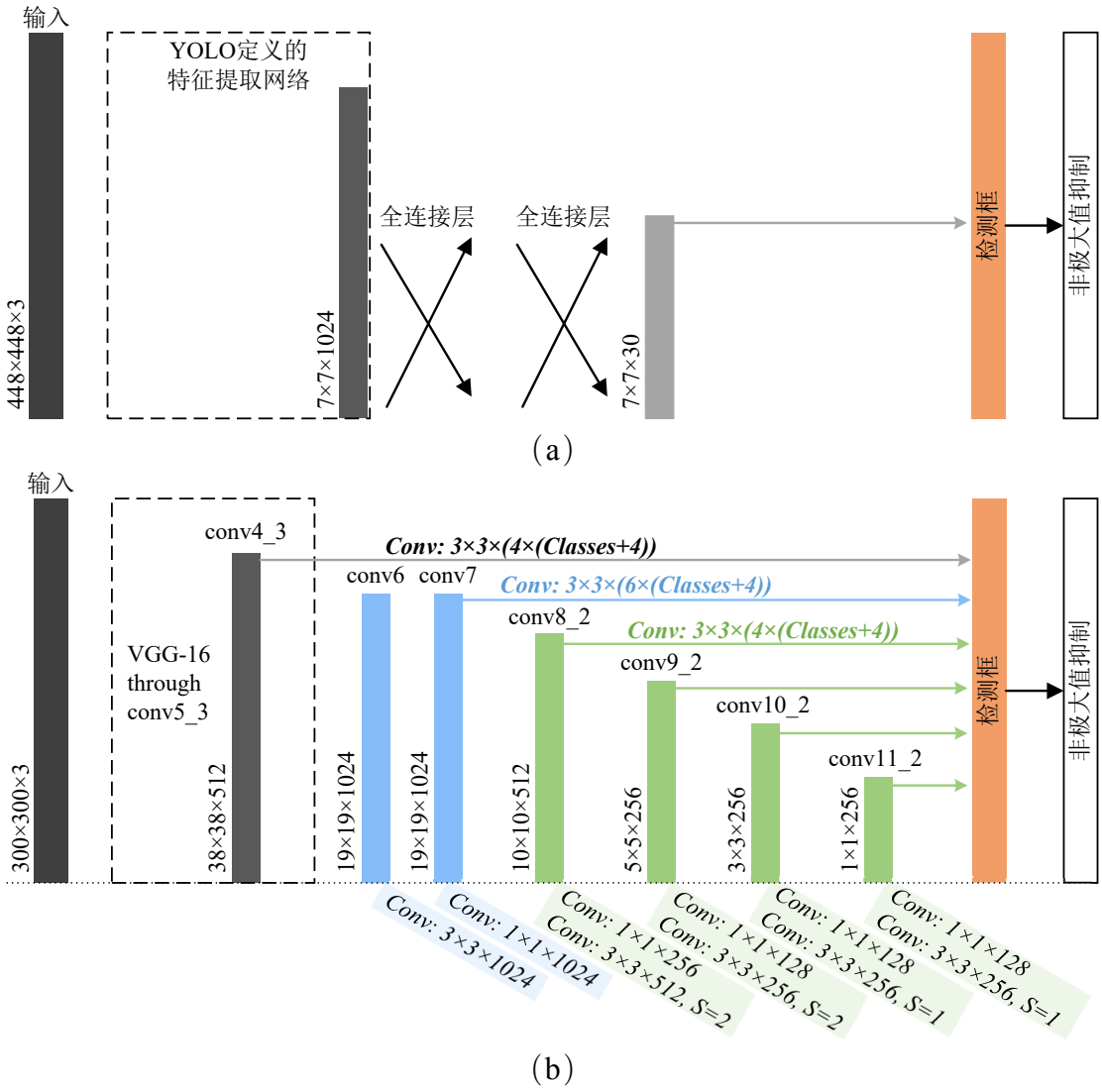


图 2.1 两种单阶段目标检测网络结构示意图。（a）YOLO[37]网络结构示意图。（b）SSD[36]网络结构示意图。

生成了一系列不同大小和长宽比的先验框，这有助于 SSD 克服 YOLO 定位准确度较低且检测小目标精度低的问题。单阶段目标检测方法采取的密集采样方式会产生大量的负样本，从而导致正负样本的不平衡。为了缓解这个问题，Lin 等人[52] 提出了 RetinaNet，其设计了一种应对类别不平衡问题的损失函数 Focal Loss，该损失函数能够在训练过程中有效地抑制负样本。此外，RetinaNet 采用了特征金字塔网络（Feature Pyramid Network，FPN）来进一步提高对不同尺寸目标的检测性能。Zhang 等人[169]结合两阶段目标检测方法的优势，提出了一种单阶段目标检测模型 RefineDet，其通过去掉负锚框同时微调锚框中心点坐标和

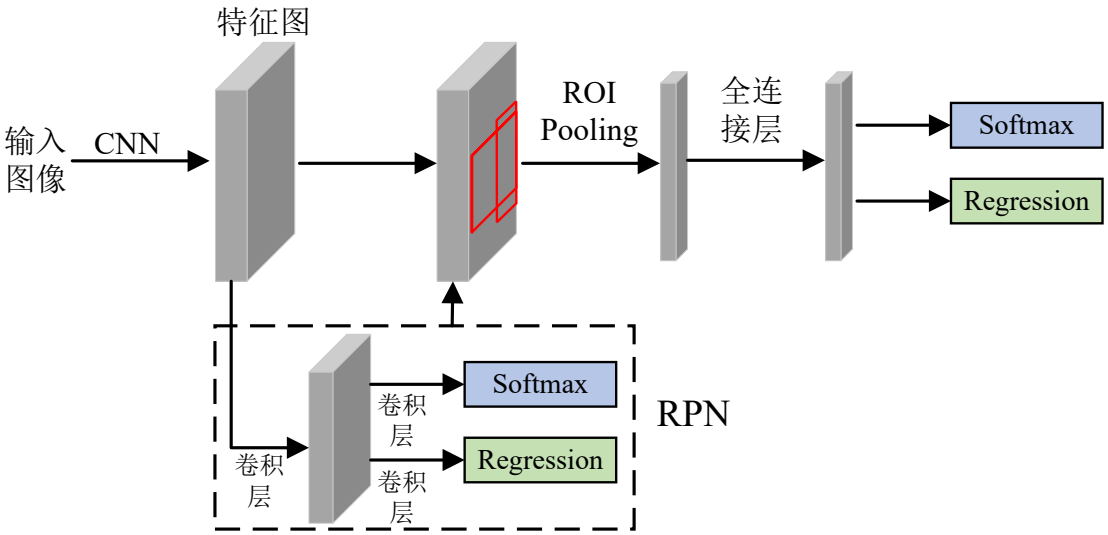


图 2.2 Faster R-CNN[35]网络结构示意图。

大小的操作，有效避免了类别不均衡的问题。

（2）两阶段目标检测。

两阶段目标检测网络首先在图像上生成候选区域，对候选区域进行特征学习后再进行位置回归和分类。R-CNN[170]是一个开创性的两阶段目标检测模型，其包含候选区域生成、对候选区域的特征提取以及区域分类三部分。R-CNN 首先利用选择性搜索算法[171]生成一系列候选区域，然后将这些候选区域裁剪并形变为固定大小，以便于输入 CNN 中提取特征。最后将提取的候选区域特征输入到支持向量机 SVM 中进行分类，并利用线性回归对候选框的位置和尺寸进行微调。R-CNN 的目标检测性能超过了所有的传统方法，然而，由于需要对所有的候选区域都进行特征提取，造成了大量的时间消耗。此外，R-CNN 的三个部分是单独训练的，无法实现端到端的训练方式。为了减少 R-CNN 的时间消耗，He 等人[172]设计了空间金字塔池化网络（Spatial Pyramid Pooling，SPPNet）。SPPNet 直接对整幅图像进行特征提取，并利用空间金字塔池化获取候选区域固定大小的特征向量。相比于 R-CNN[170]，SPPNet 无需调整输入图像的尺寸，避免了信息的损失，有效提高了目标检测的精度。然而 SPPNet 仍然需要多个阶段的训练过程，针对这个问题，Girshick 等人[173]提出了 Fast R-CNN。其在对整幅图像进行特征提取后，利用兴趣区池化层（ROI Pooling）来获取候选区域固定长度的特征向量，并通过回归层和分类层分别预测边界框的位置和类别。Fast R-CNN 可以实现端到端的训练，提高了目标检测的速度。

R-CNN[170]和 Fast R-CNN[173] 均采取选择性搜索算法来生成候选框，需要花费较多的时间。为了解决这个问题，ren 等人[35]提出了 Faster R-CNN，其网络结构如图 2.2所示。该方法设计了一种区域候选网络（Region Proposal Network，RPN）来生成候选框，利用滑动的卷积核在每个位置生成一系列的锚框，其中锚框的像素尺度和长宽比是预先定义的。RPN 的引入，使得 Faster R-CNN 在检测速度上获得了较大的提高，但是由于需要利用全连接层对候选区域的特征向量进行处理，带来了大量的计算消耗。Dai 等人[174]提出了基于区域的全卷积网络，减少了对候选框的特征学习时间。此外，针对 Faster R-CNN[35]只利用网络最后一层特征带来的难以检测小尺寸目标的问题，Lin 等人[140] 提出了特征金字塔网络。通过将深层和浅层特征相融合，来把深层丰富的语义信息融合到细节充分的浅层特征中，提高了对不同大小目标的检测性能。

2.1.2 图像语义分割

图像语义分割是对图像中每个像素的类别进行判断，其是视觉领域中一个重要的研究内容，相关技术在多个任务中都有较高的应用价值，例如遥感影像分析[6]、自动驾驶[175]、农作物监测[176]和医学影像诊断[177]等。

2.1.2.1 传统的图像语义分割方法

传统的图像语义分割方法会利用人工定义的图像特征和分类器进行分割，这类方法通常效率较低，且很大程度上依赖于所定义的特征提取规则。下面将对几种传统的图像分割技术进行简要介绍。

（1）基于边缘检测的图像分割方法。这类方法是利用图像中不同类别区域的边缘特征突变来进行图像分割的。首先利用图像的一阶或二阶梯度信息来获取区域的边缘，其中利用图像一阶梯度的有 Roberts 和 Sobel 算子等[11]，利用图像二阶梯度的有 Log 和 Laplacian 算子等[178]。在获取边缘的像素点后，通常会利用霍夫（Hough）变换[179] 将这些像素点连接从而生成最终的分割结果。然而，基于边缘检测的图像分割方法容易受到噪声干扰的影响，同时也不能很好地获取图像区域结构，限制了其鲁棒性。

（2）基于阈值的图像分割方法。基于阈值的分割方法通常较为高效，其首先确定一个或者多个像素的阈值，根据设定的阈值判断图像中每个像素的类别，从而获取分割结果。目前比较有代表性的基于阈值的分割方法有：最大类间方差[180] （即先设定一个或多个阈值，从而使得每类分割结果的类内方差最小）、

最大熵法[181]、均衡直方图[182] 和自适应阈值法[178]等。基于阈值的分割方法只关注像素的值，而忽略了像素之间的联系，因此这种方法易受噪声和光照条件等的影响。

（3）基于区域的图像分割方法。基于区域的分割方法是利用设定的规则将图像分割成不同的小区域，具体而言有两种规则，分别是区域分裂合并法[183] 和区域生长法[184]。区域分裂合并法首先将整个图像划分成多个小的部分，然后通过计算各相邻部分的相似性，将相似度高的相邻部分合并，直到所有相邻部分的相似度低于特定阈值，即停止合并。区域生长法首先初始化一个种子点，将与种子点相似度高的像素融入到该区域中，迭代直到所有像素都被融入到某个区域。基于区域的分割方法具有对噪声较高的鲁棒性，但是过程比较复杂且效率较低。

（4）基于图论的图像分割方法。基于图论的分割方法会将图像中的像素看做图中的点，像素之间的连线和距离分别当做图中边和边的权重，从而形成一个具有权重的无向图，其中边的权重定义了相邻像素之间的相似度。对此无向图根据一定的规则进行剪切，剪切得到的子图对应图像中的特定区域，即为图像分割的结果。基于图论的图像分割方法有：马尔可夫随机场（Markov Random Field）[185]、最小生成树（Minimum Spanning Tree）[186] 和随机游走等[187]。这种方法不能判别不同类别区域的相似数据，对模糊边界较为敏感。

2.1.2.2 基于深度学习的图像语义分割方法

相比于手工设计的特征，卷积神经网络（CNN）能够提取更具有辨别性的高阶特征，从而获得更好的性能，近年来在图像语义分割任务中基于 CNN 的模型获得了较大的发展。FCN（Fully Convolutional Network）[73]是图像语义分割任务中一个具有里程碑意义的研究，其将用于图像分类的 CNN 中的全连接层全部替换为卷积层，输出二维的特征图，再利用反卷积层将特征图上采样，从而得到与输入图像大小相同的语义分割结果。在 FCN 的基础上，U-Net[46]提出了一种对称的 U 型编码器-解码器（Encoder-Decoder）网络架构，来进行医学图像的语义分割。该方法采用跃层连接将编码器的特征融合到对应的解码器特征中，从而能够有效利用不同网络阶段的细节和抽象特征。由于性能出色，U-Net 已经被应用于不同场景图像的语义分割任务上[79,188]。SegNet[74]也采用编码器-解码器结构进行语义分割，并在解码器中采用相应编码器中最大池化层所保存的池化索引来对特征图进行上采样，这样的操作能够避免学习上采样的需要。

DeepLab 是用于图像语义分割的一系列研究工作，DeepLab v1[189]将空洞卷

积（Atrous Convolution）引入到 CNN 中，其能够在不改变网络参数量的情况下增大模型的感受野。此外，DeepLab v1 采用了条件随机场（Conditional Random Field，CRF）作为后处理步骤来优化边缘的细节。DeepLab v2[190]在 DeepLab v1 的基础上进一步提出了空洞空间金字塔池化模块（Atrous Spatial Pyramid Pooling，ASPP）来获取多尺度图像特征。ASPP 由四个具有不同扩张率（扩张率分别为 6、12、18 和 24）的并行空洞卷积组成，分别对输入特征图进行空洞卷积运算来获取多尺度信息，之后将四个空洞卷积的输出特征图进行融合。DeepLab v3[191]对 DeepLab v2 中的 ASPP 进行了改进：（a）将 ASPP 中扩张率为 24 的空洞卷积替换为 1×1 的卷积，从而避免当扩张率过大时导致部分卷积核权重失效的问题；（b）在四个卷积分支的基础上，添加了一个全局平均池化的分支，并将这种包含图像级别特征的输出与其他卷积分支的输出特征相融合；（c）在 ASPP 的卷积层后添加了批归一化层（BN）来使网络在训练时能够快速地收敛。在 DeepLab v3 的基础上，DeepLab v3+[75]引入了编码器-解码器结构，将编码器中的低层特征与 ASPP 模块输出的特征通过解码器进行融合，从而有效提高了图像语义分割的性能。此外，PSPNet[192]利用金字塔池化模块（Pyramid Pooling）来获取图像的全局上下文信息。Refinenet[193]采用多个分辨率的图像作为输入，同时提出了链式残差池化模块（Chained Residual Pooling），其能够从较大的图像区域获取背景语义信息。近年来，一些研究将 Transformer 结构引入到语义分割任务中。SETR[161]利用 Transformer 框架将图像编码为图像块的序列，结合一个解码器完成图像的分割。SegFormer[194]利用一个分层结构的 Transformer 编码器输出多尺度的特征，同时采用多层感知机作为解码器来融合多级别的特征并输出语义分割的结果。

第二节 注意力机制

注意力机制使得神经网络能够自动关注图像特征中的主要信息，同时抑制次要信息。从注意力机制的工作机理出发，可以将其分为选择性注意力和自注意力机制两大类。本小节主要介绍了选择性注意力中的经典模块——挤压-激励（Squeeze-and-Excitation，SE）模块和自注意力中的经典模块——非局部（Non-Local）模块，此外也对利用自注意力机制的 Transformer 结构进行了介绍。

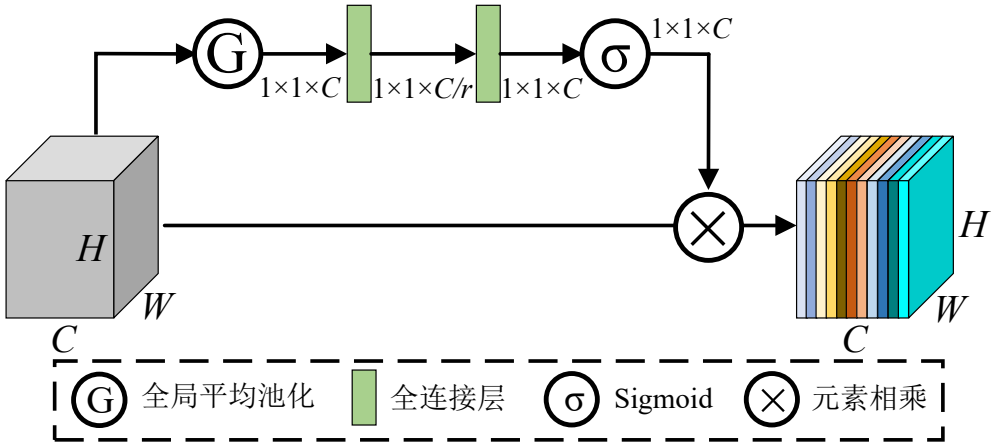


图 2.3 挤压-激励（Squeeze-and-Excitation，SE）模块[44]的结构示意图。

2.2.1 挤压-激励模块

挤压-激励（Squeeze-and-Excitation，SE）模块[44] 是一种通道注意力机制，能够根据特征图通道的重要程度对其进行重校准。SE 模块的结构如图 2.3所示，主要包含挤压（即全局信息嵌入）和激励（即自适应重校准）两个步骤。

（1）挤压。令 SE 模块的输入特征图尺寸为 $H \times W \times C$，其中 H、W 和 C 分别表示特征图的高度、宽度和通道数量。首先利用全局平均池化来生成特征图通道的向量，其中嵌入了将空间维度（$H \times W$）信息压缩成的全局空间信息，压缩后的向量尺寸为 $1 \times 1 \times C$。

（2）激励。将挤压步骤得到的 $1 \times 1 \times C$ 向量输入到两层非线性的全连接层中，其中第一个全连接层用来降维，将通道数缩放为 C/r（r 为通道数的缩放因子）来减少计算量。第二个全连接层用来升维，将通道数量恢复为 C。经过两层全连接层后，输出向量的尺寸为 $1 \times 1 \times C$。然后将该向量输入到一个 Sigmoid 激活函数中，将其映射到 [0，1] 的范围，从而得到通道注意力权重。将通道注意力权重向量（尺寸为 $1 \times 1 \times C$）与 SE 模块的输入特征图（尺寸为 $H \times W \times C$）进行元素相乘的操作，即可得到 SE 模块对通道进行重校准后的输出特征图（尺寸为 $H \times W \times C$）。

2.2.2 非局部模块

自注意力机制是能够增强输入特征图中不同部分之间相关性的注意力方法，其最早是在自然语言处理领域被提出的[146]，由于性能卓越，迅速被引入到计算机视觉领域。Wang 等人[18]提出了基于自注意力机制的非局部（Non-Local）模

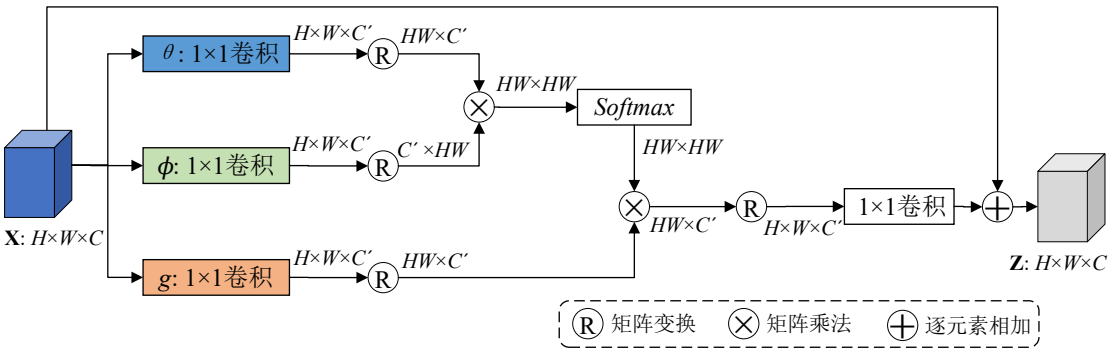


图 2.4 非局部（Non-Local）模块[18]的结构示意图。

块，其结构如图 2.4所示。

设置 $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ 表示非局部模块的输入特征图，H、W 和 C 分别表示特征图的高度、宽度和通道数量。为了获取整个特征图的长程依赖信息，Non-Local 模块利用所有像素特征的加权和来计算响应 $\mathbf{Y} \in \mathbb{R}^{H \times W \times C'}$。其定义如下：

$$\mathbf{Y} = f(\theta(\mathbf{X}), \phi(\mathbf{X}))\, g(\mathbf{X}), \tag{2.1}$$

其中，$\theta(\cdot)$、$\phi(\cdot)$、$g(\cdot) \in \mathbb{R}^{H \times W \times C'}$ 是对输入特征图 $\mathbf{X}$ 的可以进行学习的转换，Non-Local 模块中是采用 1×1 的卷积层生成的。其可以表示为：

$$\theta(\mathbf{X}) = \mathbf{W}_\theta \mathbf{X},\ \phi(\mathbf{X}) = \mathbf{W}_\phi \mathbf{X},\ g(\mathbf{X}) = \mathbf{W}_g \mathbf{X}, \tag{2.2}$$

其中，$\mathbf{W}_\theta$、$\mathbf{W}_\phi$ 和 $\mathbf{W}_g$ 为相应的权重矩阵，它们会将 $\mathbf{X}$ 的特征通道数降低以减少计算量。$\theta(\mathbf{X})$、$\phi(\mathbf{X})$ 在经过形状变换后输入到函数 $f(\cdot,\cdot)$ 中，从而获得特征图中任意位置像素间的相似度，得到的相似度之后被输入到 $Softmax$ 层中来获取空间注意力权重矩阵。Non-Local 模块原文[18]中描述了函数 $f(\cdot,\cdot)$ 的几种选择，包括：高斯（Gaussian）、嵌入式高斯（Embedded Gaussian）、点乘（Dot Product）和拼接（Concatenation）。值得注意的是，如果采用嵌入式高斯的方式，则 Non-Local 的操作与自注意力模块[146]相同。将利用函数 $f(\cdot,\cdot)$ 得到的注意力权重与经过形状变换的 $g(\mathbf{X})$ 相乘，即可得到响应 $\mathbf{Y} \in \mathbb{R}^{H \times W \times C'}$，其公式定义为：

$$\mathbf{Y} = f(\theta(\mathbf{X}), \phi(\mathbf{X}))\, g(\mathbf{X}) = softmax\left(\mathbf{W}_\theta \mathbf{X} \mathbf{X}^\top \mathbf{W}_\phi^\top\right) \mathbf{W}_g \mathbf{X}. \tag{2.3}$$

对 $\mathbf{Y}$ 进行形状变换后经过 1×1 的卷积层，使其通道数与输入特征 $\mathbf{X}$ 相同，再利用一个残差连接[137]与 $\mathbf{X}$ 进行逐元素相加，即可得到 Non-Local 模块的最终

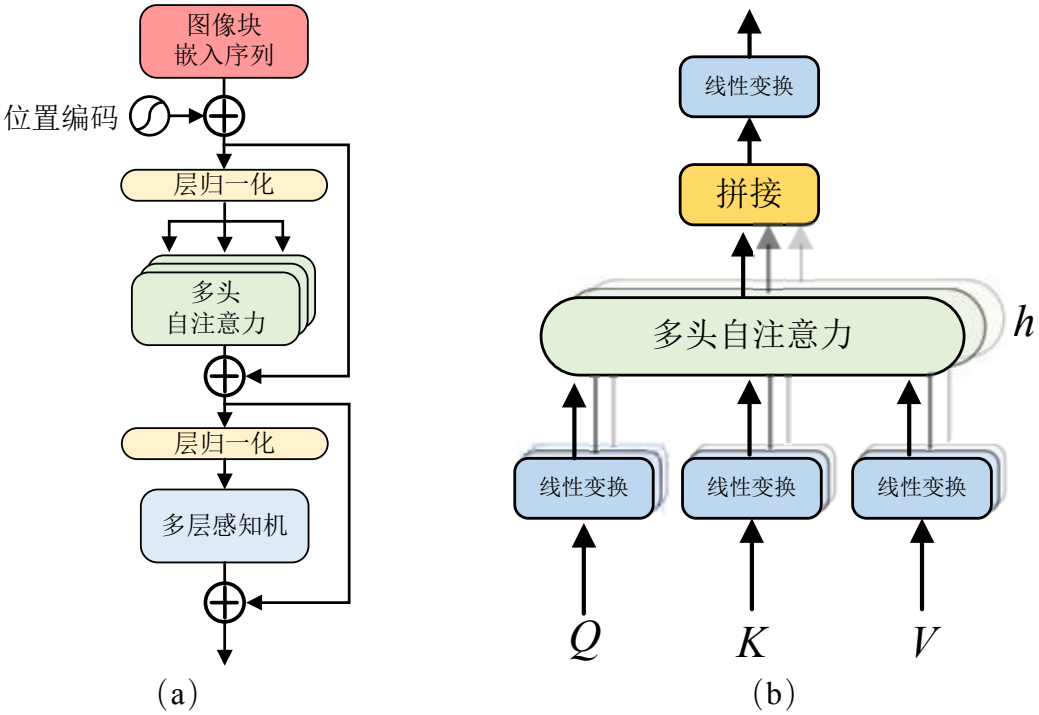


图 2.5 （a）Transformer 编码器[146]结构示意图。（b）多头自注意力（Multi-Head Self-Attention）[146]结构示意图。

输出，其定义为：

$$\mathbf{Z} = \mathbf{W}_Z \mathbf{Y} + \mathbf{X}, \tag{2.4}$$

其中，$\mathbf{W}_Z$ 为权重矩阵，$\mathbf{Z} \in \mathbb{R}^{H \times W \times C}$ 是 Non-Local 模块的输出。‘+’ 表示残差连接[137]，使用残差连接结构使得 Non-Local 模块可以插入到任何预训练模型中。

2.2.3 视觉中的 Transformer

Transformer[146]是针对机器翻译任务提出的，其在自然语言处理（Natural Language Processing，NLP）任务中取得了出色的的性能[153]。Transformer 中的关键结构是自注意力机制，由于其能够学习到输入项之间长程依赖关系的能力，近年来在计算机视觉领域的多个任务中获得了较多的关注[154,159,161]。本小节对计算机视觉中的 Transformer 编码器进行了介绍，其结构如 图 2.5（a）所示，Transformer 编码器包含层归一化（Layer Normalization，LN）、多头自注意力（Multi-Head Self-Attention，MSA）和多层感知机（Multi-Layer Perceptron，MLP）等运算操作。

标准 Transformer 的输入是一维的嵌入序列。为了能够对二维图片进行处

理，需要首先将图像 $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$ 划分成二维图像块，并展平成一维的序列 $\mathbf{x}_p \in \mathbb{R}^{N\times(P^2\cdot C)}$，其中 H、W 和 C 分别表示图像的高度、宽度和通道数量，(P,P) 表示每个图像块的分辨率大小，$N = HW/P^2$ 为图像块的数量。将 $\mathbf{x}_p$ 输入到一个可以学习的线性层从而获得特征嵌入序列 $e_f \in \mathbb{R}^{N\times D}$，其中 D 是隐藏向量的大小。由于 Transformer 中自注意机制的本质是无序的，为了保留图像的空间信息，需要学习一维的位置嵌入序列 e_{pos}，并将其与特征嵌入序列 e_f 相加从而获得 Transformer 的输入图像块序列 $E \in \mathbb{R}^{N\times D} = e_f + e_{pos}$。

在单个自注意力模块中，输入的图像块序列 E 被线性变换为三个部分，即查询（Query）$Q \in \mathbb{R}^{N\times d_k}$，键（Key）$K \in \mathbb{R}^{N\times d_k}$，以及值（Value）$V \in \mathbb{R}^{N\times d_v}$，其中 d_k、d_v 分别是查询 Query（键 Key）和值 Value 的维数。然后将尺度点乘注意力（Scaled Dot-Product Attention）应用于 Q、K 和 V，其定义为：

$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V, \tag{2.5}$$

其中，‘softmax’ 是指 softmax 函数，$\sqrt{d_k}$ 是一个缩放因子。将公式 (2.5) 与 Non-Local 模块中的 公式 (2.3) 相比，Q、K 和 V 分别对应 Non-Local 模块中的 $\theta(\mathbf{X})$、$\phi(\mathbf{X})$ 和 $g(\mathbf{X})$，两个公式相似，只是 Transformer 中的尺度点乘注意力多了一个缩放因子 $\sqrt{d_k}$。

多头自注意力是包含多个独立自注意力模块的拓展，其结构如图 2.5（b）所示。多头自注意力将 Q、K 和 V 进行多次拆分，并行计算公式 (2.5) 中的尺度点乘注意力函数，之后将所有头的输出进行拼接并线性转换得到最终的输出。多头自注意力模块的输出在经过层归一化后，被输入到多层感知机中进行特征转换，从而得到一层 Transformer 编码器的输出。

第三节 本章小结

本章针对本文研究内容所涉及的相关理论技术进行了介绍。首先针对与本文研究内容相关的计算机视觉两个任务，即图像目标检测和图像语义分割，介绍了这两个任务的传统方法和基于深度学习模型的发展现状，并对其中的经典模型进行了回顾和比较。此外，详细地介绍了注意力机制的相关理论技术，包括选择性注意力机制中的经典模块——挤压-激励（squeeze-and-excitation，SE）模块和自注意力机制中的经典模块——非局部（Non-Local）模块，也对利用自注意力机制的 Transformer 结构进行了介绍。

第三章　基于切片关联注意力的 CT 图像肺结节检测

CT 图像在通道维度包含多张连续的断层扫描切片，不同切片间关联性很大，共同组成了人体的组织，属于立体空间序列图像。本章基于肺部 CT 图像，针对其立体空间维度较高的难点，提出了一种切片关联注意力网络进行肺结节的检测。本章的章节安排为：第一节介绍了本章的研究背景、研究内容和创新点；第二节介绍了所提出的基于切片关联注意力的肺结节检测网络；第三节描述了新的大规模肺结节检测数据集 PN9；第四节是对本章所提出的肺结节检测模型的对比实验及相关分析；第五节是对本章内容进行的总结。

第一节　引言

肺癌已经成为世界上致死率最高的癌症之一[195-196]。肺结节是肺部的一种病变，有很高的风险发展成为恶性肿瘤，而针对肺结节的早期诊断和及时治疗是防治肺癌的有效手段。胸部计算机断层扫描（即 CT）是一种早期诊断肺结节的方式[197]，近年来在降低肺癌的死亡率方面发挥了重要的作用[198]。在肺部 CT 图像当中，肺结节与其他组织的 X 光（X-Ray）吸收水平通常是相近的。然而，与血管、支气管的连续管状结构不同的是，肺结节通常是孤立的球状结构，这种结构区别为肺结节的识别提供了基础。在对 CT 图像进行分析时，医生需要同时查看数百张图片，即使是经验丰富的医生也通常需要十分钟左右的时间来对一个病人进行一次全面的分析。此外，肺部会存在很多的小尺寸结节，不同类型的肺结节会有不同的形态表现，因此准确的识别和诊断肺结节对于医生来说是一个较大的挑战[199]。

目前，计算机辅助诊断（Computer-Aided Diagnosis，CAD）系统已经得到较好的发展，其能够协助医生更加有效和准确地解析 CT 图像[47,200]。传统的计算机辅助诊断系统主要依靠形态学操作或者低阶的特征映射来检测肺结节[25-26,28]，然而由于肺结节的大小、形状和类型等变化多样，这些方法的检测结果往往较差。得益于深度学习技术的发展，基于卷积神经网络（Convolutional Neural Network，CNN）的模型，例如 Faster R-CNN[35]、SSD[36]、YOLO[37]等方法被陆续提出并在目标检测任务中取得了较好的性能。与此同时，CNN 也被应用于医

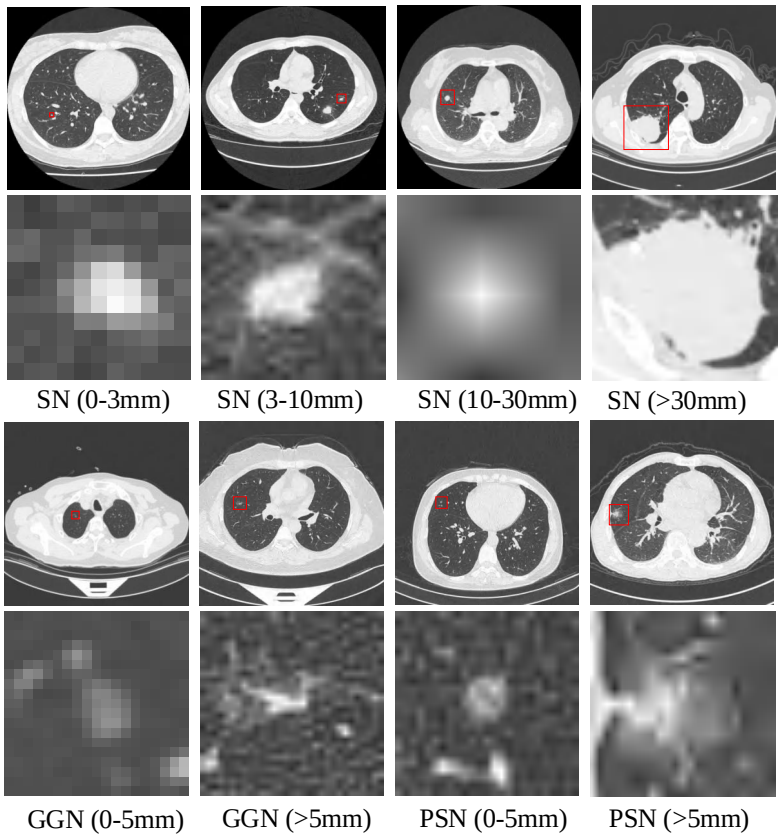


图 3.1 本章所提出的 PN9 数据集中的肺结节图像示例。其中每张图像属于不同的肺结节类别，SN、PSN、GGN 分别为实性结节（Solid Nodule）、部分实性结节（Part-Solid Nodule）和磨玻璃结节（Ground-Glass Nodule），括号中标注了每个结节的大小。第一行和第三行是完整的切片，其余两行分别是放大后的图像。

疗影像处理任务中[201-204]。对于肺结节的检测来说，基于 CNN 的方法[38,45,205]要比传统方法的性能更好。相比于自然图像中的二维目标检测，肺结节检测是利用三维的 CT 图像数据进行三维目标检测的任务，因此具有更大的难度。一些研究[41,56]首先利用二维区域候选网络（Region Proposal Network，RPN）获取三维 CT 图像中每一个二维切片的候选框，再将二维候选框进行跨切片的融合以生成三维的候选框。近年来，越来越多的方法[38,45,206]将三维卷积神经网络应用到 CT 图像数据中来直接生成三维的候选结节框。相比于二维卷积神经网络，三维卷积神经网络的参数量更大，导致其需要更长的时间和更多的 GPU 显存来进行训练，但是其也在针对 CT 图像的分析中获得了比二维卷积神经网络更好的结果[56]。

利用目前已公开的几个 CT 图像数据集，例如 LIDC-IDRI[207]和 LUNA16[208]，

基于 CNN 的方法已成为肺结节检测的主流。这些数据集支持研究人员在相同的评价指标下设计并评估不同的肺结节检测算法，进一步推动了计算机辅助诊断系统在临床医学中的应用。然而，即使是 LUNA16[208]这样目前应用最广泛的肺结节检测数据集，也只有 888 张 CT 图像，同时标注的肺结节数量和种类都有限。这些数据对于三维卷积神经网络的训练来说是不充足的，限制了神经网络在肺癌诊断中的应用，因此一个数据量更大且种类更多的肺结节检测数据集是有必要的。

本章提出了一个新的大规模肺结节检测数据集 PN9（Pulmonary Nodule Dataset with 9 Classes），其中包含 8798 张来自不同病人的 CT 图像和 40439 个标注的肺结节，部分示例如图 3.1所示。参考现有的医学指南并考虑到医院和医生在临床诊断时的需求，将所有的结节根据其类型和大小划分为 9 个不同的类别。与现有的肺结节检测数据集相比，PN9 包含了数据量更大的 CT 图像和种类更多的标注肺结节，使得研究人员能够基于肺结节的丰富属性设计更加有效的算法。与此同时，PN9 所标注的更多小尺寸肺结节有助于提高小结节诊断的准确性，从而能够在更早的阶段对患者进行治疗。

此外，本章提出了一种切片关联注意力网络（Slice-Aware Network，SANet），基于 CT 图像进行肺结节的检测。由于肺结节的尺寸要比一般的自然目标小很多，例如 图 3.1中的小尺寸结节，所以本章首先引入一个编码器-解码器网络进行肺结节特征的学习。借鉴医生的临床诊断方法，本章提出了一种切片分组非局部模块（Slice Grouped Non-Local，SGNL）并将其添加到编码器网络中，其能够捕获一组切片特征图中任意位置和任意通道之间的长程依赖信息。三维区域候选网络对肺结节的检测具有较高的灵敏度，但通常会带来大量的假阳性样例，所以本章设计了基于多尺度特征图的假阳性抑制模块（False Positive Reduction，FPR）来减少假阳并进一步提高肺结节检测的性能。在 PN9 数据集和公开数据集上，本章将所提出的方法与几种基于二维卷积神经网络和三维卷积神经网络的检测模型进行了比较，实验充分验证了 SANet 的性能。

本章研究工作的主要贡献包括：

- 提出了切片关联注意力网络进行肺结节的检测，其在多个数据集上都取得了较好的性能。
- 受到医生诊断方式的启发，设计了切片分组非局部模块，将切片维度分组的思想引入自注意力机制中，来充分学习 CT 图像中的立体空间序列信息。

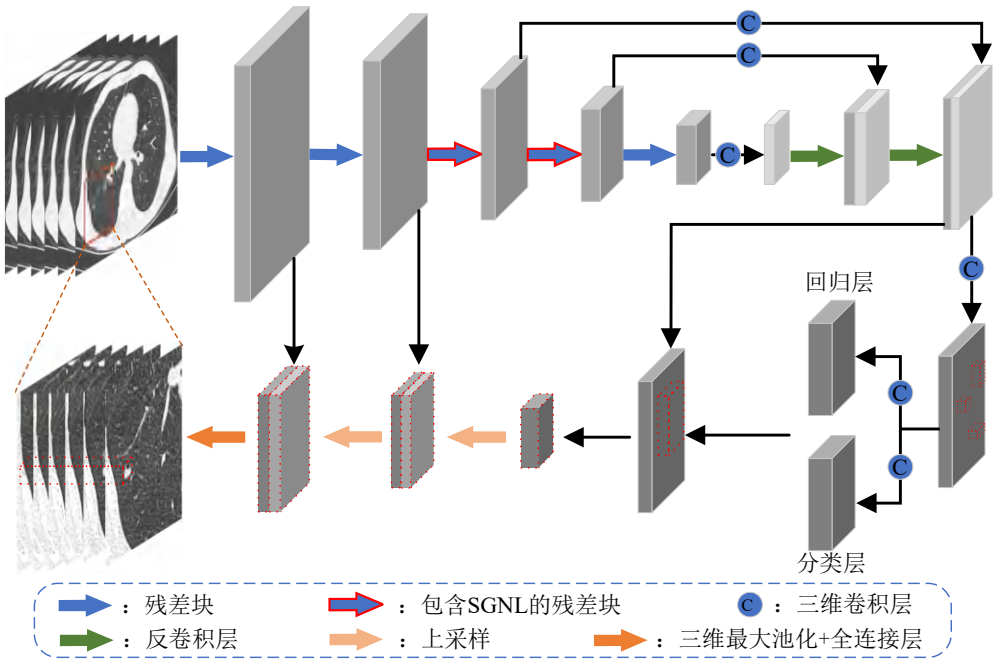


图 3.2 本章所提出的切片关联注意力网络（Slice-Aware Network，SANet）的总体架构。图中红色虚线框表示候选肺结节。

- 构建了医学影像领域目前为止规模最大、种类最丰富的肺结节检测数据集，其中包含 8798 张 CT 图像和 40439 个属于 9 种不同类别的标注肺结节。

第二节 基于切片关联注意力的肺结节检测

不同于一般自然图像中的二维目标检测，肺结节检测是利用三维 CT 图像进行三维目标检测的任务。为了充分利用 CT 图像中不同切片之间的三维立体空间序列信息，本章提出了切片关联注意力网络（SANet），如图 3.2所示。在切片关联注意力网络中，本章设计了一种切片分组非局部（SGNL）模块，用于获取一组切片特征图中任意位置和任意通道之间的长程依赖关系，同时提出了一种假阳性抑制（FPR）模块，以进一步提高肺结节检测的性能。

3.2.1 网络结构

在本章所提出的模型 SANet 中，考虑到三维 ResNet-50[137]在特征提取方面的出色性能，采用其作为编码器。然而，与自然图像中的普通物体相比，CT 图像中的肺结节尺寸较小且变化较大。三维 ResNet-50 采用 5 个残差阶段来对 CT 图像进行编码，不能准确地描述不同尺寸肺结节的特征，从而导致较差的肺结

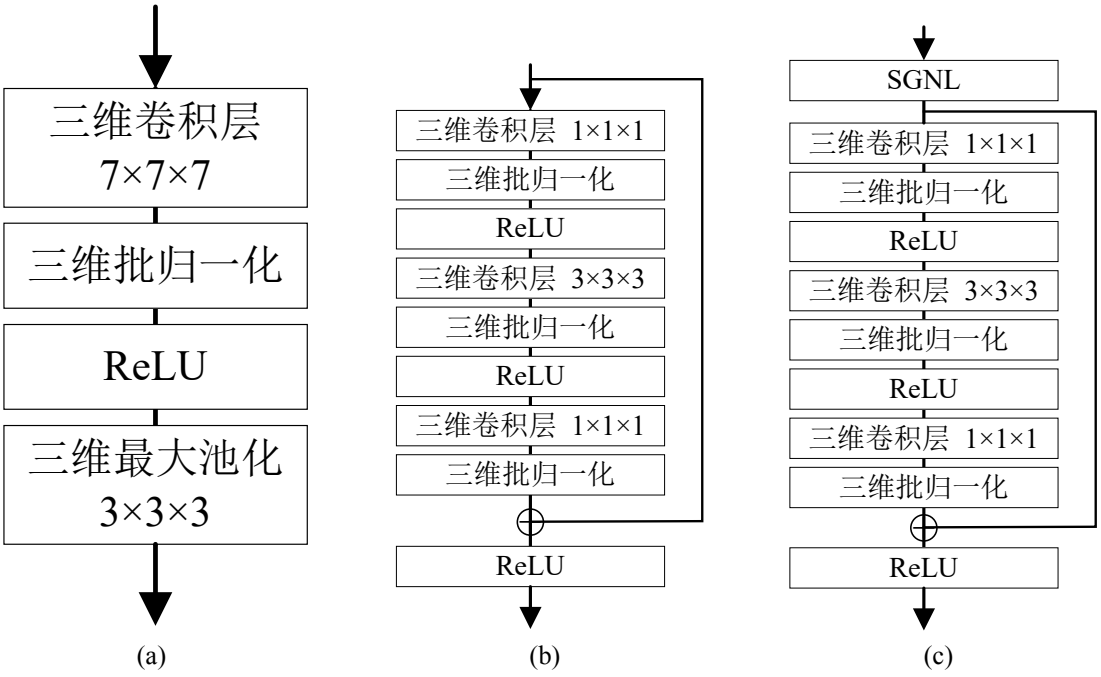


图 3.3 三维 ResNet-50 残差块的结构示意图。（a）残差阶段 *res*1 的结构；（b）残差阶段 *res*2-*res*5 中残差块的结构；（c）残差阶段 *res*3 和 *res*4 中包含 SGNL 残差块的结构。

节检测性能。为了解决这一问题同时使网络能够获取多尺度信息，本章采用了 U 型编码器-解码器架构[46]。其中解码器包含两个 $2 \times 2 \times 2$ 的反卷积层，将特征图上采样到合适的大小。反卷积层的每个输出特征图与编码器网络中对应的输出相连接，其中后者的输出会经过一个 $1 \times 1 \times 1$ 的三维卷积层对通道数量进行调整。为了方便后续的描述，将编码器-解码器网络产生的特征图分别定义为 $\{M_{res1}, M_{res2}, M_{res3}, M_{res4}, M_{res5}, M_{de1}, M_{de2}\}$，其中 *res*1 到 *res*5 为编码器中的 5 个残差阶段，*de*1 和 *de*2 为解码器中的两个反卷积层。图 3.3（a）和（b）分别展示了三维 ResNet-50 *res*1 和其他阶段残差块的结构。

为了生成候选肺结节，对特征图 M_{de2} 应用 $3 \times 3 \times 3$ 的三维卷积层，其后面跟随两个并行的 $1 \times 1 \times 1$ 卷积层来回归出三维候选框（即 图 3.2中的回归层）并且生成分类结果（即图 3.2中的分类层）。根据肺结节尺寸大小的分布，本章设计了尺寸为 5、10、20、30、50 的五种锚框（Anchor）。其中每个锚框被指定了 6 个回归参数，分别是：中心点的 z、y、x 坐标、深度、高度和宽度。本章多任务损失函数的定义为：

$$L_{RPN} = L(\{p_i\}, \{t_i\}) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*), \quad (3.1)$$

其中，i 是三维图像块中第 i 个锚框的索引，N_{cls} 和 N_{reg} 分别为计算分类损失和回归候选框损失时锚框的数量，λ 是一个用来平衡这两个损失的参数。p_i 是第 i 个锚框为肺结节的预测概率，p_i^* 为相应的真值，如果锚框是肺结节则 p_i^* 为 1，否则 p_i^* 为 0。在本章方法中，如果一个锚框与真值中任何肺结节的三维交并比（3D Intersection over Union，3D IoU）大于或等于 0.5，这个锚框被认为是真阳性（即 $p_i^* = 1$）。如果没有满足上述条件的锚框，则三维交并比最高的锚框会被分配真阳性标签。另一方面，如果一个锚框与真值中任何肺结节的三维交并比都小于 0.02，则其被认为是非肺结节。

t_i 是一个向量，用来表示预测肺结节的 6 个参数化位置坐标，t_i^* 为相应的真值向量。t_i 和 t_i^* 的定义如下所示（为了符号表示的方便，在后文中下标 i 被省略。）：

$$\begin{aligned} t &= \left(\frac{z - z_a}{d_a}, \frac{y - y_a}{h_a}, \frac{x - x_a}{w_a}, \log\frac{d}{d_a}, \log\frac{h}{h_a}, \log\frac{w}{w_a} \right), \\ t^* &= \left(\frac{z^* - z_a}{d_a}, \frac{y^* - y_a}{h_a}, \frac{x^* - x_a}{w_a}, \log\frac{d^*}{d_a}, \log\frac{h^*}{h_a}, \log\frac{w^*}{w_a} \right), \end{aligned} \tag{3.2}$$

其中，x、y、z、w、h 和 d 分别表示预测框的中心坐标、宽度、高度和深度。x^*、y^*、z^*、w^*、h^* 和 d^* 是真值框的参数，x_a、y_a、z_a、w_a、h_a 和 d_a 表示锚框的参数。另外，本章对分类层应用了加权二值交叉熵损失函数，对回归层应用了平滑 L_1 损失[173]。

3.2.2 切片分组非局部模块

在肺部 CT 图像中，血管、支气管等组织是连续的管状结构，而肺结节多为孤立的球状结构。为了从其他组织中识别出肺结节，医生需要查看多个连续的切片来获取它们之间的相关性，从而识别出与其他组织结构不同的肺结节。借鉴医生的临床诊断方式，本章提出了由非局部操作（Non-Local）[18]构成的切片分组非局部模块（如 图 3.4所示），其可以学习跨切片像素之间的显式相关性。

3.2.2.1 非局部操作

设置 $\mathbf{X} \in \mathbb{R}^{D \times H \times W \times C}$ 表示非局部模块的输入特征映射，其中 D、H、W、C 表示特征图的深度、高度、宽度和通道数量。原始非局部操作[18]的定义如下：

$$\mathbf{Y} = f(\theta(\mathbf{X}), \phi(\mathbf{X}))\, g(\mathbf{X}), \tag{3.3}$$

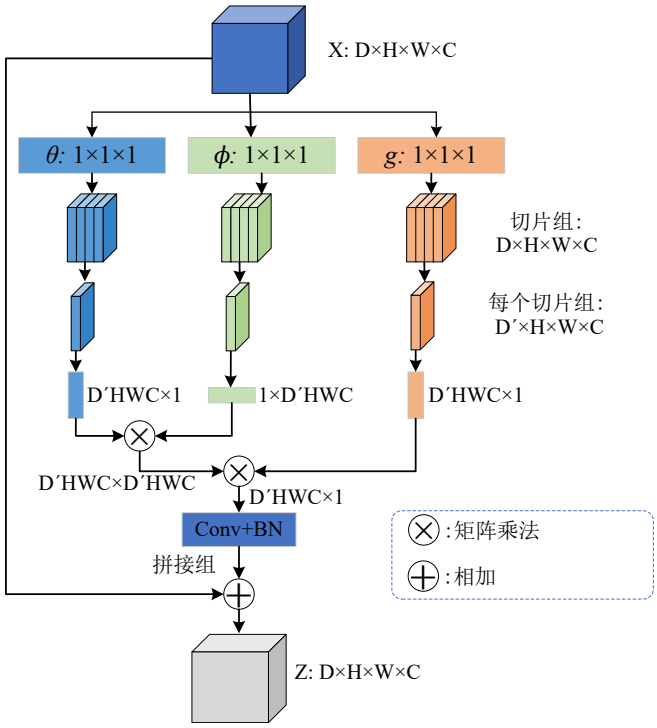


图 3.4 切片分组非局部模块（Slice Grouped Non-Local，SGNL）的结构示意图。在三个 $1\times1\times1$ 的三维卷积层之后，特征图沿着深度维度被划分为多个组。深度维度被分组为 $D'=D/G$，其中 G 是分组数。

其中，$\mathbf{Y}\in\mathbb{R}^{D\times H\times W\times C}$。$\theta(\cdot)$、$\phi(\cdot)$、$g(\cdot)\in\mathbb{R}^{DHW\times C}$ 由 $1\times1\times1$ 的卷积层生成，其可以表示为：

$$\theta(\mathbf{X})=\mathbf{X}\mathbf{W}_\theta,\ \phi(\mathbf{X})=\mathbf{X}\mathbf{W}_\phi,\ g(\mathbf{X})=\mathbf{X}\mathbf{W}_g, \tag{3.4}$$

其中 $\mathbf{W}_\theta,\mathbf{W}_\phi,\mathbf{W}_g$ 为需要学习的权重矩阵，函数 $f(\cdot,\cdot)$ 用于计算特征图中所有位置之间的相似度。非局部操作[18]中描述了函数 f 的几种选择，点乘是其中最简单的形式，其定义为：

$$f(\theta(\mathbf{X}),\phi(\mathbf{X}))=\theta(\mathbf{X})\phi(\mathbf{X})^{\top}. \tag{3.5}$$

3.2.2.2 切片分组非局部

原始的非局部操作可以捕获特征图中任意位置之间的长程依赖关系，但是通道之间的相似性对于区别细粒度物体来说也很重要，如文献[209-210]所述。本章考虑了非局部操作中的跨通道信息，来建模特征图中任意位置和任意通道之间的长程依赖关系。通过将通道维度合并到其他维度中来将公式 (3.4) 的输出变

形，得到 $\theta(\cdot)$、$\phi(\cdot)$、$g(\cdot) \in \mathbb{R}^{DHWC}$，切片分组非局部模块利用如下公式来计算 $\mathbf{Y}$:

$$\mathbf{Y} = f\left(vec\left(\theta(\mathbf{X})\right), vec\left(\phi(\mathbf{X})\right)\right) vec\left(g(\mathbf{X})\right), \tag{3.6}$$

其中，vec 表示变形后的向量。

直接利用公式 (3.6) 来对变形后的 $\theta(\cdot)$、$\phi(\cdot)$、$g(\cdot)$ 进行运算会带来远高于原始非局部操作的计算复杂度，因为其中存在 $DHWC \times DHWC$ 的成对矩阵，因此直接实现切片分组非局部模块是不可行的。近期的一些研究探索了分组卷积的思想，例如 Xception[211]、MobileNet[212]、ResNeXt[213]和 Group normalization[214]，这些方法验证了将特征图的通道分组能够有效提高卷积神经网络的性能。本章在切片分组非局部模块中引入了分组的思想，同时考虑到 CT 图像中肺结节的结构属性，将深度维度 D 划分为 G 个组。如图 3.4所示，特征图中每个组的深度维度为 $D' = D/G$。每个组通过公式 (3.6) 单独执行以计算 $\mathbf{Y}'$，并且将所有组的结果沿着深度维度拼接以获得 $\mathbf{Y}$。在 CT 图像中，一个结节通常存在于连续的几个切片中，因此利用所有深度维来检测每一个结节是不必要的。切片分组操作可以获取一个组特征图中任意位置、任意通道之间的相似性，这增强了对一个切片组中不同尺寸结节的判别能力。

图 3.4展示了切片分组非局部模块中每组的实现流程，公式 (3.6) 中的切片分组非局部操作被整合到切片分组非局部残差块中，其定义为:

$$\mathbf{Z} = concatenate\left(BN\left(\mathbf{Y}'\mathbf{W}_z\right)\right) + \mathbf{X}, \tag{3.7}$$

其中，$\mathbf{W}_z$ 代表 $1 \times 1 \times 1$ 的三维卷积层，BN 是批正则化[215]，“*concatenate*” 表示将所有组的结果沿着深度维度拼接。残差连接 “$+\mathbf{X}$” 使得切片分组非局部模块与现有的神经网络残差块兼容。对于切片分组非局部残差块的配置，本章根据文献[18]，将 5 个切片分组非局部残差块（2 个在 *res*3，3 个在 *res*4，分别间隔一个残差块）替换到 ResNet-50 中。其中，包含 SGNL 残差块的结构如图 3.3（c）所示。

3.2.3 假阳性抑制模块

三维区域候选网络的引入是为了筛选具有高灵敏度的候选肺结节，然而其中通常包含较多的假阳性样本。一些胸部组织，如结节状结构、纵隔结构、大血管和疤痕等，会比较容易被误诊为假阳性肺结节。本章进一步提出了一个假

阳性抑制模块（False Positive Reduction，FPR），以减少候选结节中的假阳性结节数量，并生成最终的结果。

如图 3.2所示，考虑到 ResNet 中的浅层残差块能够产生包含丰富空间细节的高分辨率特征图，本章利用多尺度特征图来减少假阳性。通过使用结节候选框裁剪特征图 M_{res1}、M_{res2}、M_{de2}，可以获得三种不同比例的兴趣区域（Regions of Interest，RoI）：R_{res1}、R_{res2}、R_{de2}。其中 R_{de2} 被上采样并与 R_{res2} 拼接，再次上采样后与 R_{res1} 拼接。最终的兴趣区域通过三维最大池化，再利用两个全连接层以获得分类概率和边界框回归偏差。利用假阳性抑制模块，可以减少假阳性，并进一步优化候选框的回归参数。

第三节　肺结节数据集

本章收集并标注了一个新的大规模肺结节检测数据集，命名为 PN9（Pulmonary Nodule Dataset with 9 Classes），其中包含了 8798 个肺结节病例和 40439 个标注的肺结节。本小节详细介绍了数据的采集和标注过程，分析了本数据集的属性。同时对现有的肺结节数据集进行了回顾，并与其进行了对比来说明本数据集 PN9 的优势。

3.3.1　数据收集与标注

3.3.1.1　数据的收集

本数据集 PN9 中的 CT 图像主要收集于两家三甲医院，覆盖了医院中的门诊部、住院部和体检部等不同部门的图像。CT 图像的采集时间跨度为 2015 年到 2019 年，共 5 年时间。对于通过电子计算机断层扫描仪扫描获得的初始 CT 图像，首先进行质量控制以确保挑选的医学图像均为医学数字成像和通信格式（Digital Imaging and Communications in Medicine，DICOM）。DICOM 是医学图像的国际标准，其定义了质量能够满足临床诊断使用并且适用于数据交换的医学图像格式，被广泛应用于心血管成像、放射医疗以及放射诊断设备（CT、X 射线和核磁共振等）。同时，对每个 CT 图像进行筛查，以排除其中含有物体干扰或严重呼吸运动伪影的图像，这些噪声会影响放射科医生对肺结节的诊断。此外为了保护医院、医生和患者的隐私，CT 图像中包含的部分信息被移除，包括医院的名称、就诊医生的姓名、患者的姓名等。

3.3.1.2 数据的标注

为了尽可能准确地标注所收集 CT 图像中的肺结节，本章采用了两阶段的标注过程，其中所有参与标注医学影像的医生均为各大医院的主治医师。第一个标注阶段是在医院进行的，主治医师会对每个患者的 CT 图像进行诊断，并由另一名医生进行核查，从而生成患者的 CT 检查报告。报告中包含患者肺部 CT 影像中每一个肺结节的类型、大小和大致位置，其不仅能给出每个患者的 CT 扫描检查结果，而且便于实施后续的标注过程。第二个阶段是详细标注，主治医师会参考第一标注阶段生成的医院检查报告，对每个病例的所有 CT 图像切片进行分析并标注出每张切片中的肺结节。对于医生识别出的一张 CT 图像切片上的肺结节，其边界框和类别信息会存储在一个单独的可扩展标记语言（Extensible Markup Language，XML）格式的文件中。通过参考现有的医学指南[23,216-217]，并考虑到医院和医生在临床诊断时的需求，本章根据肺结节的类型和大小将其分为 9 个不同的类别。主治医师在标注过程中，会根据所制定的分类标准对肺结节进行分类。一名医生标注完成的肺结节位置和类别信息，会由另一名医生进行审核和修改，从而形成最终的标注结果。如果第二个标注阶段的两位主治医生在标注肺结节时有意见不一致的地方，他们将通过进一步的讨论和筛查以确定最终的标注。

通过实施上述两阶段的标注过程，本章最终获得了 8798 个患者的 CT 扫描图像和 40439 个标注的肺结节。其中所有的 CT 图像都是从三甲医院中收集的，而且数据的统计分布也与临床情况一致。

3.3.2 数据集分析

3.3.2.1 电子计算机断层扫描机

为了提高本数据集的普适性，PN9 所收集的 CT 图像是利用一系列不同的 CT 设备扫描得到的，如图 3.5 （c）所示。PN9 数据集包含从通用电气医疗系统有限公司（GE Medical Systems）制造的 10 种 CT 扫描机获取的 2652 张 CT 图像、从西门子有限公司（Siemens）制造的 11 种 CT 扫描机获取的 2305 张 CT 图像、从东芝公司（Toshiba）制造的 3 种 CT 扫描机获取的 2224 张 CT 图像、从联影医疗科技股份有限公司（United Imaging Healthcare）制造的 2 种 CT 扫描机获取的 800 张 CT 图像、以及从飞利浦公司（Philips）制造的 6 种 CT 扫描机获取的 817 张 CT 图像。

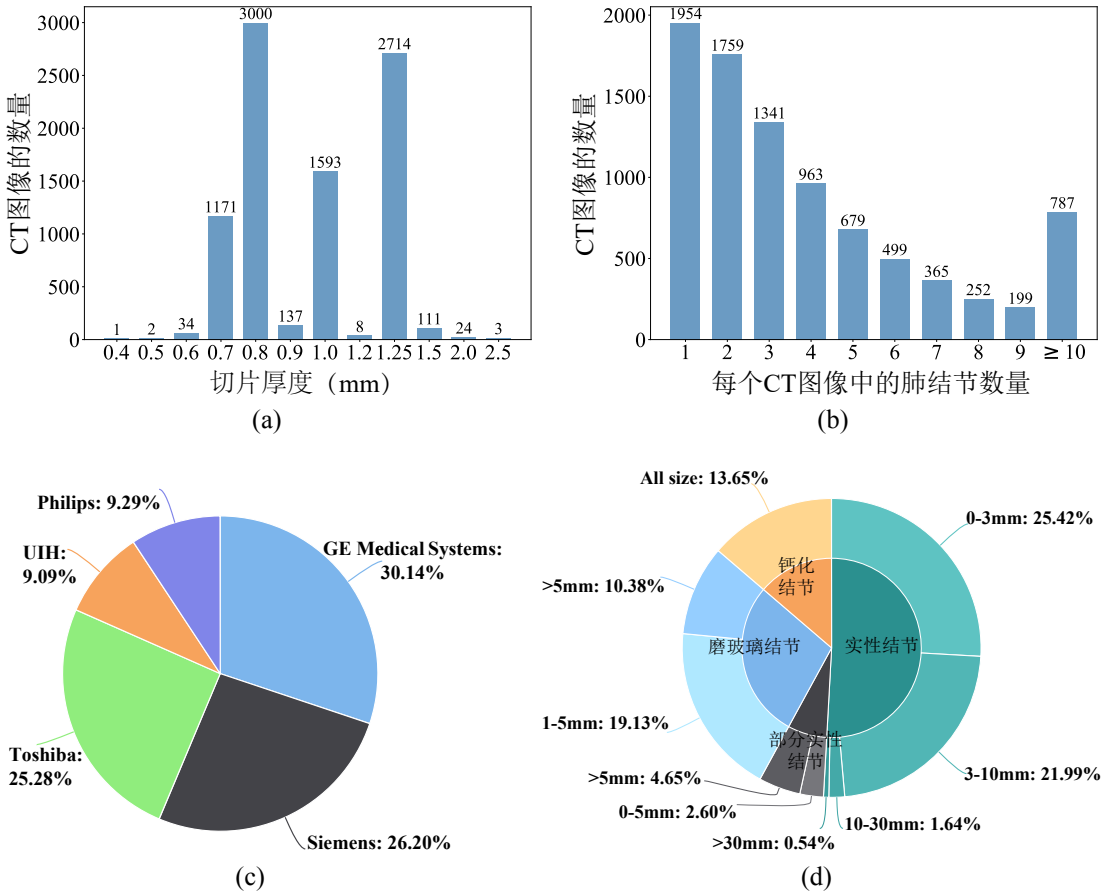


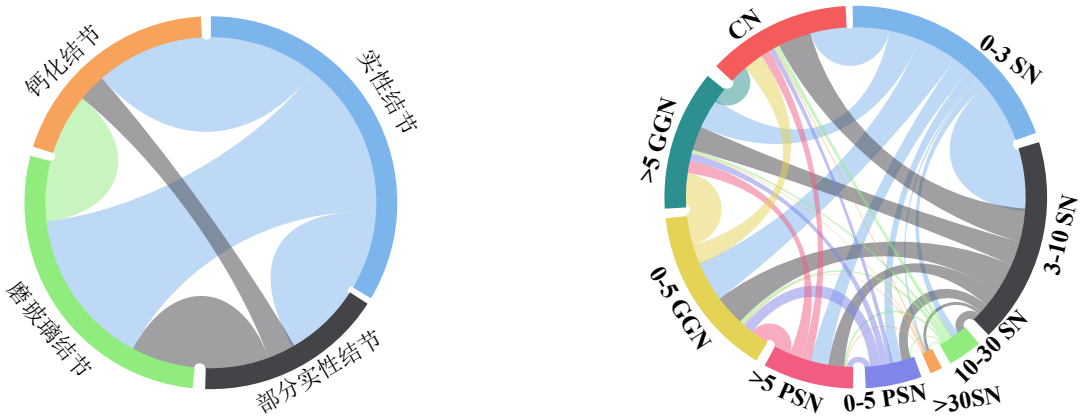
图 3.5 PN9 数据集的统计分析。（a）CT 图像的切片厚度分布；（b）每个 CT 图像中的肺结节数量分布；（c）所采用的 CT 设备的分布；（d）PN9 数据集的类别统计，其包含 4 个大类和 9 个子类，百分比表示某一类肺结节的数量占所有肺结节数量的比例。

3.3.2.2 CT 图像切片厚度

CT 图像中的切片厚度如果过厚，会对临床中医生对肺结节的诊断造成干扰[218-219]，因此本章主要收集了薄切片的 CT 图像。如图 3.5（a）所示，切片厚度的分布从 0.4mm 到 2.5mm 不等，大多数 CT 图像的切片厚度是 0.7、0.8、1.0 或 1.25mm，此外，CT 图像的像素间距分布从 0.310mm 到 1.091mm，平均像素间距为 0.706mm。

3.3.2.3 单张 CT 图像的结节数量

在图 3.5（b）中展示了单张 CT 图像中的肺结节数量分布。从图中可以观察到，PN9 数据集中大约 68% 的患者的肺结节数量小于 5 个，但是大约有 9% 的患者的肺结节数量超过 10 个，将这些肺结节全部检测出来的难度较大。



(a) (b)

图 3.6 PN9 数据集中的类别依赖关系。(a) 大类之间的相互依赖关系；(b) 子类之间的相互依赖关系。

3.3.2.4 类别

PN9 数据集具有分层次的类别结构，其详细分类如图 3.5 (d) 所示。根据肺结节的性质，PN9 数据集中的所有肺结节首先被分为四个大类，包括实性结节（Solid Nodule，SN）、部分实性结节（Part-Solid Nodule，PSN）、磨玻璃结节（Ground-Glass Nodule，GGN）和钙化结节（Calcific Nodule，CN）。同时，为了满足医院和医生在临床诊断时的实际需求，本章参照现有的医学指南[23,216-217]进一步细分大类。基于肺结节的尺寸，每个结节被分配为属于某个大类的子类。例如，0-3mm 实性结节子类（0-3mm Solid Nodule，0-3SN）定义为平面内尺寸分布在 0-3mm 范围内的实性结节。通过划分，得到了 9 种不同的子类，这些类别涵盖了临床中最常见的肺结节类型。然而，由于现实生活中患某些类型肺结节的病例较少，PN9 数据集中的数据分布是不平衡的。如图 3.5（d）的肺结节类别统计数据所示，小尺寸肺结节的数量比较多，同时一些类别比如部分实性结节的数量相对较少。这种不平衡的数量分布会导致模型对样本相对较多的肺结节有所偏向，此外大量的小尺寸结节也给肺结节的准确检测带来了挑战。

在图 3.6 (a-b) 中，本章分别分析了 PN9 数据集中大类和子类之间的相互依赖关系。其中两个类别之间的连接宽度越大，表明这两个类别的肺结节同时出现在一个患者身上的概率越高。例如，诊断为磨玻璃结节的患者有较大的概率同时患有实性结节。

表 3.1 与现有肺结节数据集的比较。

数据集	发表时间	CT 图像数量	肺结节数量	类别数	是否开源可用
ANODE09[200]	2010	55	710	4	是
LIDC-IDRI[207]	2011	1018	2562	3	是
LUNA16[208]	2016	888	1186	2	是
DSB 2017[47]	2017	2101	N/A	2	否
PN9	2020	8798	40439	9	是

3.3.3 与其他数据集的对比

目前已经有一些开源可用的肺结节数据集，例如 LIDC[220]、ANODE09[200]、LIDC-IDRI[207]、TCIA[221]以及 LUNA16[208]，这些数据集为研究人员提供了评价肺结节检测模型的统一指标。2010 年，Van 等人提出了 ANODE09[200]，其仅仅包含 55 张 CT 图像，且均是使用一种 CT 扫描仪扫描得到的。此外，尺寸较大的肺结节通常更有可能是恶性结节，及早的诊断出恶性结节有助于预防肺癌的发生。而 ANODE09 数据集中包含的大尺寸肺结节数量较少，这限制了其在实际临床中的应用。在此之后，几个数据量更大且包含大尺寸肺结节的数据集相继被提出，LIDC-IDRI 数据集[207]包含了由四个经验丰富的放射科医生标注的 1018 张 CT 扫描图像，这些 CT 图像是从七个不同的学术机构以及利用一系列 CT 扫描仪器收集得到的。其中的肺结节被划分为三个类别：尺寸 ⩾3mm 的肺结节、尺寸 <3mm 的肺结节、非肺结节。在对数据集进行标注时，LIDC-IDRI 数据集仅将尺寸 ⩾3mm 的肺结节进行了人工的三维分割。LUNA16 数据集[208]是基于 LIDC-IDRI 数据集[207] 中的 CT 图像收集得到的，其中 CT 图像切片厚度大于 3mm 的数据被丢弃。实际上，LUNA16 数据集[208]中所有 CT 图像的切片厚度都小于 2.5mm。此外，CT 图像缺少切片或者像素间距不一致的数据也被移除。LUNA16 数据集最终包含 888 个 CT 扫描图像，其中有 1186 个被医生标注出来的肺结节，而且所有肺结节的尺寸都 ⩾3mm。DSB 2017[47]包含 2101 个 CT 扫描图像，但是这个数据集只包括两个类别的标注，用于指示一个病例是否被诊断为肺癌。由于不同类型的肺结节有不同的形态表现和癌变概率，目前已有的这些数据集包含的标注肺结节较少，而且肺结节的种类也很有限，因此在实际应用中不足以支持临床肺癌诊断的需求。

在表 3.1中，本章将 PN9 数据集与现有的几个肺结节检测数据集进行了比

较。与目前广泛使用的数据集 LUNA16[208]相比，PN9 数据集包含数量超过其 10 倍的 CT 图像和数量超过其 30 倍的标注肺结节。至于肺结节类别的多样性，其他数据集通常只有三个类别：尺寸 ⩾3mm 的肺结节、尺寸 <3mm 的肺结节和非结节[207-208]。由于上述这些限制，大多数现有的肺结节检测数据集难以应用于临床实践。相比之下，PN9 数据集包含大量的 CT 扫描图像和 9 个肺结节类别，这将有助于肺结节的检测和分类任务，允许研究人员基于不同类型的肺结节设计更加有效的检测算法。此外，PN9 数据集包含有更多的小尺寸肺结节，如 0-3mm 实性结节和 0-5mm 毛玻璃结节。因此其有助于更准确地识别小尺寸肺结节，这样医生就可以更早地诊断和治疗患者。总之，本章提出的 PN9 数据集不仅数据量比现有的肺结节检测数据集更大，而且具有更丰富的多样性和更大的检测难度。

第四节 实验结果与分析

本小节将介绍本章实验的评测指标和实现细节，同时也在本章提出的 PN9 数据集和公开数据集上与现有的肺结节检测方法进行了对比。此外，为了更好的理解本章模型中各个模块的重要性，本小节也详细地对模型中的每个模块进行了消融实验。

3.4.1 评测指标

FROC（Free-Response Receiver Operating Characteristic）是 LUNA16 数据集[208]提出的官方评测指标，其定义为每张 CT 图像在假阳性为 0.125，0.25，0.5，1，2，4，8 时的平均召回率。当一个候选肺结节位于距离真值中任何肺结节中心 R 的距离内时，将其视为真阳性，不在真值中任何肺结节范围内的候选结节被视为假阳性，其中 R 表示真值中肺结节的半径。本章的实验中采用了这一评测指标，并进一步将 FROC 拓展为 FROC_{IoU}，其定义为候选肺结节和真值中任何结节的三维交并比（3D Intersection over Union，3D IOU）高于一个阈值（在本章实验中，三维交并比的阈值被定义为 0.25）时的真阳性。

此外，本章还采用了三维平均精度均值（3D mean Average Precision，3D mAP）作为肺结节检测的评测指标。根据三维目标检测的特点和 PN9 数据集的属性，本章定义了以下五个评测指标：AP@0.25（3D IoU = 0.25 时的 AP）、AP@0.35（3D IoU = 0.35 时的 AP）、AP_s（对应于尺寸为 0-5mm 且体积 <512 的

小肺结节的 AP）、AP_m（对应于尺寸为 5-10mm 且 512< 体积 <4096 的中肺结节的 AP）、AP_l（对应尺寸 >10mm 且体积 >4096 的大肺结节的 AP）。由于肺结节检测是一项三维目标检测任务，对于比较实验中的几种二维目标检测方法，需要使用类似于文献[222]中的方法将二维目标检测结果合并成三维的检测结果。

3.4.2 实现细节

3.4.2.1 数据预处理

对于 PN9 数据集的 8798 张 CT 图像，本章将其分为 6707 张训练图像和 2091 张测试图像。在训练过程中，从训练集中划分出 670 张 CT 图像作为验证集，以监测模型的收敛性。此外，需要对原始的 CT 图像进行预处理：第一步是将所有的原始数据都转换成亨氏单位（Hounsfield Unit，HU），因为亨氏单位是描述无线电密度的标准定量值；第二步是将 CT 图像依据窗宽窗位剪切到 [-1200，600]；最后将 CT 图像的数据范围线性转换到 [0，255]。

3.4.2.2 图像块输入

对于三维卷积神经网络来说，由于计算机硬件 GPU 的显存限制，在训练过程中使用整张 CT 图像作为输入是不可行的。本章从 CT 图像中裁剪出小的三维图像块，并将它们单独输入网络，其大小为 $128 \times 128 \times 128 \times 1$（深度 × 高度 × 宽度 × 通道）。如果一个图像块超出了 CT 图像的尺寸范围，就用值 170 进行填充，这个值是常见组织的亮度且可以和肺结节区分开来。在测试阶段，本章将一张 CT 图像全部输入网络，不对其进行裁剪。为了避免整个三维 CT 图像的尺寸大小为奇数，在输入到模型之前，这些图像被填充了值 170。

3.4.2.3 网络配置

对于本章提出的基于切片关联注意力的肺结节检测模型，使用批大小（Batch Size）为 16 的随机梯度下降（Stochastic Gradient Descent，SGD）优化器。初始学习率（Initialization Learning Rate）被设置为 0.01，动量（momentum）和权重衰减系数（Weight Decay Coefficient）分别被设置为 0.9 和 1×10^{-4}。设置训练 200 个 epoch（当一个数据集的图片全部输入神经网络一次并返回一次，这个过程被称为一个 epoch。），100 个 epoch 后学习率降至 0.001，再过 60 个 epoch 学习率降至 0.0001。本章的模型是基于深度学习框架 pytorch[223]实现的，并利用 4 块 24GB 显存的 NVIDIA RTX TITAN GPU 进行训练和测试。

表 3.2 本章方法 SANet 和其他方法在 PN9 数据集上基于评测指标 FROC 的比较结果。表中数值为肺结节检测的敏感度（单位：%），每列表示每张 CT 图像的平均假阳性。

方法	0.125	0.25	0.5	1.0	2.0	4.0	8.0	平均值
二维卷积神经网络								
Faster R-CNN[35]	10.79	15.78	23.22	32.88	46.57	61.94	75.52	38.10
RetinaNet[52]	8.42	13.01	20.13	29.06	40.41	52.52	65.42	32.71
SSD512[36]	12.26	18.78	28.00	40.32	56.89	73.18	86.48	45.13
三维卷积神经网络								
Leaky Noisy-OR[47]	28.08	36.42	46.99	56.72	66.08	73.77	81.71	55.68
3D Faster R-CNN[45]	27.57	36.59	46.76	58.00	70.00	80.02	88.32	58.18
DeepLung[45]	28.59	39.08	50.17	62.28	72.60	82.00	88.64	60.48
NoduleNet (N_2)[224]	27.33	38.25	49.40	61.09	73.11	83.28	89.83	60.33
I3DR-Net[50]	23.99	34.37	46.80	60.04	72.88	83.60	89.57	58.75
DeepSEED[43]	29.21	40.64	51.15	62.20	73.82	83.24	89.70	61.42
SANet	**38.08**	**45.05**	**54.46**	**64.50**	**75.33**	**83.86**	**89.96**	**64.46**

表 3.3 本章方法 SANet 和其他方法在 PN9 数据集上基于评测指标 FROC_{IoU} 的比较结果（%）。

方法	0.125	0.25	0.5	1.0	2.0	4.0	8.0	平均值
二维卷积神经网络:								
Faster R-CNN[35]	3.41	6.97	12.26	20.58	33.05	46.41	57.90	25.80
RetinaNet[52]	2.60	5.56	10.95	19.25	29.29	40.49	51.05	22.74
SSD512[36]	4.62	8.48	14.76	25.06	40.32	57.27	70.80	31.61
三维卷积神经网络:								
NoduleNet (N_2)[224]	21.17	30.23	40.38	51.02	61.26	70.70	76.93	50.24
I3DR-Net[50]	15.64	23.13	37.00	51.54	64.54	72.91	77.53	48.90
SANet	**26.72**	**36.03**	**47.46**	**56.99**	**66.35**	**73.52**	**78.32**	**55.06**

3.4.3 与现有方法的对比

本小节将本章的基于切片关联注意力的肺结节检测方法（SANet）与现有的几种性能较好的方法进行了比较，包括基于二维卷积神经网络的方法 Faster R-CNN[35]、Retianet[52]、SSD512[36]和基于三维卷积神经网络的方法 noise-OR[47]、3D Faster R-CNN[45]、DeepLung[45]、NoduleNet(N_2)[224]、I3DR-Net[50]以及 DeepSEED[43]。

表 3.4 本章方法 SANet 和其他方法在 LUNA16 数据集[208]上基于评测指标 FROC 的比较结果（%）。

方法	0.125	0.25	0.5	1.0	2.0	4.0	8.0	平均值
Leaky Noisy-OR[47]	59.38	72.66	78.13	84.38	87.50	89.06	89.84	80.13
3D Faster R-CNN[45]	66.20	74.60	81.50	86.40	90.20	91.80	93.20	83.40
DeepLung[45]	69.20	76.90	82.40	86.50	89.30	91.70	93.30	84.20
NoduleNet (N_2)[224]	65.18	76.79	83.93	87.50	91.07	92.86	93.75	84.43
I3DR-Net[50]	63.56	71.31	79.84	85.27	87.60	89.92	91.47	81.28
DeepSEED[43]	**73.90**	**80.30**	85.80	88.80	90.70	91.60	92.00	86.20
SANet	71.17	80.18	**86.49**	**90.09**	**93.69**	**94.59**	**95.50**	**87.39**

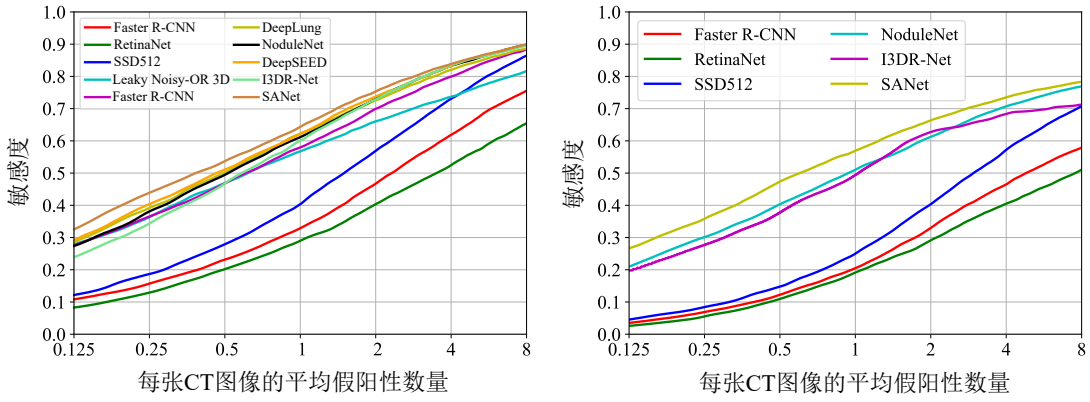
表 3.5 本章方法 SANet 和 NoduleNet[224]基于评测指标 AP 的比较结果。

方法	AP@0.25	AP@0.35	AP_s	AP_m	AP_l
NoduleNet (N_2)[224]	46.7	30.2	12.8	45.3	46.4
SANet	**52.2**	**36.6**	**14.1**	**47.6**	**48.6**

3.4.3.1 基于评测指标 FROC 的比较

本章方法 SANet 和其他方法在 PN9 数据集上基于评测指标 FROC 的比较结果如 表 3.2，FROC 曲线如 图 3.7（a）所示。可以看到，SANet 获得了优于其他所有方法的最好结果，与性能第二好的方法 DeepSEED[43]相比，SANet 的平均 FROC 分数提高了 3.04%。同时在每张 CT 图像的平均假阳性数量小于 2 时，SANet 的性能要明显优于其他检测方法。此外，其他基于三维卷积神经网络的方法，比如 NoduleNet (N_2)[224] 和 DeepLung[45]，也获得了较好的实验结果。可以看出，基于三维卷积神经网络方法的 FROC 分数要明显优于基于二维卷积神经网络的方法。例如，本章提出的方法 SANet 将 SSD512[36]和 Faster R-CNN[35]的平均 FROC 分数分别提高了 19.33% 和 26.36%。由于基于二维卷积神经网络的方法仅利用了包含三个通道的输入图像，其学习的空间信息不足，因此它们对于三维肺结节检测的性能较差。

为了进一步验证所提出方法 SANet 的性能，本章在目前被广泛使用的肺结节检测数据集 LUNA16[208]上进行了 10 折交叉验证的实验。如表 3.4所示，SANet 在检测肺结节时的性能是最好的，其获得了 87.39% 的平均 FROC 分数，比第二好的方法 DeepSEED[43] 高了 1.19%。此外，对于平均假阳性数量大于 1 的 CT 图像，SANet 的性能显著优于其他的肺结节检测方法。



(a) (b)

图 3.7 本章方法 SANet 和其他方法的 FROC、FROC$_{IoU}$ 曲线比较。(a) 本章方法 SANet 和其他方法的 FROC 曲线比较；(b) 本章方法 SANet 和其他方法的 FROC$_{IoU}$ 曲线比较。

表 3.6 对本章提出的切片分组非局部模块和假阳性抑制模块的消融实验（%）。基准模型为基于三维 ResNet-50 和三维 RPN 的检测框架（编号 1），将切片分组非局部模块和假阳性抑制模块分别加入基准模型来验证模块的有效性（编号 2 和编号 3），编号 4 是本章提出的方法 SANet 的实验结果。p 为 FROC 的统计学显著性检验（与 SANet 相比）。

编号	SGNL	FPR	FROC	FROC$_{IoU}$	AP@0.25	p 值
1			61.29	51.96	49.0	0.009
2	✔		64.34	53.44	51.3	0.081
3		✔	62.69	52.51	50.2	0.015
4	✔	✔	**64.46**	**55.06**	**52.2**	–

3.4.3.2 基于评测指标 FROC$_{IoU}$ 的比较

在评测指标 FROC 当中，如果候选肺结节位于距真值中任何肺结节中心一定距离的位置，则该候选肺结节被定义为真阳性。本章进一步将评测指标 FROC 拓展为 FROC$_{IoU}$，其基于候选肺结节和真值肺结节的 3D IoU 来定义真阳性，实验结果如表 3.3和 图 3.7（b）所示。可以看出，本章方法 SANet 在评测指标 FROC$_{IoU}$ 上达到了 55.06%，这一结果优于其他方法，并高于 Nodulenet(N$_2$)[224]4.82%。值得注意的是，基于三维卷积神经网络的方法在评测指标 FROC$_{IoU}$ 上也优于基于二维卷积神经网络的方法。此外，leaky noisy-OR 的[47]、3D Faster R—CNN[45]、DeepLung[45]和 DeepSEED[43]等方法的 FROC$_{IoU}$ 结果并没有被列出，因为这些方法预测的中心坐标和直径与真值中肺结节的三维位置不匹配。

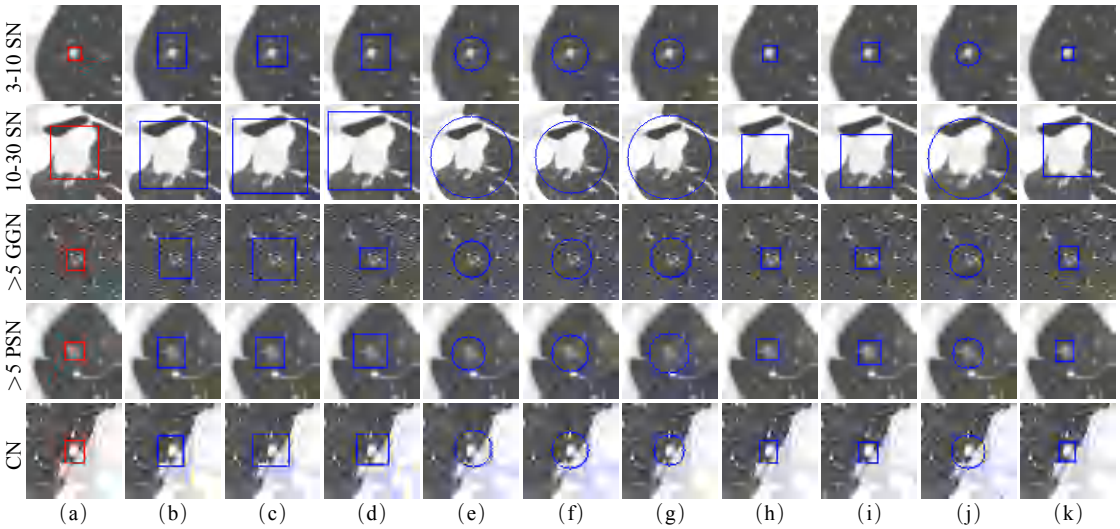


图 3.8 本章方法 SANet 和其他方法对于肺结节中心切片的定性化比较。第一行到第五行展示了不同类别肺结节的比较结果，分别是：3-10 SN、10-30 SN、>5 GGN、> 5 PSN 和 CN。每列分别表示（a）真值；（b）-（k）Faster R-CNN[35]、RetinaNet[52]、SSD512[36]、Leaky Noisy-OR[47]、3D Faster R-CNN[45]、DeepLung[45]、NoduleNet (N_2)[224]、I3DR-Net[50]、DeepSEED[43]和 SANet 的检测结果。

3.4.3.3 基于评测指标 AP 的比较

由于尺寸较大的肺结节通常更有可能是恶性肿瘤，所以结节的大小对肺癌的诊断来说很重要。本章基于 PN9 数据集，采用 3D AP 定义了几个评测指标，表 3.5中列出了 NoduleNet (N_2)[224]和本章方法 SANet 的结果。可以看出，SANet 在 AP@0.25 指标上将 NoduleNet 提高了 5.5。此外，考虑到肺结节的尺寸大小，本章定义了三个指标 AP_s、AP_m 和 AP_l 来分别评估小、中和大尺寸肺结节的检测性能。SANet 在所有三个指标上都取得了较好的结果，证明了其在检测不同大小的肺结节时是有效的。

3.4.3.4 可视化结果比较

本章方法 SANet 和其他方法对于肺结节中心切片的可视化结果比较如图 3.8所示。对于五种类型的肺结节，SANet 检测到的肺结节位置与真值标注基本一致，而其他方法得到的检测结果通常会偏离或大于标注框，尤其是基于二维卷积神经网络的方法。这些实验结果验证了本章提出的 SANet 在肺结节检测任务中的优越性。

表 3.7 将本章提出的切片分组非局部模块和假阳性抑制模块加入到其他方法的消融实验（%）。

模型	SGNL	FPR	FROC	FROC$_{IoU}$
DeepLung[45]			60.48	–
DeepLung[45]	✔		61.67	–
DeepLung[45]		✔	61.15	–
DeepLung[45]	✔	✔	62.06	–
NoduleNet (N_2)[224]			60.33	50.24
NoduleNet (N_2)[224]	✔		62.35	51.13
NoduleNet (N_2)[224]		✔	61.18	50.99
NoduleNet (N_2)[224]	✔	✔	**62.69**	**52.37**

表 3.8 对切片分组非局部模块的不同配置进行的消融实验（%）。p 为 FROC 的统计学显著性检验（与 5 个 SGNL 残差块相比）。

编号	SGNL 残差块	FROC	FROC$_{IoU}$	AP@0.25	p-value
1	4 个残差块	62.97	53.45	50.8	0.023
2	5 个残差块	**64.46**	**55.06**	**52.2**	–
3	6 个残差块	63.12	53.26	51.2	0.067
4	7 个残差块	63.68	53.51	51.5	0.057
5	10 个残差块	64.20	53.92	51.1	0.079

3.4.4 消融实验

3.4.4.1 模块的有效性

在本章所提出的方法 SANet 中，有两个重要的模块：切片分组非局部模块（Slice Grouped Non-Local，SGNL）和假阳性抑制模块（False Positive Reduction，FPR）。为了验证这两个模块的性能，本小节进行了不同的实验，如 表 3.6所示。其中编号 1 是基于三维 ResNet-50 和三维 RPN 的基准检测模型，将本章提出的切片分组非局部模块和假阳性抑制模块分别加入到基准模型后，评测指标 FROC 分数获得了 3.05% 和 1.40% 的提高，上述结果证实了这两个模块都有助于肺结节的检测。此外，如果将切片分组非局部模块和假阳性抑制模块结合，在 FROC 分数方面取得了 3.17% 的提高，同时评测指标 FROC$_{IoU}$ 和 AP@0.25 也通过应用这两个模块得到了提高。本小节还将这两个模块加入到其他方法中来进一步验证其性能，包括：DeepLung[45] 和 Nodulenet(N2)[224]。如表 3.7中所列，应用本章提出的切片分组非局部模块和假阳性抑制模块，NoduleNet(N_2)[224]的 FROC

表 3.9 对切片分组非局部模块进行不同数量分组 G 的消融实验（%）。

分组数 G	FROC	FROC_{IoU}	AP@0.25	AP_s	AP_m	AP_l
1	62.88	53.73	50.6	10.5	**48.2**	48.5
4	**64.46**	**55.06**	**52.2**	14.1	47.6	**48.6**
8	63.80	54.62	51.2	**15.6**	46.6	47.8

分数分别提高了 2.02% 和 0.85%，DeepLung[45]的性能也通过增加这两个模块得到了提高。这些结果验证了本章提出的切片分组非局部模块和假阳性抑制模块在肺结节检测任务中是有效的。

3.4.4.2 切片分组非局部模块中不同参数的影响

表 3.8中展示了具有不同配置的切片分组非局部模块的实验结果，本小节将 4 个 SGNL 残差块（3D ResNet-50 的第二个残差阶段 *res*2 到第五个残差阶段 *res*5 的倒数第二个残差块）、5 个 SGNL 残差块（2 个在 *res*3，3 个在 *res*4，分别间隔一个残差块）、6 个 SGNL 残差块（*res*4 中的每个残差块）、7 个 SGNL 残差块（1 个在 *res*2，2 个在 *res*3，3 个在 *res*4，1 个在 *res*5，分别间隔一个残差块）和 10 个 SGNL 残差块（*res*3 和 *res*4 中的每个残差块）替换到 3D ResNet-50 中。与表 3.6 中的编号 3 相比，添加不同配置的 SGNL 残差块在三个评测指标上都带来了改进。其中 5 个 SGNL 残差块的结果最好，在 FROC 上提高了 1.77%，在 FROC_{IoU} 上提高了 2.55%。

本小节还分析了切片分组非局部模块中不同分组数量 G 的影响，如 表 3.9所示。当分组数 $G=4$ 时，评测指标 FROC 达到了 64.46%，比另外两个配置 $G=1$ 和 $G=8$ 分别提高了 1.58% 和 0.66%。同时可以看到，当 $G=8$ 时 AP_s 分数最好，当 $G=1$ 时 AP_m 最高，这些实验结果符合预期。切片分组操作是为了获取一组特征图中任意位置和任意通道之间的关系，增强对不同尺寸肺结节的辨别能力。组数越少，每组包含的连续切片越多，有利于检测大肺结节，但限制了小肺结节的检测。当划分太多组数时，每组包含很少的切片，限制了大尺寸肺结节的检测。由于分组数 $G=4$ 时模型的整体表现最好，不同大小肺结节的 AP 都相近，所以本章将切片分组非局部模块中的分组数设置为 $G=4$。

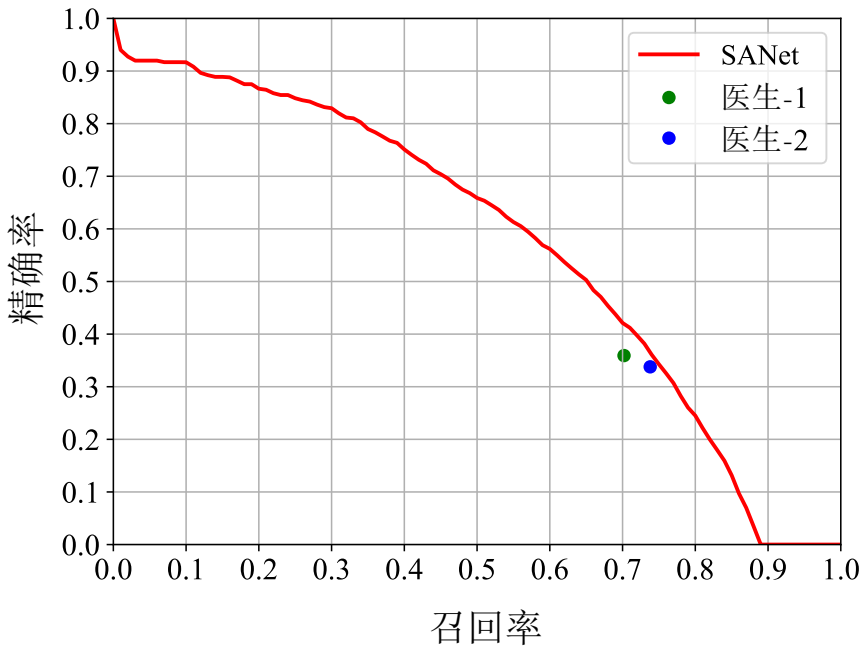


图 3.9 基于小尺度肺结节检测数据集的结节检测精确率-召回率曲线（Precision-Recall Curve，PR Curve）。“医生-1”和“医生-2”表示两位经验丰富医生的检测结果。

表 3.10 不同的 CT 设备制造厂商对模型性能的影响分析。

CT 设备制造厂商	GE	Philips	Siemens	Toshiba	UIH
测试集中的 CT 图像数量	657	174	523	557	180
FROC	66.63	65.07	65.73	60.65	65.50

3.4.5 讨论

3.4.5.1 与专业医生的比较

为了进一步验证本章方法 SANet 在检测肺结节时的性能，本小节将 SANet 与两名具有至少 10 年临床经验的医生进行了比较。本章收集了一个额外的小规模肺结节测试数据集，其中包含 120 张 CT 图像，经过三甲医院几名主治医生的精确注释，这个测试数据集最终包含 2137 个带有金标准的标注肺结节。另外两名未参与诊断小规模测试数据集的经验丰富的医生，被邀请单独识别 CT 图像中的肺结节，每个医生将识别出来的肺结节标注一个三维边界框和类别名称。如果一个候选结节和金标准的 3D IoU 高于一个阈值，则该候选结节将被认为是真阳性。然后就可以得到两位经验丰富医生的肺结节检测结果：医生 1 的精确率为 35.92%，召回率为 70.20%；医生 2 的精确率为 33.78%，召回率为 73.80%。可以看出，两名医生的检测精确率都不高，这是目前肺结节诊断和治疗的主要

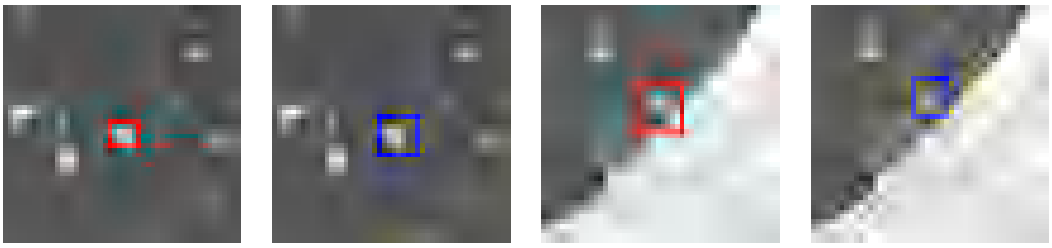


图 3.10 本章方法 SANet 的失败案例分析。前两列为类别 0-3 SN 的肺结节，后两列为类别 0-5 PSN 的肺结节。

挑战之一。至于本章的 SANet，是在 PN9 数据集上训练，在这个小规模测试数据集上测试，使用指标 AP@0.25 进行评测。

本章方法 SANet 的精确率-召回率曲线和两位经验丰富医生的检测结果如位图 3.9所示。可以看出，SANet 的性能要优于两位医生单独诊断肺结节的结果，这验证了 SANet 的性能可以超过医生水平且适用于肺结节检测。临床诊断中，有些肺结节的尺寸很小，而且不同的肺结节有不同的形态。医生通常可以识别出 CT 图像中相对明显的肺结节，但却不能识别出全部的肺结节，尤其是大量的小尺寸肺结节。对于包含在多个连续切片中的肺结节，医生通常会遗漏一些切片进而影响肺结节检测的性能。本章提出的肺结节检测模型 SANet 能够花费比医生更少的时间来识别结节，因此医生可以利用 SANet 从而以更高的效率和准确性来诊断肺结节，这将进一步帮助肺癌的早期诊断和治疗。同时，希望本章提出的 PN9 数据集可以促进未来对肺结节检测的研究，并有助于其在临床中的应用。

3.4.5.2 失败案例分析

如上述实验所示，本章所提出的方法 SANet 在 PN9 数据集和公开数据集 LUNA16[208]上都取得了比其他检测方法更好的结果。同时 SANet 检测肺结节的性能也要比两位经验丰富的医生更好，然而本章的方法仍然有一些失败的案例。如图 3.10所示，当识别 0-3 SN 类别的肺结节时，SANet 会生成比真实标注更大的边界框。此外，由于部分实性结节通常具有模糊的边界，SANet 可能无法识别完整的结节，并且产生比真实标注更小的边界框。下一步可以考虑结合不同类别的属性来更好地检测肺结节。

3.4.5.3 不同 CT 设备厂商的影响

本小节分析了不同 CT 设备厂商对肺结节检测的影响。表 3.10中是本章方法 SANet 在不同 CT 设备制造厂商的测试集中进行实验的结果，其中除了东芝之外，其他设备厂商的结果差别都很小。其原因是利用东芝 CT 设备获得的测试集中 CT 图像恰好比其他厂商包含更多的小结节，影响了其性能。总的来说，因为本章提出的 PN9 数据集中包含足够多来自不同设备制造商的 CT 图像，不同 CT 设备制造厂商对肺结节检测结果的影响很小。

第五节 本章小结

本章提出了一个新的大规模肺结节检测数据集 PN9，其包含 8798 张 CT 图像、9 种常见的肺结节类别和 40439 个标注肺结节。与现有肺结节数据集相比，PN9 包含了数据量更大且肺结节种类更多的 CT 图像，其能够使得研究人员基于肺结节的丰富属性设计更加有效的算法。同时，PN9 所标注的更多小尺寸肺结节有助于提高在小结节上的诊断准确性，进而能够更早的发现患者的疾病并及早的进行治疗。此外，本章还提出了一种切片关联注意力网络用于肺结节的检测。借鉴医生临床诊断肺结节的方式，本章设计了一种切片分组非局部模块并将其添加到编码器网络中，其能够捕获一个切片组内特征图中任意位置和任意通道之间的长程依赖信息。三维区域候选网络对肺结节的检测具有较高的灵敏度，但通常会带来大量的假阳性样例，所以本章提出了基于多尺度特征图的假阳性抑制模块来进一步优化肺结节检测的结果。本章在 PN9 数据集和公开数据集 LUNA16 上，与几种目前较好的基于二维卷积神经网络和三维卷积神经网络的方法进行了比较，SANet 均取得了更优越的性能。同时本章也开展了大量的消融实验来验证所提出的模块在肺结节检测任务中的有效性。希望本章提出的 PN9 数据集和肺结节检测模型 SANet 能够促进未来的肺结节检测研究，并进一步帮助人工智能在医疗图像中的应用。

第四章　基于拓扑连通注意力的遥感图像道路提取

道路在平面维度上连续分布，会呈现细长的形状。道路遥感图像可以看做是多张顺序排列的道路子图像的集合，同时不同子图像共同形成了完整且连通的道路，属于平面空间序列图像。在识别道路的研究中，如何缓解其他地物遮挡或复杂交通场景的影响，并利用道路的拓扑序列信息来保证道路的连通性，是一个难点。本章提出了拓扑连通注意力网络，能够直接从高分辨率遥感影像中提取出连通性较好的道路。本章的章节安排为：第一节是对本章研究背景、研究内容和创新点的介绍；第二节介绍了所提出的基于拓扑连通注意力网络的道路提取模型；第三节给出了本章所提出的道路提取模型的对比实验及相关分析；第四节是对本章内容进行的总结。

第一节　引言

道路网络的创建是多个应用领域基础且必不可少的步骤，包括：自动驾驶、城市规划、车辆导航和地理信息更新等。目前一些地图公司采用的地图采集方法通常很耗时，例如从激光雷达点云中提取道路、全球定位系统（Global Positioning System，GPS）轨迹的聚合以及人工的道路标注等。这些方法不适用于大面积区域内道路的提取，且不足以在快速变化的现实环境中更新道路网络的动态变化[88,225]。高分辨率遥感影像不仅可以呈现道路的几何特征，还可以提供多个时期甚至实时的影像，这为道路的快速更新提供了条件，因此从遥感影像中提取道路网络[57-59]已成为目前主流的方法。

传统的研究侧重于利用手工设计的特征并定义一些准则来从遥感图像中提取道路[65-66,70,226]。然而在处理大范围区域的遥感图像时，这些方法通常效率比较低。近年来，卷积神经网络（CNN）特别是基于全卷积网络（Fully-Convolutional Network，FCN）[73]架构的模型被提出并在图像语义分割任务中取得了较好的性能[46,74-75,85,190-191]。一些研究将具有编码器-解码器架构的 CNN 应用于道路分割任务[79,81-83]，这些方法通常能够获得良好的道路分割结果。然而，从遥感图像中提取道路是具有挑战性的任务，由于下列几项原因：（a）建筑物、树木以及阴影对道路区域的遮挡；（b）复杂的城市环境和交通场景；（c）道路与其他

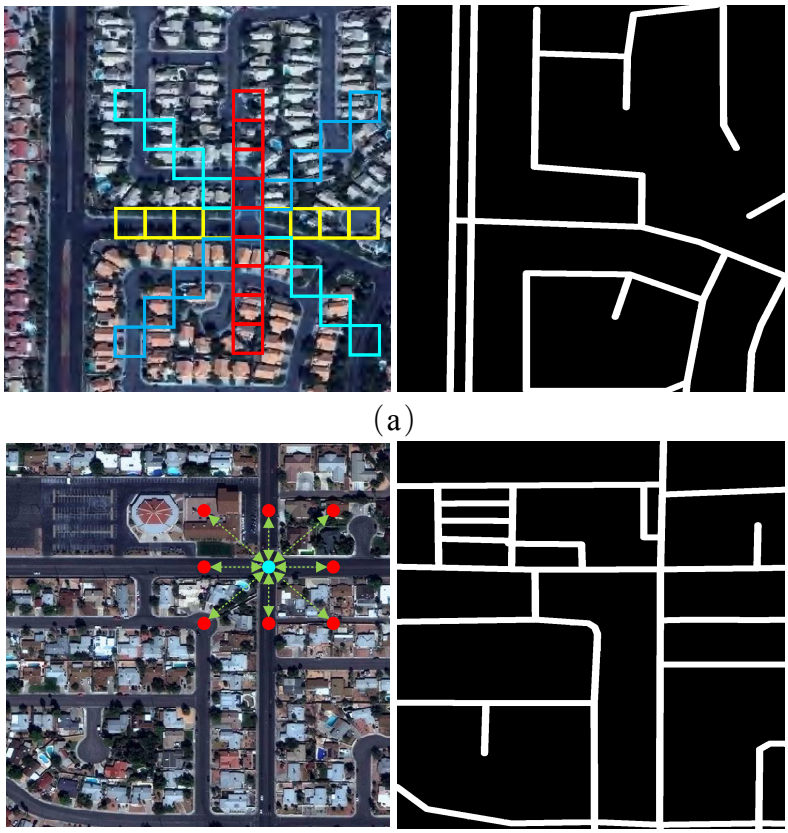


图 4.1 本章提出的条形卷积模块和连通性注意力模块示意图。从左到右：遥感影像、真值。(a) 用于学习道路线性特征的具有四种形状的条形卷积。(b) 预测一个像素与其相邻像素的连通性以捕获局部的成对依赖关系并确保道路拓扑的正确性。

地物（例如建筑屋顶以及停车场等）之间的相似性。这些困难会导致道路分割结果碎片化，因此上述方法很难保证道路的连通性。一些方法[87-88,227]通过使用道路分割和后处理步骤来优化道路缺失连接的情况，其中最短路径算法通常被用作后处理操作，但是这些算法不能适用于复杂的道路环境。为了直接获得连通性较好的道路，Bastani 等[58]采用迭代搜索过程自动提取道路网络，Mosinska 等[89]利用 U-Net[46]并结合多种损失函数对道路轮廓进行迭代优化。此外，Liu 等[90]集成了包括道路面、边缘和中心线在内的多级别特征，以提高道路预测的结果。然而，这些方法通常比较耗时而且需要复杂的步骤来训练。

本章提出了一种用于从遥感影像中提取道路的拓扑连通注意力网络（Connectivity Attention Network，CoANet）。首先引入编码器-解码器架构网络来学习道路的特征，其中采用空洞空间金字塔池化模块（Atrous Spatial Pyramid

Pooling，ASPP）来增加特征点的感受野并获取多尺度特征。由于道路的形状通常呈现为跨度大、狭窄且连续分布，利用道路的这一特性本章设计了一种条形卷积模块（Strip Convolution Module，SCM），并将其放置于解码器网络中。如图 4.1（a）所示，条形卷积模块利用水平、垂直、左对角线和右对角线等四个条形卷积从四个不同方向来获取道路的长距离上下文信息，同时其还可以抑制不相关区域对特征学习的干扰。为了缓解建筑物或树木对道路区域的遮挡问题，本章提出了一种连通性注意力模块（Connectivity Attention Module，CoA）来探索图像中相邻像素之间的关系。如图 4.1 （b）所示，连通性注意力模块能够预测给定像素与其周围八个相邻像素的连通性，从而实现道路拓扑连接的正确性。本章利用基于像素和基于图的评测指标在多个公开数据集上进行了大量实验，验证了所提出的 CoANet 与其他道路提取方法相比的优越性。

本章研究工作的主要贡献包括：

- 提出了拓扑连通注意力网络，联合学习道路分割和相邻像素之间的图关系来提高道路连通性，在公开数据集上取得了相比于其他方法的显著改进。
- 设计了一种条形卷积模块，利用四个不同方向的条形卷积来捕获长距离上下文信息并避免来自不相关区域的干扰。
- 提出了一种连通性注意力模块，通过利用成对相邻像素之间的依赖关系来结合图信息并提高道路的连通性。

第二节 基于拓扑连通注意力网络的道路提取

道路连通性是一个重要的道路特征，然而基于语义分割的方法通常会产生支离破碎的道路，不能满足实际需求。为了缓解这个问题，本章提出了一种拓扑连通注意力网络（Connectivity Attention Network，CoANet），用于从遥感影像中提取道路，如图 4.2 所示。在拓扑连通注意力网络中，本章设计了一种条形卷积模块（Strip Convolution Module，SCM）来适应道路的细长形状并提取其线性特征。同时进一步提出了一个连通性注意力模块（Connectivity Attention，CoA），以预测相邻像素之间的道路连通性。

4.2.1 网络结构

在拓扑连通注意力网络中，本章采用了在 ImageNet[228]上预训练的 ResNet-101[137]作为编码器。考虑到空洞卷积是控制卷积核感受野和调整特征图分辨率

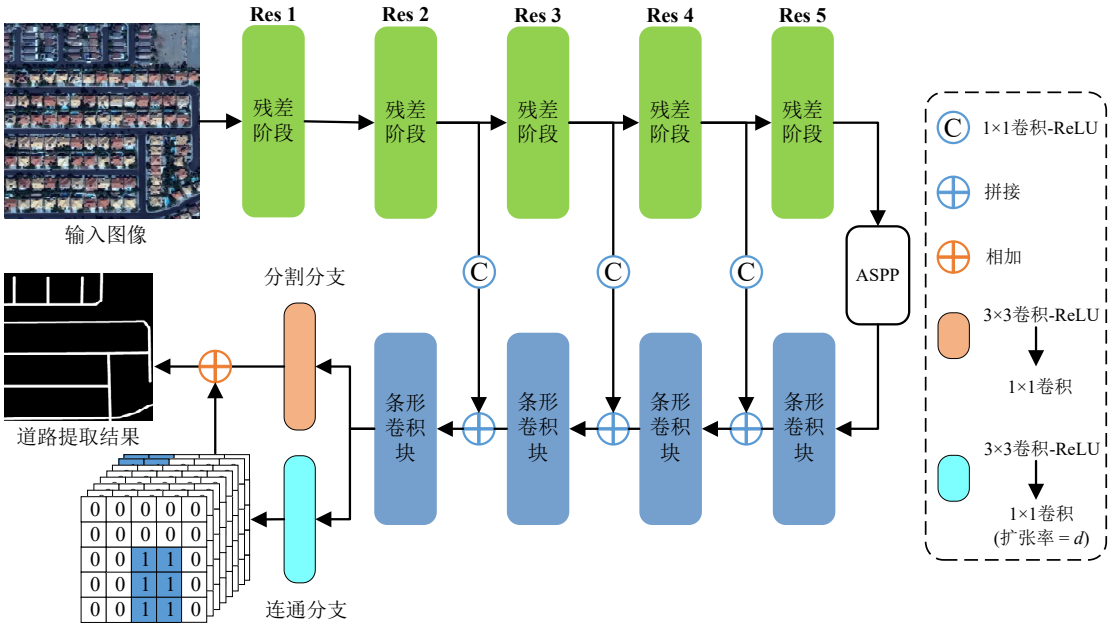


图 4.2 本章所提出的拓扑连通注意力网络（Connectivity Attention Network，CoANet）的整体结构图。其中编码器网络包含 5 个残差阶段，解码器网络包含 4 个条形卷积块，d 表示给定像素与其相邻像素的间隔。ASPP 是空洞空间金字塔池化模块（Atrous Spatial Pyramid Pooling Module），其通过应用多尺度的空洞卷积来学习多尺度特征。

的强大工具，参考文献[75]，本章将扩张率为 $r = 2$ 和 $r = 4$ 的空洞卷积应用于 ResNet-101 中的最后两个残差阶段，其能够提取更密集的特征。扩张率为 r 的空洞卷积在两个连续的卷积滤波器值之间插入 $r - 1$ 个零，而 $r = 1$ 的标准卷积是一种特殊的情况。为了在多个尺度上有效地学习图像特征，本章采用了空洞空间金字塔池化模块（Atrous Spatial Pyramid Pooling Module，ASPP）[191]。ASPP 引入了几个具有不同扩张率的并行空洞卷积滤波器来捕获多尺度特征，并最终与来自全局平均池化的特征相融合。由于现实中的道路呈现狭窄、复杂且跨度大的形状，ASPP 的使用将增加特征点的感受野并提高道路的连通性。此外，解码器模块包含四个条形卷积块，用于将特征图上采样到合适的大小并提取道路的线性特征。每个条形卷积块包含四个不同方向的条形卷积来捕获长距离上下文信息，包括水平、垂直、左对角线和右对角线等四个方向。每个条形卷积块的输出特征图都会经过一个 1×1 的卷积层进行调整，然后与编码器中相应残差阶段的输出特征图相拼接。

在本章所提出的拓扑连通注意力网络中，在解码器模块之后有两个分支：分别是分割分支和连通分支。连通分支对应于本章所设计的连通性注意力模块，

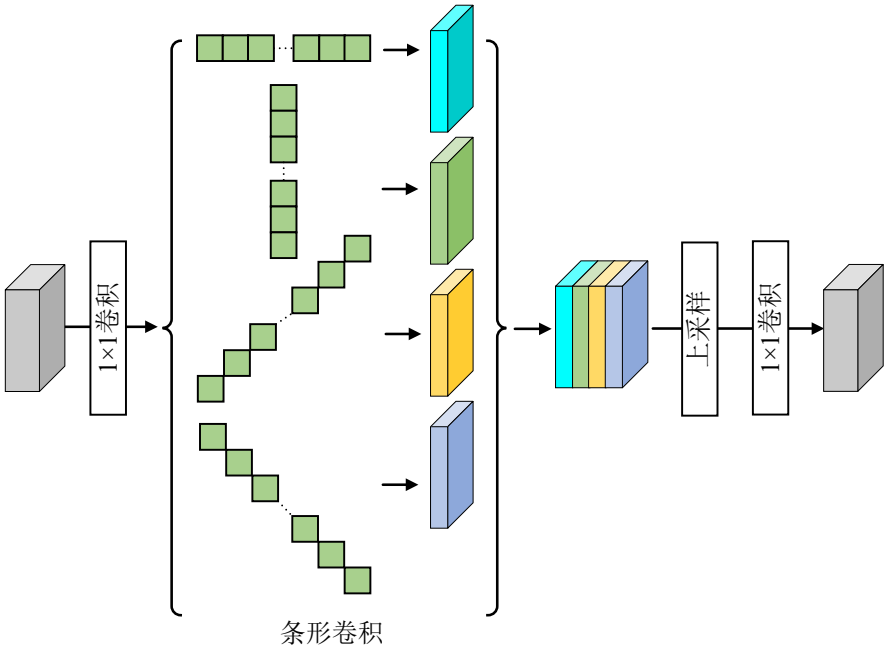


图 4.3 条形卷积模块（Strip Convolution Module，SCM）的结构示意图。其中条形卷积包含四个不同的形状：水平、垂直、左对角线和右对角线。

其通过预测给定像素与八个相邻像素的连通性来结合图形信息并保证道路的拓扑正确性。对于分割分支，其包含一个 3×3 的卷积层和一个用于将通道数量减少到 1 的 1×1 卷积层。分割分支的损失函数被定义为：

$$L_{seg} = L_{BCE} + \alpha(1 - L_{Dice}), \tag{4.1}$$

其中，L_{BCE} 是二元交叉熵，L_{Dice} 是 Dice 系数，其定义为：

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)], \tag{4.2}$$

$$L_{Dice} = \frac{2\sum_{i=1}^{N}(y_i\hat{y}_i)}{\sum_{i=1}^{N}y_i^2 + \sum_{i=1}^{N}\hat{y}_i{}^2}, \tag{4.3}$$

其中，α 是一个常数。N 表示 $H \times W$ 图像中元素的数量，y_i 表示位置 i 处给定像素的道路或背景的真值，$\hat{y}_i$ 是分割分支对应的预测概率。

4.2.2 条形卷积模块

卷积神经网络架构中的卷积通常具有方形内核并在方形窗口内学习特征图，这适用于大多数具有块状形状的自然物体。然而，现实中的道路往往呈现狭窄、

跨度大且连续分布的形状。利用方形卷积不能很好地捕捉道路的线性特征，并且不可避免地会包含来自相邻像素的干扰信息。条形卷积更符合道路的形状，其利用沿一个空间方向的条形内核来捕获道路区域中的长距离依赖信息，同时沿另一个空间方向捕获局部上下文信息并防止不相关区域对特征学习的干扰。

受上述事实的启发，本章提出了一种新颖的条形卷积模块（Strip Convolution Module，SCM）。如图 4.3所示，条形卷积模块利用水平、垂直、左对角线和右对角线等四个条形卷积从四个不同的方向捕获远距离上下文信息。设置 $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ 为条形卷积模块的输入特征图，其中 H、W 和 C 表示特征图的高、宽和通道数量。在条形卷积模块中，$\mathbf{X}$ 在经过一个 1×1 的卷积层后被输入到四个平行的路径，每个路径都包含一个不同方向的条形卷积。然后将四个条形卷积的输出特征图拼接起来，再经过一个上采样操作和一个 1×1 的卷积层获得条形卷积块的输出。

令 $\mathbf{w} \in \mathbb{R}^{2k+1}$ 是大小为 $2k+1$ 的条形卷积滤波器，$\boldsymbol{D} = (D_h, D_w)$ 是滤波器 $\mathbf{w}$ 的方向，$\mathbf{Z}_{\boldsymbol{D}} \in \mathbb{R}^{H \times W \times C'}$ 表示条形卷积的输出特征。则条形卷积的定义为：

$$\mathbf{Z}_{\boldsymbol{D}}[i,j] = (\mathbf{X} * \mathbf{w})_{\boldsymbol{D}}[i,j] \quad = \sum_{l=-k}^{k} x[i + D_h l, j + D_w l] \cdot w[k-l], \tag{4.4}$$

其中，$\mathbf{X} * \mathbf{w}$ 表示卷积操作。$\boldsymbol{D}$ 是条形卷积的方向向量，取值 $(0,1)$、$(1,0)$、$(1,1)$ 和 $(-1,1)$ 分别表示水平、垂直、左对角线和右对角线方向的条形卷积。对于滤波器 $\mathbf{w}$，本章设置 $k = 4$ 使得每个方向的条形卷积有 9 个参数，与 3×3 卷积核相同。

在上述的条形卷积模块中，输出特征图中每个位置的像素都能够与输入特征图中四个方向上多个位置的像素建立联系。本章选择的四个不同的条形卷积方向，与遥感图像中大部分道路的分布是一致的，同时也相对容易实现。

4.2.3 连通性注意力模块

由于建筑物和树木等造成的遮挡会对道路的连通性造成干扰，因此从遥感影像中提取道路非常具有挑战性。为了缓解这个问题，本章提出了一种连通性注意力模块（Connectivity Attention，CoA）来有效地预测相邻像素之间的连通关系。连通性注意力模块能够探索像素对之间的关系，且能够与特征学习过程无缝地结合。这一模块使得本章提出的模型能够整合通常在图形模型中学到的信息，并带来更好的道路连通性。

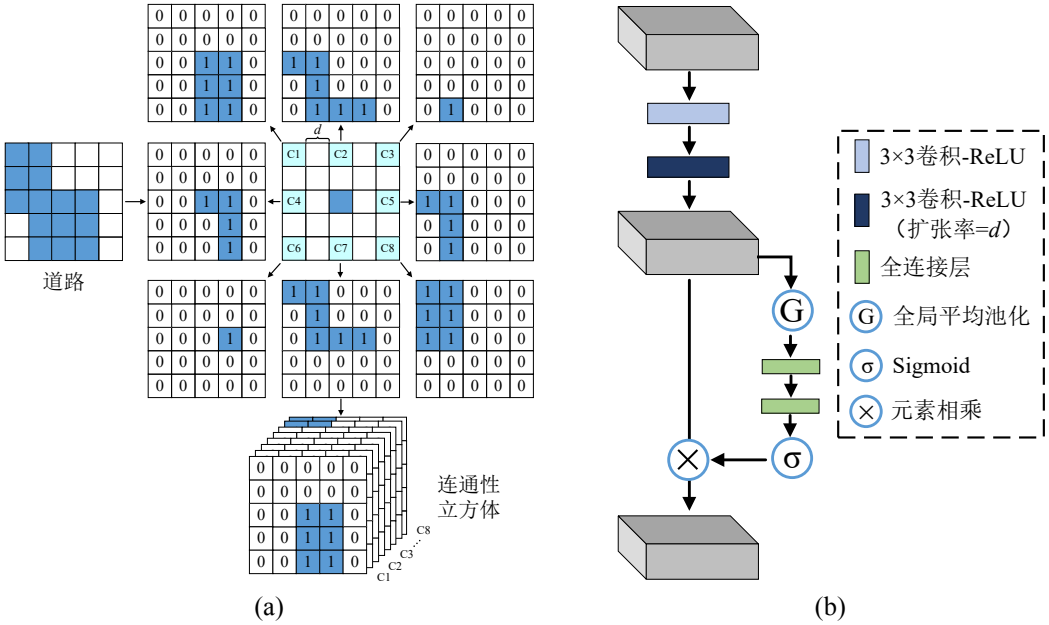


图 4.4 （a）连通性立方体生成示意图，采样像素之间的间隔 $d=1$。在道路图像中，白色像素表示背景，蓝色像素表示道路。在连通性立方体中，所有像素都有二元值，其中 1 表示该像素在一个方向上与相邻像素相连，0 表示未连接的像素。（b）连通性注意力模块（Connectivity Attention，CoA）结构示意图。

利用二值化真值，首先生成连通性立方体 $O \in \mathbb{R}^{H\times W\times C_o}$，其中 C_o 表示给定像素的相邻采样像素的数量，本章设置 $C_o=8$。在连通性立方体中，$O_{i,j,c}$ 表示一个像素与特定位置相邻像素的连通性，其中 i,j 表示该像素的位置，c 表示其相邻像素的位置。如果两个像素相连，则 $O_{i,j,c}=1$，表示这两个像素都是道路像素，背景像素不会相互连接以减少不相关的噪声。对于相邻的像素，本章对给定像素选择周围间隔为 $d=1$ 的像素。如图 4.4（a）所示，$C1-C8$ 位置上的像素被选为相邻像素。通过检查每个像素与其在特定位置的相邻像素是否连接并拼接位置 $C1-C8$ 的连接图，可以获得连通性立方体 O 的真值。

在连通性注意力模块中，如图图 4.4（b）所示，特征图首先被输入一个 3×3 的卷积层，其次是一个扩张率 $r=d$ 的 3×3 空洞卷积层，其中空洞卷积层的扩张率与相邻像素的间隔相同以增加感受野并学习相邻像素之间的关系。然后采用挤压-激励（Squeeze-Excitation，SE）模块[44]，通过使用通道注意力机制来重新校准预测的连通性立方体。输入特征图在全局平均池化后再依次经过两个全连接层和一个 Sigmoid 函数，即可以获得一个范围为（0，1）之间的向量，其中每个因子乘以输入特征图中的相应通道。连通性注意力模块的最终输出是一个

$H \times W \times C_o$ 的连通性立方体，用于预测相邻像素之间的连通性。

本章所提出的拓扑连通注意力网络中的连接分支包含两个连通性注意力模块，其中一个是上述 $d=1$ 的模块，另一个是 $d=3$ 的模块。对于 $d=3$ 的连通性注意力模块，给定像素与其相邻像素的间隔被设置为 3，其模块中的 3×3 空洞卷积扩张率相应的也被设置为 $r=3$。采用不同设置的两个连通性注意力模块来捕获多尺度连通性信息，可以提高预测道路的连通性。本章将在实验中对连通性注意力模块不同配置下的性能提供更多的分析。

连通分支的损失函数被定义为：

$$L_{con} = L_{d1} + \beta L_{d3}, \tag{4.5}$$

$$L_{d1} = -\frac{1}{C_o \times N} \sum_{c=1}^{C_o} \sum_{i=1}^{N} [y_i^c \cdot \log(\hat{y_i^c}) + (1 - y_i^c) \cdot \log(1 - \hat{y_i^c})], \tag{4.6}$$

其中，β 是一个常数。C_o 表示给定像素的相邻采样像素的数量，N 是 $H \times W$ 图片中元素的数量。y_i^c 表示位置 i 处的给定像素与其在位置 c 处的相邻像素的连通性或非连通性的真值，$\hat{y_i^c}$ 是连通分支所对应的预测连通性。损失函数 L_{d3} 与 L_{d1} 相同。

本章方法的整体损失函数可以被定义为：

$$L_{CoANet} = L_{seg} + \lambda L_{con}, \tag{4.7}$$

其中 λ 是一个常数。

第三节 实验结果与分析

本小节将介绍本章实验中所使用的两个公开数据集和评估指标，在此基础上与多个方法进行了定量和定性的比较实验。此外，本小节也详细地对模型中的模块进行了消融实验，来验证各个模块的有效性和重要性。

4.3.1 数据集

为了验证本章所提出的方法在道路提取任务上的性能，本小节在两个不同的公开道路数据集上开展了实验：分别是 SpaceNet[229]和 DeepGlobe[230]。在两个数据集中，本小节均使用了三通道的 RGB 图像。

SpaceNet[229]: 该数据集提供了来自巴黎、拉斯维加斯、上海和喀土穆等四个不同城市的遥感图像。其中的图像是从 DigitalGlobe WorldView-3 卫星收集的，每张图像的像素分辨率为 1300 × 1300，空间分辨率为 30 厘米/像素。SpaceNet 提供了线串形式（Line-String）的道路标注，用来表示道路的中心线。为了方便与其他数据集进行比较，本节将线串标注以固定的宽度进行栅格化，从而获得道路分割的真值。在本节的实验中，道路中心线的缓冲区被设置为 3 米（10 个像素）。此数据集包含 2780 张图像，参考文献[231]，本小节将其分为 2213 张训练图像和 567 张测试图像。为了增强训练图像，本小节将原图像裁剪为 650× 650 的图像，同时采用全尺寸图像来测试。

DeepGlobe[230]: 该数据集包含来自三个不同地区的图像：分别是泰国、印度尼西亚和印度，其中的图像是从 DigitalGlobe +Vivid 数据集中收集的。每张图像的像素分辨率为 1024 × 1024，空间分辨率为 50 厘米/像素。DeepGlobe 提供了像素级标注，指示每个像素属于道路或者背景类别。此数据集包含 6226 张图像，参考文献[231]，将其分成 4696 张训练图像和 1530 张测试图像。此外，通过创建 512 × 512 的裁剪图像来扩充训练数据集。

4.3.2 评测指标

4.3.2.1 基于像素的评测指标

从遥感图像中提取道路可以被看做是图像分割问题，即识别图像中的每个像素属于道路或非道路类别。本小节采用像素级的交并比（Intersection over Union，IoU）和 $F1$ 分数来评估道路分割的性能，其定义如下：

$$IoU = \frac{TP}{TP + FP + FN}, \tag{4.8}$$

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall}, \tag{4.9}$$

其中，$Precision = TP/(TP + FP)$, $Recall = TP/(TP + FN)$。TP、FP 和 FN 分别代表真阳性（True Positive）、假阳性（False Positive）和假阴性（False Negative）的像素数量。

4.3.2.2 基于图形的评测指标

上述两个基于像素的评测指标严重惩罚了道路宽度的偏差，但轻微惩罚了预测道路的碎片化，因此这两个指标无法正确激励连通道路的生成。本小节实

验中采用了 SpaceNet 数据集[229]中的平均路径长度相似度（Average Path Length Similarity，APLS）来衡量预测道路的拓扑正确性和连通性。评测指标 APLS 衡量了预测道路图 $\hat{G}$ 中所有节点对之间最短路径距离和真值道路图 G 的偏差。任何带有缺失边的预测道路图 $\hat{G}$（例如，如果道路被建筑物或树木遮挡，则可能会造成所提取道路的缺失。）将受到该评测指标的惩罚，以确保正确连接的道路可以获得高分。评测指标 APLS 的定义为：

$$APLS = 1 - \frac{1}{n}\sum min\left\{1, \frac{\left|L(a,b) - L(\hat{a},\hat{b})\right|}{L(a,b)}\right\}, \tag{4.10}$$

其中，$\hat{a},\hat{b}$ 分别为预测图 $\hat{G}$ 中最接近真值图 G 中节点 a,b 位置的节点。$L(\hat{a},\hat{b})$ 和 $L(a,b)$ 分别表示预测图 $\hat{G}$ 和真值图 G 中对应节点之间的路径长度。n 是唯一路径的数量。

4.3.3 实现细节

在本章所提出的道路提取模型拓扑连通注意力网络 CoANet 中，使用了批大小（Batch Size）为 16 的随机梯度下降（Stochastic Gradient Descent，SGD）优化器，动量（Momentum）和权重衰减系数（Weight Decay Coefficients）分别被设置为 0.9 和 5×10^{-4}。初始学习率被设置为 0.01，采用“poly”策略逐渐降低学习率，其中学习率被乘以 $(1 - \frac{iter}{maxiter})^{power}$，$power = 3$。本章的模型是使用深度学习框架 PyTorch[223]来实现的，实验是在 4 个显存为 24GB 的 NVIDIA RTX TITAN GPU 上实现的。在训练过程中，应用随机旋转、水平翻转、尺度调节和高斯模糊等数据增强手段来提高模型的泛化能力，并将两个数据集的图像裁剪为固定大小 512×512 作为模型的输入。

在推理阶段，本小节利用连通性分支的预测输出来增强道路提取的结果。连通性注意力模块的输出中有八个通道，其中每个通道的预测可以被看作是分割任务的一个子问题。例如，给定预测的连通性立方体 O，如果 $\sigma(O_{i,j,c}) > t$，则 (i,j) 位置处的像素与位置 c 处的相邻像素是相连的，并且它们两个都是道路像素。$\sigma()$ 是 Sigmoid 非线性函数，t 是一个阈值。通过估计 O 的每个通道，并沿通道维度求和，就可以得到一个一维的道路掩膜。将其加入到分割分支的输出中，得到最终的道路提取结果。

4.3.4 与现有方法的对比

本小节将本章的拓扑连通注意力网络在 SpaceNet[229]和 DeepGlobe[230]数据集上与几种目前较好的道路提取方法进行了比较：包括 DeepRoadMapper[88]、Topology Loss[89]、LinkNet34[85]、D-LinkNet[86]、RoadCNN[58]、ImprovedConnectivity[59]和 VecRoad[232]。DeepRoadMapper[88]采用卷积神经网络从遥感图像中获得道路的初始分割结果，并利用后处理步骤推断出丢失的路段。Topology Loss[89]提出了拓扑损失和迭代优化的方法来提高预测道路的性能。LinkNet34[85] 将空间信息从编码器传递到相应的解码器层，D-LinkNet[86] 建立在 LinkNet 架构之上，并在其中心部分增加了空洞卷积层。RoadCNN[58] 使用卷积神经网络获得分割输出，并应用一组启发式方法将分割结果转换成道路网络图。ImprovedConnectivity[59] 提出了一个堆叠的多分支卷积模块来同时利用道路分割和方向学习任务的信息，其进一步开发了连通性细化方法来增强预测道路网络的连通性。VecRoad[232] 设计了一种基于点的迭代探索方法以增强道路的连通性。本节实验在相同的数据集上重新执行了上述方法以进行公平的比较。

4.3.4.1 SpaceNet 数据集上的比较

SpaceNet 数据集[229]上的定量实验结果如 表 4.1所示。值得注意的是，本章所提出的拓扑连通注意力网络（CoANet）在基于像素和基于图形的评测指标上都优于其他方法。例如，CoANet 取得了 76.91% 的 F1 得分和 62.48% 的 IoU 得分，分别比 ImprovedConnectivity[59]高了 1.00% 和 1.31%。对于用来评估道路拓扑正确性的评测指标 APLS，CoANet 将次优方法 LinkNet34[85]提高了 2.4%。由于本章提出的 CoANet 将像素分割和相邻像素之间的关系集成到一个框架中，因此其可以从遥感图像中提取更准确的道路，同时连通性注意力模块的使用可以进一步提高道路的拓扑连通性。本小节还比较了一种基于图的方法 VecRoad[232]，其引入了具有灵活步骤的迭代图探索模型。相比于 VecRoad[232]，CoANet 在 IoU 分数上获得了 15.83% 的改进，在 APLS 分数上获得了 3.89% 的提高。基于图的方法通常能更好地保证所提取道路的连通性，但可能存在大量缺失的道路，因此本章的 CoANet 也比基于图的方法更具有优势。此外，本小节通过在推理过程中使用连通分支的真值来得到所提出方法 CoANet 的上限。上限结果表明连通分支仍有较大的改进空间，一种可能的改进是在未来的工作中使用更大的像素间隔。

表 4.1 本章提出的拓扑连通注意力网络 CoANet 在 SpaceNet 数据集上与其他道路提取方法的定量比较（%）。CoANet-UB 表示 CoANet 在连通性分支是真值情况下的上限。

方法	F1	IoU	APLS
DeepRoadMapper[88] $_{ICCV17}$	71.47	55.61	46.76
Topology Loss[89] $_{CVPR18}$	58.44	41.29	39.08
LinkNet34[85] $_{VCIP17}$	73.96	58.68	63.12
D-LinkNet[86] $_{CVPRW18}$	69.77	53.57	50.20
RoadCNN[58] $_{CVPR18}$	73.74	58.40	59.39
ImprovedConnectivity[59] $_{CVPR19}$	75.91	61.17	62.81
VecRoad[232] $_{CVPR20}$	63.63	46.65	61.64
CoANet	**76.91**	**62.48**	**65.53**
CoANet-UB	**85.54**	**74.73**	**76.98**

SpaceNet[229]数据集中的遥感影像主要采集于城市区域，包含了城市中的多种道路类型，如高速公路、住宅区道路和主干道等。在这个数据集中的图像中，会存在建筑物、建筑物阴影以及树木等的遮挡。本章方法在基于像素和基于图形的评测指标上都取得了最好的结果，这表明 CoANet 能够处理复杂的城市交通环境并提取具有更好连通性的道路。

4.3.4.2 DeepGlobe 数据集上的比较

表 4.2展示了本章所提出的 CoANet 与其他目前较好的道路提取方法在 DeepGlobe[230]数据集上的定量比较结果。由于 VecRoad[232]需要道路线串的真值，而 DeepGlobe 数据集只有像素级的标注，所以这里不展示 VecRoad 的实验结果。值得注意的是，本章的 CoANet 实现了 68.37% 的 IoU 分数，优于其他所有方法，并将第二好的方法 ImprovedConnectivity[59]提高了 1.79%。此外，CoANet 还获得了最佳的 APLS 分数，其性能优于 LinkNet34[85]0.55%，高于 D-LinkNet[86]1.67%。

DeepGlobe[230]数据集中的遥感图像主要来自乡村地区，其中包含大量的乡村道路，并且道路的宽度在不断变化，这意味着一条道路可能有不同的宽度，并同时存在树木和树木阴影造成的严重遮挡。实验结果表明，本章的 CoANet 对乡村地区的道路也有效。结合在 SpaceNet[229]数据集上的结果，表明本章提出的方法对不同地区的不同道路类型都具有鲁棒性。

表 4.2 本章提出的拓扑连通注意力网络 CoANet 在 DeepGlobe 数据集上与其他道路提取方法的定量比较（%）。CoANet-UB 表示 CoANet 在连通性分支是真值情况下的上限。

方法	F1	IoU	APLS
DeepRoadMapper[88] $_{ICCV17}$	78.04	63.98	58.85
Topology Loss[89] $_{CVPR18}$	56.07	38.95	46.99
LinkNet34[85] $_{VCIP17}$	79.65	66.18	72.93
D-LinkNet[86] $_{CVPRW18}$	77.49	63.26	71.81
RoadCNN[58] $_{CVPR18}$	79.08	65.40	71.15
ImprovedConnectivity[59] $_{CVPR19}$	79.93	66.58	71.69
CoANet	**81.22**	**68.37**	**73.48**
CoANet-UB	**89.25**	**80.58**	**85.14**

表 4.3 本章提出的条形卷积模块（Strip Convolution Module，SCM）和连通性注意力模块（Connectivity Attention Module，CoA）的消融实验（%）。基准模型是基于 ResNet-101（编号 1）的分割模型。通过添加条形卷积模块和连通性注意力模块来验证这两个模块的有效性（编号 2 和编号 3）。编号 4 是本章提出的 CoANet 的完整版本。

			SpaceNet		DeepGlobe	
编号	SCM	CoA	IoU	APLS	IoU	APLS
1			59.57	58.69	63.09	69.13
2	✔		61.84	63.93	63.89	70.09
3		✔	61.10	61.27	64.32	70.45
4	✔	✔	**62.48**	**65.53**	**68.37**	**73.48**

4.3.4.3 可视化结果比较

本章模型 CoANet 和其他方法的定性比较结果如图 4.5所示，其中展示了来自 SpaceNet[229]数据集中不同城市的四个示例和 DeepGlobe[230]数据集中的四个示例。值得注意的是，本章方法提取的道路基本与真值道路一致，并且存在很少的假阳性像素。在一些存在遮挡的区域，其他方法提取的道路可能会存在断开的情况，而 CoANet 很好地保持了道路的连通性。例如，在拉斯维加斯的结果（图 4.5中的第一行）中，图像的左下角有一个停车场，其中停放着很多车辆并存在几棵树。利用 DeepRoadMapper[88]、Topology Loss[89]、LinkNet34[85]和 ImprovedConnectivity[59]等方法提取的道路有很多缺失，未能保持道路连通性，但是通过 CoANet 得到的结果基本与真值一致。在 DeepGlobe（图 4.5中第八行）数据集的结果中，道路来自乡村地区且存在树木造成的严重遮挡。与本章的方法相比，其他方法提取的道路连通性较差。这些可视化结果验证了本章的 CoANet

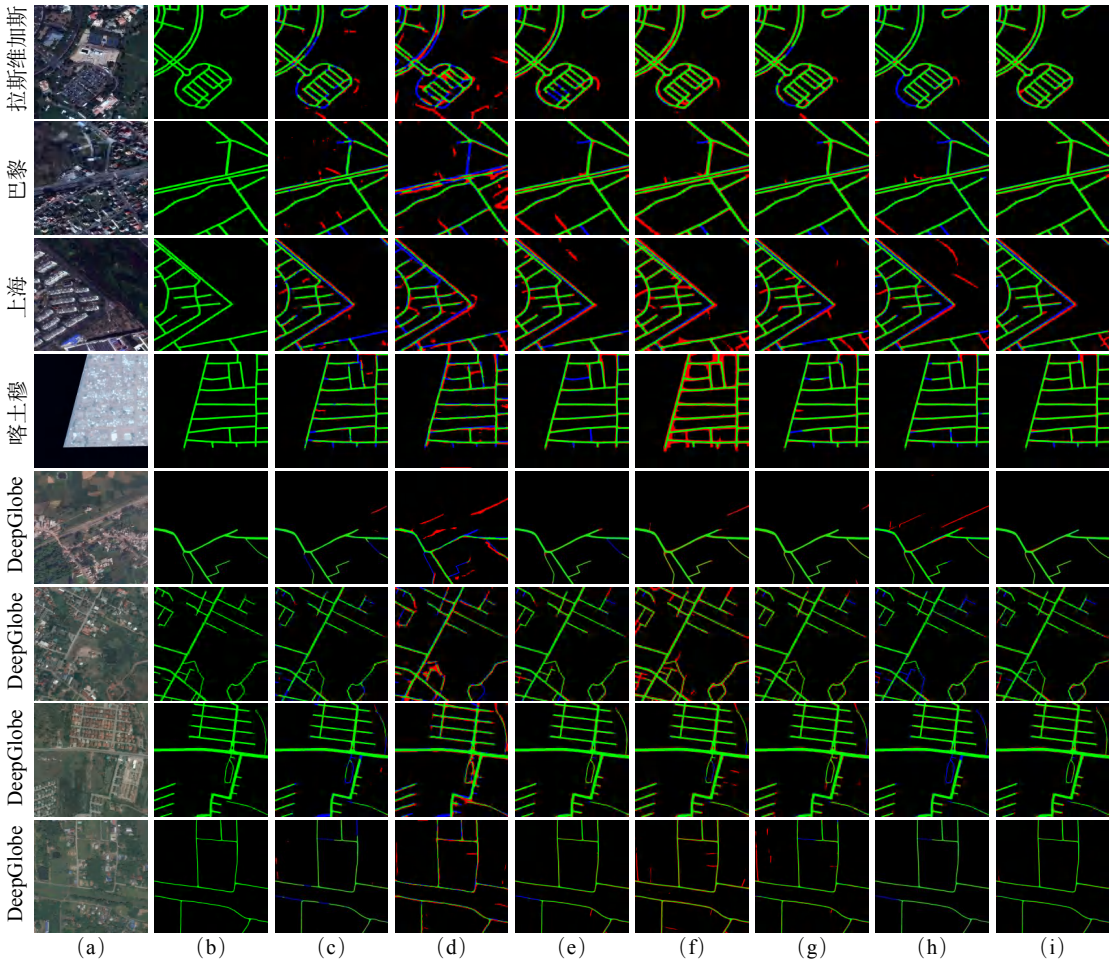


图 4.5 本章方法 CoANet 与其他道路提取方法的定性比较。其中绿色为真阳性，红色为假阳性，蓝色为假阴性。第一行到第四行分别是拉斯维加斯、巴黎、上海、喀土穆等 SpaceNet[229]数据集中不同城市的对比结果。第五行到第八行是 DeepGlobe[230]数据集的对比结果。(a) 遥感图像；(b) 真值图；(c) - (i) DeepRoadMapper[88]、Topology Loss[89]、LinkNet34[85]、D-LinkNet[86]、RoadCNN[58]、ImprovedConnectivity[59]和本章方法 CoANet 的道路提取结果。

在基于遥感影像提取道路任务中的优越性。

4.3.5 消融实验

4.3.5.1 模块的有效性

本章的拓扑连通注意力网络中提出了条形卷积模块（Strip Convolution Module，SCM）和连通性注意力模块（Connectivity Attention Module，CoA）来分别获取道路区域的长距离信息并探索相邻像素之间的依赖关系。为了验证这两个模块的有效性，本小节进行了模型在不同配置下的实验，如表 4.3所示。基

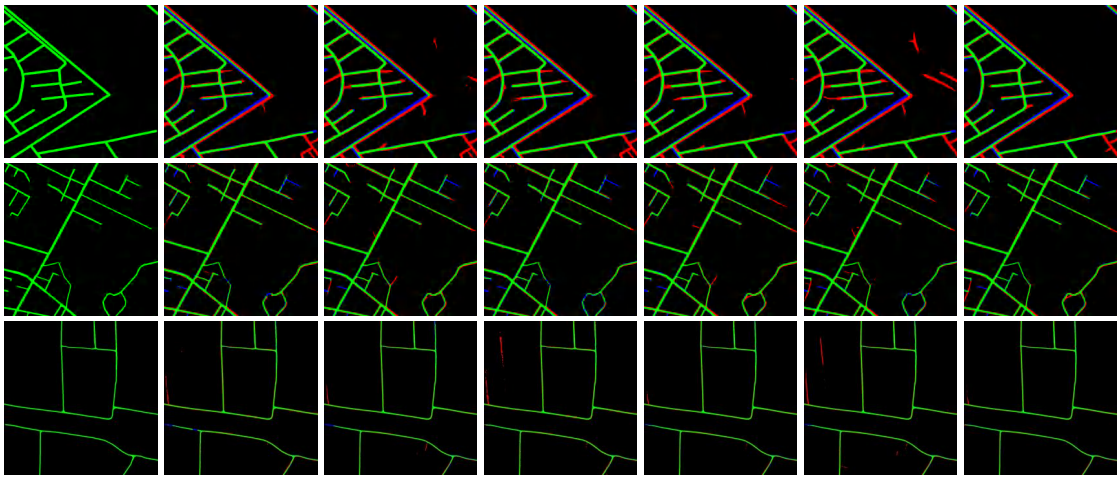


图 4.6 本章方法 CoANet 在不同模型配置下的可视化结果。绿色为真阳性，红色为假阳性，蓝色为假阴性。第一行到第三行显示了来自 SpaceNet[229]和 DeepGlobe[230]数据集的三个样本。(a) 真值图；(b) 基准模型；(c) 基准模型 + 条形卷积模块；(d) 基准模型 + 连通性注意力模块；(e) 只包含配置 $d = 1$ 的连通性注意力模块的 CoANet；(f) 包含三个连通性注意力模块的 CoANet；(g) 本章方法 CoANet。

表 4.4 本章提出的连通性注意力模块 CoA 的消融实验（%）。CoANet-Affinity 表示将 CoANet 中的连通性注意力模块替换为文献[233]中的亲和力模块（Affinity Module）。

	SpaceNet		DeepGlobe	
	IoU	APLS	IoU	APLS
CoANet-Affinity	61.93	62.67	62.43	69.54
CoANet	**62.48**	**65.53**	**68.37**	**73.48**

准模型（编号 1）是基于 ResNet-101 的全卷积网络模型，将本章提出的 SCM 和 CoA 分别加入到基准模型后，在 SpaceNet[229]数据集上，IoU 得分分别从 59.57% 提高到 61.84% 和 61.10%。对于 SpaceNet 数据集上的评测指标 APLS 分数，添加 SCM 和 CoA 后将基准模型分别提高了 5.24% 和 2.58%。在结合 SCM 和 CoA 之后，本章方法在 IoU 分数上实现了 2.91% 的提高，在 APLS 分数上实现了 6.84% 的提高。通过添加 SCM 和 CoA 这两个模块，本章方法在 DeepGlobe 数据集[230]上的实验性能也得到了提升。这些结果验证了本章提出的条形卷积模块和连通性注意力模块对于道路提取是有效的。

本章方法在不同模型配置下的定性结果如图 4.6所示。可以看到，基准模型的可视化结果中存在一些断掉的道路，尤其是被树木遮挡的区域。添加 CoA 模块后，基准模型中的大部分断裂路段被连接，但是道路的边缘是粗糙的，并且

表 4.5 不同配置下连通性注意力模块（Connectivity Attention Module，CoA）的消融实验。$d1$、$d3$ 和 $d5$ 分别表示 $d=1$、$d=3$ 和 $d=5$ 配置下的连通性注意力模块。编号 1 是只包含一个连通性注意力模块（$d=1$）的 CoANet；编号 2 表示本章方法 CoANet 的完整版本；编号 3 表示由三个连通性注意力模块组成的连通分支。

				SpaceNet		DeepGlobe	
编号	$d1$	$d3$	$d5$	IoU	APLS	IoU	APLS
1	✔			62.04	64.87	64.39	70.50
2	✔	✔		**62.48**	**65.53**	**68.37**	**73.48**
3	✔	✔	✔	62.09	65.13	64.89	70.83

表 4.6 不同配置下条形卷积模块（Strip Convolution Module，SCM）的消融实验（%）。'H'、'V'、'L' 和 'R' 分别表示具有不同形状的四个条形卷积：水平、垂直、左对角线和右对角线。编号 1 是包含两个条形卷积的条形卷积模块：水平和垂直；编号 4 代表本章方法的完整版本；编号 5 表示条形卷积模块中每个形状包含两个卷积，总共有 8 个卷积。

		SpaceNet		DeepGlobe	
编号	SCM	IoU	APLS	IoU	APLS
1	H&V	61.85	64.58	64.57	70.65
2	H&V&L	62.01	65.01	64.69	70.92
3	H&V&R	61.93	64.85	64.97	70.74
4	H&V&L&R	**62.48**	**65.53**	**68.37**	**73.48**
5	2×H&V&L&R	61.98	65.41	65.19	71.03

有一些离散像素点被识别为道路。SCM 模块有助于改善道路的连通性，同时使提取的道路更加平滑。然而，由于 SCM 是为了获取远距离依赖关系而设计的，因此其他类别的某些区域可能会被识别为道路，这会使得提取的道路比实际道路更长。需要注意的是，如果单独使用本章提出的两个模块，提取的道路都会有一些缺陷。SCM 模块能够连接 CoA 模块中相邻像素无法到达的断裂路段，而 CoA 模块可以防止来自背景区域的无关噪声。因此，本章的 CoANet 通过结合这两个模块获得了最好的道路提取结果。

为了验证本章提出的连通性注意力模块在道路提取任务上的有效性，本小节将其与学习像素亲和力的模块进行了比较。AffinityNet[233]被提出来学习图像中相邻像素之间的类别无关语义亲和性，其类似于本章所提出的 CoA 模块。本小节将 CoA 模块替换为 AffinityNet[233]中的亲和力模块，实验结果如 表 4.4所示。可以看到对于 DeepGlobe[230]数据集，CoANet 在 IoU 分数上比 CoANet-Affinity 高 5.94%，在 APLS 分数上高 3.94%。在 AffinityNet[233]中，如果两个相邻像素

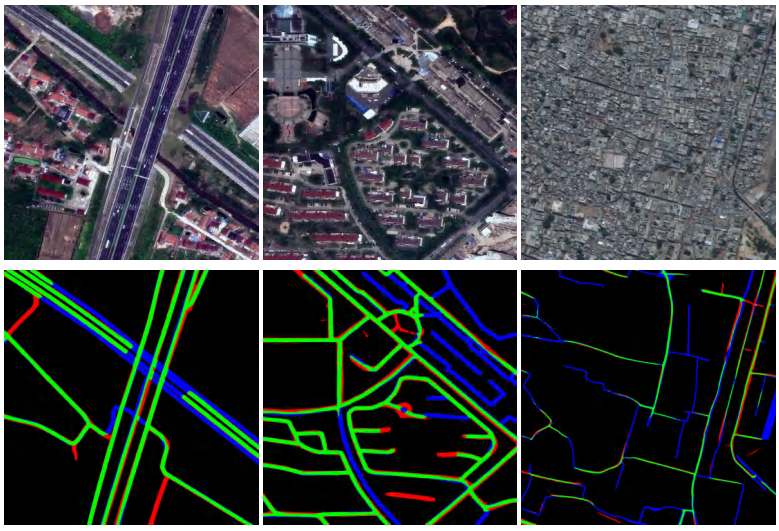
的类别是相同的，则将它们的亲和力标签分配为 1，而且像素对是在小半径区域内采样的。然而，本章的 CoA 模块探索了相邻道路像素之间的连通性并抑制了背景区域的干扰，并且相邻像素是从周围的八个特定方向上采样的，这些方向与遥感图像中大多数道路的分布是一致的。此外，CoA 模块中采样像素的间隔有两种不同的设置，其可以捕获多尺度的连通性信息，而 AffinityNet[233]中的采样半径是固定的。这些优势使得本章的 CoA 模块在道路提取任务上取得了更好的性能。

4.3.5.2 连通性注意力模块中不同配置的影响

如章 4.2.3中所描述的，CoANet 中的连通分支包含两个 CoA 模块，其中一个是给定像素与邻域像素的间隔 $d=1$，另一个是 $d=3$。本小节分析了连通分支中 CoA 模块不同配置的影响，结果如表 4.5中所示。定义 CoANet-*d1* 表示本章提出的 CoANet 只包含一个 $d=1$ 的 CoA 模块，CoANet-*d5* 表示 CoANet 由 $d=1$、$d=3$ 和 $d=5$ 等三个 CoA 模块组成。CoANet-*d1* 在 SpaceNet 数据集上[229]上获得了 62.04% 的 IoU 分数和 64.87% 的 APLS 分数。添加了 $d=3$ 的 CoA 模块之后，性能在 IoU 分数上提高到 62.48%，在 APLS 分数上提高到 65.53%。然而，如果 CoANet 由三个 CoA 模块组成（表 4.5 中的编号 3），其性能高于 CoANet-*d1*（编号 1），但是低于有两个 CoA 模块的 CoANet （编号 2）。从可视化结果中可以分析其原因，如 图 4.6所示，CoANet-*d1* 的结果中还存在一些断裂的道路。此外，由于给定像素与其相邻像素的间隔更大，CoANet-*d5* 的道路提取结果中包含更多的背景区域并且提取的道路比真值更长，对其结果造成了影响。因此本章选择具有两个 CoA 模块的配置，其获得的道路提取结果更符合真值。

4.3.5.3 条形卷积模块中不同配置的影响

如表 4.6所示，本小节分析了条形卷积模块中不同配置对道路提取结果的影响。当 SCM 包含水平和垂直方向的条形卷积时，其在 SpaceNet 数据集[229]上的 IoU 得分为 61.85%，在 APLS 分数上获得了 64.58%。通过添加额外一种类型的条形卷积，能够提高其性能，如表 表 4.6中的编号 2 和编号 3。在添加左对角线和右对角线方向的条形卷积后，CoANet 在 IoU 得分上达到了 62.48%，在 APLS 得分上达到了 65.53%。以上实验结果符合预期，使用不同方向的 4 个条形卷积，本章方法可以捕获多个空间方向上的长距离上下文信息。值得注意的是，这四个方向与遥感图像中大部分道路的分布是一致的，并且相对容易实现。本小节



（a） （b） （c）

图 4.7 本章方法 CoANet 的失败案例分析。绿色为真阳性，红色为假阳性，蓝色为假阴性。第一行是遥感图像，第二行是 CoANet 的道路提取结果。（a）-（c）是来自 SpaceNet[229]和 DeepGlobe[230] 数据集中的三个样本。

表 4.7 本章方法 CoANet 和其他道路提取方法在相同条件下的运行时间分析。表中列出了相同批量大小图像的训练时间和一张图像的推理时间（s）。

	训练时间	推理时间
DeepRoadMapper[88] $_{ICCV17}$	0.579	0.121
Topology Loss[89] $_{CVPR18}$	0.168	0.049
LinkNet34[85] $_{VCIP17}$	0.201	0.066
D-LinkNet[86] $_{CVPRW18}$	0.218	0.071
RoadCNN[58] $_{CVPR18}$	0.156	0.028
ImprovedConnectivity[59] $_{CVPR19}$	0.361	0.074
CoANet	0.146	0.022

还列出了 SCM 模块中每个形状包含两个卷积，总共有八个条形卷积的配置结果，如编号 5 所示，然而额外的条形卷积损害了模型性能的提高并增加了计算成本。因此，本章在 SCM 模块中应用了四个不同方向的条形卷积。

4.3.6 讨论

4.3.6.1 失败案例分析

如上述实验所验证的，本章提出的道路提取模型 CoANet 在两个公共数据集 SpaceNet[229]和 DeepGlobe[230]上都实现了最好的性能。但是，本章方法仍然

存在一些失败的案例，如 图 4.7所示，其中展示了来自两个数据集的三个不同的样本。对于图 4.7（a），遥感图像中有一个隧道和一个立交桥。此外，树木和建筑物会对道路区域造成严重的遮挡，如 图 4.7（b）和（c）所示。由于遮挡区域非常大，并且这些区域的道路在遥感图像中可能不可见，本章的 CoANet 无法在这些区域生成道路并保持道路的连通性。下一步的工作将考虑加入其他的信息来提取这些挑战性区域的道路，例如行人和汽车的 GPS 轨迹。此外，遥感图像中的道路网络可以被视为具有边和节点的图，可以利用图卷积网络来提取道路，这可能会对提取遮挡区域的道路有效。

4.3.6.2 运行时间分析

如表 4.7所示，本小节在 SpaceNet[229]数据集上分析了 CoANet 和其他道路提取方法的运行时间。所有方法的对比实验均是在具有 4 个 NVIDIA RTX TITAN GPU 的工作站上执行的。为了公平地进行比较，本小节列出了相同批量大小图像的训练时间和大小为 512×512 图像的推理时间。可以看出，本章的 CoANet 实现了最快的训练时间和推理时间。此外，利用后处理步骤的方法，如 DeepRoadMapper[88] 和 ImprovedConnectivity[59]，需要更多的时间来进行训练和推理。凭借更好的性能和更快的执行速度，本章提出的 CoANet 更适合从遥感图像中提取道路。

第四节 本章小结

本章提出了一种用于从遥感影像中提取道路的拓扑连通注意力网络（Connectivity Attention Network，CoANet），其共同学习了图像分割和像素的成对依赖关系。现实中的道路大多是大跨度、狭窄且连续分布的，受到道路形状的启发，本章提出了一种条形卷积模块（Strip Convolution Module，SCM）来提取道路的线性特征。条形卷积模块利用四个条形卷积从四个不同的方向捕获远距离上下文信息，并防止不相关区域对特征学习的干扰。此外，为了缓解由建筑物或树木引起的道路区域的遮挡问题，本章设计了连通性注意力模块（connectivity attention module，CoA）来探索相邻像素之间的关系，其能够对给定像素周围八个相邻像素的连通性进行预测，这其中包含了图形信息，使道路的连通性更好地得到保留。在两个公开数据集（SpaceNet 和 DeepGlobe 数据集）上的大量实验验证了本章提出的 CoANet 与其他道路提取方法相比的优越性。同时还进行

了不同模型配置下的消融实验以验证条形卷积模块和连通性注意力模块在道路提取任务中的有效性。

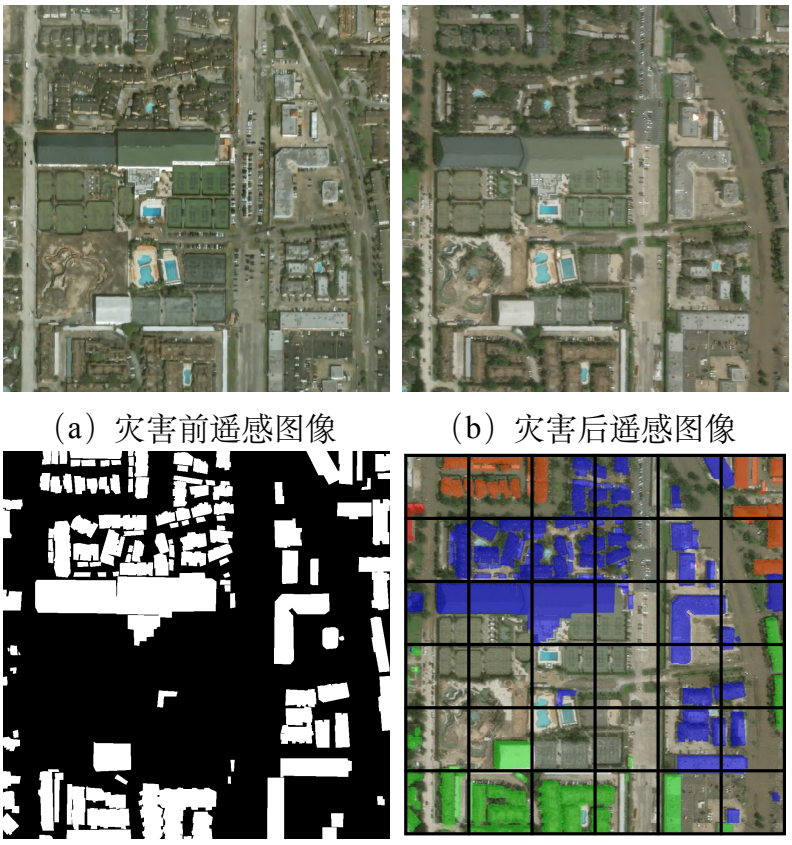
第五章　基于差异感知注意力的双时序图像变化检测

双时序遥感图像包含相同区域两个不同时间的图像，属于时间序列图像。本章基于双时序遥感图像，通过探索时序信息来检测不同图像间发生的多级别变化。本章的章节安排为：第一节介绍了本章的研究背景、研究内容和创新点；第二节介绍了所提出的基于差异感知注意力网络的变化检测模型；第三节给出了本章所提出的变化检测模型的对比实验及相关分析；第四节是对本章内容进行的总结。

第一节　引言

变化检测是一项用于识别不同时序图像间差异的技术，其在多个领域得到了广泛的研究：例如：环境监测[101]、异常检测[234]、灾害评估[100]、恶意软件检测[235] 以及医疗保健[236]等。变化检测其中一个重要的应用是分析地表的动态变化，遥感图像不仅提供了覆盖同一片区域的多时序图像并且相对容易获取，这为变化检测提供了必要的支持。当自然灾害发生时，尽快评估灾害损毁的严重性和损毁范围对于援助受灾人员和分配救灾物资是非常重要的。发生灾害的地区通常比较危险而且很难实地勘察，所以高分辨率遥感图像是非常有价值的评估灾害影响的工具。然而，现有的灾害评估方法通常需要人为分析双时序遥感图像，即灾害前和灾害后的遥感图像。这些方法需要大量人力物力的投入，并且无法快速应用于大范围的灾害区域。为了减轻人力劳动并加快灾害评估的过程，近年来越来越多的研究关注利用双时序遥感图像进行自动检测变化的方法。

传统的变化检测方法通常基于像素之间的差异来检测多个时间段内图像的变化[105,113,117,237]。但是，同一个区域内双时序遥感图像之间的配准误差和光照变化给变化检测算法带来了较大的挑战。图像中的这些误差通常是由不同的卫星成像参数造成的，难以避免，而基于像素差异的变化检测方法往往是针对特定数据设计的，不能有效地解决这些问题[238]。随着深度学习的发展，基于卷积神经网络（CNN）的模型在语义分割任务中取得了较好的性能[75,191,239]。一些研究采用具有编码器-解码器架构的卷积神经网络，基于分割的方法来完成变化检测的任务[125,240]。考虑到灾害损毁评估的重要性，近年来越来越多的研



（a）灾害前遥感图像　（b）灾害后遥感图像

（c）建筑物分割真值　（d）多级别变化真值

图 5.1　双时序遥感图像之间多级别变化的示意图。（d）中的四种颜色代表四种不同的变化等级：绿色、蓝色、橘红色和红色分别代表无损毁，轻微损毁，严重损毁和完全损毁。本章提出了一个差异感知注意力网络来激活双时序图像之间的全局变化模式，并获取多级别变化之间丰富的局部依赖性。如图中（d）所展示的，将特征图划分成一些小的特征立方体来学习局部信息。

究[132-134] 关注由自然灾害造成的变化。这些方法结合灾害前和灾害后的遥感图像来检测受损的建筑物，但是它们通常被设计成只能识别灾害造成的单一变化。最近，为了推动变化检测和建筑物损毁评估的研究，Gupta 等人[7]发布了一个包含 19 种自然灾害和 4 种不同受损程度的双时序遥感图像数据集 xBD。基于 xBD 数据集，Gupta 等人[7] 进一步提出了一个基于 U-Net[46]架构的网络来进行建筑物的分割，同时利用 ResNet-50 网络来进行灾害损毁的分类。Shen 等人[241] 提出了一个包含两个网络的模型来完成这两个任务。但是，这些方法通常将建筑物分割和变化检测分成两个单独的部分来进行，这限制了网络从多任务学习中受益并且需要复杂的训练步骤。

本章提出了一个差异感知注意力网络（Difference-Aware Attention Network，

D2ANet），基于双时序高分辨率遥感图像同时进行建筑物分割和多级别变化检测的任务。首先应用卷积神经网络作为编码器从双时序遥感图像中提取特征，灾害前的图像在经过编码器的特征学习后作为一个解码器的输入来执行建筑物分割的任务。对于多级别变化检测任务，本章提出了一个差异感知注意力块（Difference-Aware Attention，D2A），基于灾前和灾后的图像特征探索不同级别变化之间的关系。差异感知注意力块包含一个双时序聚合（Dual-Temporal Aggregation Module，DTA）模块和差异注意力（Difference-Attention Module，DA）模块。灾前和灾后图像的特征在不同通道上可能呈现出不同的信息，双时序聚合模块基于成对的图像特征来激活变化敏感的通道，同时学习全局变化信息。除此之外，不同的自然灾害通常会对建筑物造成不同程度的损毁，如 图 5.1 中所示，充分利用多级别变化之间的相关性能帮助模型识别不同等级的建筑物损毁。本章进一步提出了一个差异注意力模块来获取一个特征立方体组内任意位置和任意通道之间的依赖关系，其中每个组中的小特征立方体都有表示多级别差异的潜力。在大规模建筑物变化检测数据集 xBD 上的大量实验表明，相比于其他的变化检测方法，本章所提出的差异感知注意力网络具有较大的优越性。

本章研究工作的主要贡献包括：

- 提出了差异感知注意力网络，基于双时序遥感图像同时进行建筑物分割和多级别变化检测两个任务，并在公开数据集上取得了较好的性能。
- 提出了一种双时序聚合模块，来激活图像特征中变化敏感的通道并捕获全局变化信息。
- 设计了一种差异注意力模块，通过充分利用多级别差异之间的局部关系来提高辨识不同级别变化的能力。相比于其他方法，其能够利用小特征立方体的划分来有效学习不同变化之间的相似性。

第二节 基于差异感知注意力网络的变化检测

考虑到实际复杂的环境和各种各样的自然灾害，不同区域内灾前和灾后图像之间的变化可能并不相同。为了研究不同变化之间的关系，本章提出了一种差异感知注意力网络（Difference-Aware Attention Network，D2ANet）用于从双时序遥感图像中进行建筑分割和多级别变化检测的任务，如图 5.2所示。在差异感知注意力网络中，本章提出了一个双时序聚合（Dual-Temporal Aggregation Module，DTA）模块用于学习不同时序图像的全局变化模式。同时进一步提出

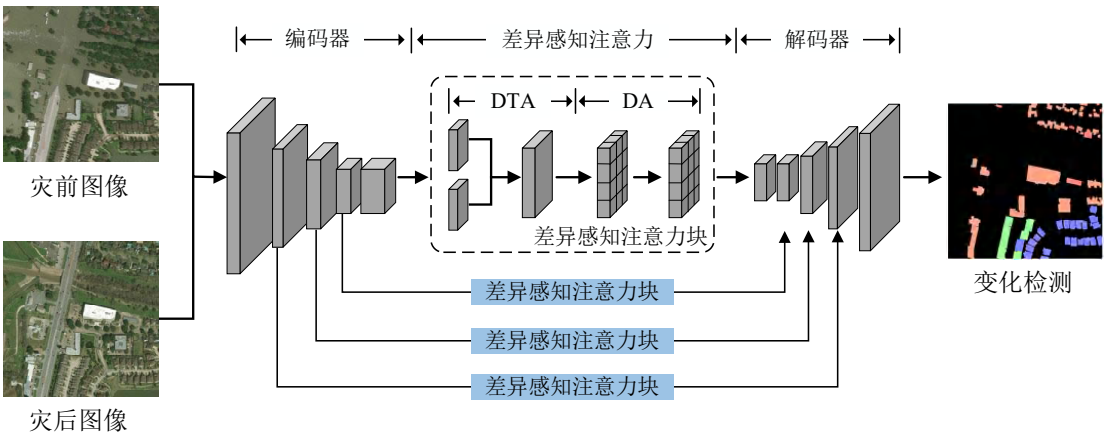


图 5.2 差异感知注意力网络（Difference-Aware Attention Network，D2ANet）的整体架构图。差异感知注意力块包含两个模块：双时序聚合（Dual-Temporal Aggregation Module，DTA）模块和差异注意力（Difference-Attention Module，DA）模块，用于探索全局变化模式和多级别变化之间的局部依赖性。

了一个差异注意力（Difference-Attention Module，DA）模块来捕获多级别变化之间的局部依赖性。

5.2.1 网络结构

考虑到残差网络[137]对于图像具有出色的特征提取能力，本章采用 ResNet-101 网络作为差异感知注意力网络的编码器。因为空洞卷积可以控制卷积核的感受野并调整特征图的分辨率，同时无需引入额外的参数，参考文献[75]，本章将扩张率为 $rate = 2$ 和 $rate = 4$ 的空洞卷积分别应用于编码器的最后两个残差阶段来提取更密集的特征。除此之外，本章采用了空洞空间金字塔池化模块（Atrous Spatial Pyramid Pooling Module，ASPP）[191]来增加特征点的视觉感受野，有效地学习多尺度特征。空洞空间金字塔池化模块利用四个具有不同扩张率的并行空洞卷积来捕获多尺度特征，然后与一个来自全局平均池化层的特征相融合。本章模型的输入是双时序遥感图像，即一对灾前和灾后的高分辨率遥感图像，在执行编码器网络后可以获取两组特征。

为了有效地处理建筑物分割和多级别变化检测两个任务，本章采用了两个解码器。每一个解码器包含五个块，其中每个块包含一个上采样层和一个卷积层。将灾前图像的特征输入到一个解码器中来生成一个二分类的建筑物掩码，完成建筑分割的任务。对于变化检测任务，考虑到不同时序图像对之间的全局和局部变化信息，本章提出了差异感知注意力块（Difference-Aware Attention

Block，D2A）来处理两组特征图。差异感知注意力块包含一个双时态聚合模块和一个差异注意力模块，这两个模块将会在后续的章 5.2.2和 章 5.2.3中详细介绍。如图 5.2所示，将成对的特征图输入到解码器之前，差异感知注意力块将对其进行处理以激活双时序遥感图像之间的全局变化模式，同时捕获多级别变化之间丰富的依赖关系。

对于这两个任务，本章在建筑物分割任务使用 Combo 损失函数[242]，在变化检测任务中使用交叉熵（Cross-Entropy）损失函数。本章所提出模型的总体损失函数可以被定义为：

$$L_{D2ANet} = L_{CE} + \lambda_1 L_{Combo}, \tag{5.1}$$

其中，λ_1 是一个常量。Combo 损失函数的定义为：

$$L_{Combo} = L_{Focal} + \lambda_2 L_{Dice}, \tag{5.2}$$

其中，λ_2 是用于平衡对应项的常量。Combo 损失函数由 Focal[52]损失函数和 Dice[243]损失函数的加权和得到，这两个损失函数的定义如下：

$$L_{Focal} = -\frac{1}{N}\sum_{i=1}^{N} \alpha y_i (1-\hat{y}_i)^{\gamma} \log(\hat{y}_i) - (1-\alpha)\hat{y}_i{}^{\gamma}(1-y_i)\log(1-\hat{y}_i), \tag{5.3}$$

$$L_{Dice} = 1 - \frac{2\sum_{i=1}^{N}(y_i \hat{y}_i)}{\sum_{i=1}^{N} y_i^2 + \sum_{i=1}^{N} \hat{y}_i{}^2}, \tag{5.4}$$

其中，y_i 表示位置 i 处像素为建筑物或背景的真值，$\hat{y}_i$ 为本章方法中对应建筑物分割任务的预测概率。N 是特征图中像素的数量。在 Focal 损失函数中，α 代表加权因子，γ 是可调节的聚焦参数，这两个参数被用于处理类别不平衡的问题。

5.2.2 双时序聚合模块

当检测灾前和灾后遥感图像中的变化时，双时序特征图之间特征级的差异是值得探索的。在双时序特征图中，不同的通道可能会呈现出不同的信息。具体来说，一些通道主要反映了差异性的变化模式，而另外的一些通道可能倾向于描述与背景相关的信息。本章提出了一个双时序聚合模块（Dual-Temporal Aggregation Module，DTA）来探索双时序特征图中对变化敏感的通道，同时获取全局的变化模式。

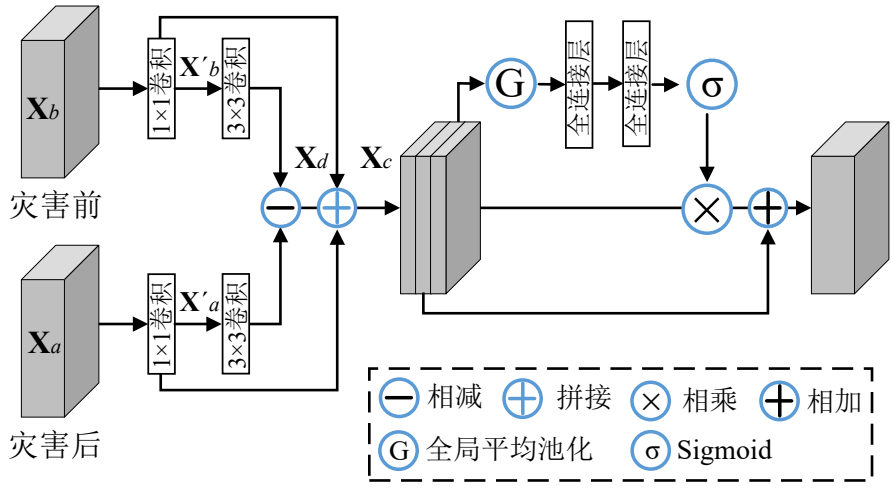


图 5.3 双时序聚合（Dual-Temporal Aggregation Module，DTA）模块的结构示意图。其融合了灾前图像的特征 $\mathbf{X}_b$ 和灾后图像的特征 $\mathbf{X}_a$，并利用时序差异来探索变化敏感的通道，激活全局变化模式。

双时序聚合模块的整体结构如图 5.3所示。令 $\mathbf{X}_b$ 和 $\mathbf{X}_a$ 表示灾害前后一对输入的特征图，这对特征图首先被送入 1×1 的卷积层来减少特征图通道数量以提高计算效率，之后即可以获得其输出特征 $\mathbf{X}'_b$ 和 $\mathbf{X}'_a$。再采用 3×3 的卷积层对特征图进行通道级的变换，并利用变换后的特征来计算时序差异。

$$\mathbf{X}_d = \mathbf{W}_{trans} * \mathbf{X}'_a - \mathbf{W}_{trans} * \mathbf{X}'_b, \tag{5.5}$$

其中，$\mathbf{W}_{trans}$ 是 3×3 通道卷积层的权重矩阵，用于对每一个通道执行变换。$\mathbf{X}_d$ 表示双时序特征图之间的差异，其有助于学习全局的变化信息。

相减的计算操作将会抑制背景的信息，这不利于建筑物的识别和变化检测。因此，本章沿着通道维度将特征图之间的差异 $\mathbf{X}_d$ 和 $\mathbf{X}'_b$、$\mathbf{X}'_a$ 拼接起来，以增强全局变化模式并且保存场景信息。这个操作的定义如下：

$$\mathbf{X}_c = concatenate(\mathbf{X}'_b, \mathbf{X}_d, \mathbf{X}'_a), \tag{5.6}$$

其中，“*concatenate*” 表示拼接操作。

最后，本章采用文献[44]中的挤压-激励模块（Squeeze-Excitation，SE）来激活变化敏感的通道，其通过通道注意力机制重新校准了拼接的特征 $\mathbf{X}_c$。

5.2.3 差异注意力模块

自然灾害通常会对建筑物造成不同程度的损坏。这意味着，与灾前的遥感图像相比，灾后的遥感图像中会存在不同等级的变化。评估灾害损毁的严重性

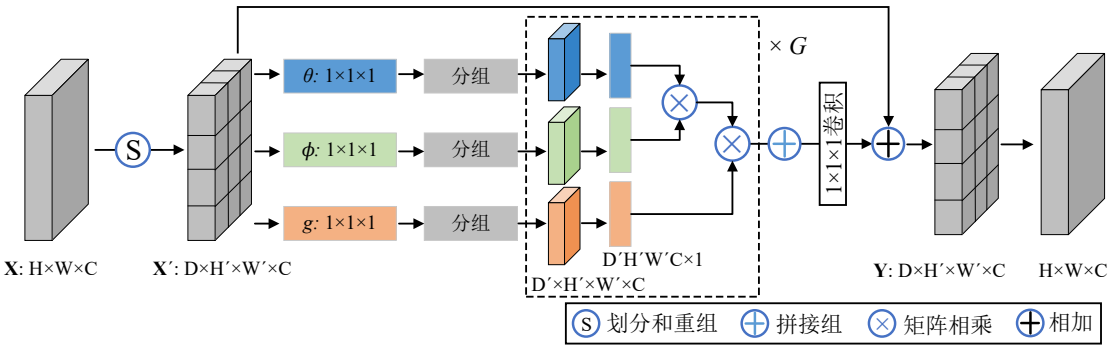


图 5.4 差异注意力（Difference-Attention Module，DA）模块的结构示意图。首先将输入的特征图 $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$ 沿高和宽两个维度进行分割，以获得形状为 $H' \times W' \times C$ 的小特征立方体，然后将所有小的特征立方体重新排列成 $\mathbf{X}' \in \mathbb{R}^{D\times H'\times W'\times C}$。图中虚线框代表每一组的特征进行矩阵乘法，即公式 (5.8)，G 是组的数量，“$\times G$” 表示公式 (5.8) 中的计算被重复了 G 次。

是灾害救援和物资援助的先决条件，而研究多级别变化之间的关系能提高判别不同灾害等级的能力。本章提出了一个差异注意力模块（Difference-Attention Module，DA）来学习不同级别变化之间的局部依赖性，如图 5.4中所示。

令 $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$ 表示差异注意力模块的输入特征图，其中 H、W 和 C 分别代表特征图的高度、宽度和通道数量。特征图 $\mathbf{X}$ 首先沿着高和宽维度被分成 $p \times p$ 块小立方体。每一块小特征立方体的形状大小为 $H' \times W' \times C$，其中 $H' = H/p$，$W' = W/p$。参数 p 是划分的数量，在本章开展的实验中设定 $p = 8$。每一个特征立方体假定包含灾前和灾后遥感图像之间一个等级的变化。之后所有的小特征立方体被重新排列组成 $\mathbf{X}' \in \mathbb{R}^{D\times H'\times W'\times C}$，其中 $D = p^2$ 表示小特征立方体的数量。

本章基于非局部模块（Non-Local Module）[18]提出了一个分组的自注意力机制来高效地学习多个特征立方体中任意位置和任意通道之间的相似性。$\mathbf{X}'$ 首先被送入三个 $1 \times 1 \times 1$ 的卷积层来学习变换参数，其定义如下：

$$\theta\left(\mathbf{X}'\right) = \mathbf{X}' * \mathbf{W}_\theta,\ \phi\left(\mathbf{X}'\right) = \mathbf{X}' * \mathbf{W}_\phi,\ g\left(\mathbf{X}'\right) = \mathbf{X}' * \mathbf{W}_g, \tag{5.7}$$

其中，$\mathbf{W}_\theta$、$\mathbf{W}_\phi$ 和 $\mathbf{W}_g$ 是卷积层的权重矩阵，$*$ 代表卷积操作。

直接在 $\theta(\mathbf{X}')$、$\phi(\mathbf{X}')$ 和 $g(\mathbf{X}')$ 上执行一些矩阵转换操作是不可行的，因为它们的维度是 $D \times H' \times W' \times C$，这会带来很高的计算复杂度。近年来，一些研究探索了将通道分成若干组的想法，并证明了这对于提高卷积神经网络的性能是有效的操作，例如 Xception[211]、ResNeXt[213]和组归一化（Group

Normalization)[214]。本章在差异注意力模块中引入了分组的操作，将维度 D 分成 G 个组。分组后特征图的形状大小为 $D' \times H' \times W' \times C$，其中 $D' = D/G$。每一个组由如下的矩阵乘法方程独立计算以得到 $\mathbf{Z}$。

$$\mathbf{Z} = f\left(vec\left(\theta\left(\mathbf{X}'\right)\right), vec\left(\phi\left(\mathbf{X}'\right)\right)\right) vec\left(g\left(\mathbf{X}'\right)\right), \tag{5.8}$$

其中，vec 表示特征图经过分组和改变形状后的向量，$vec\left(\theta\left(\mathbf{X}'\right)\right)$、$vec\left(\phi\left(\mathbf{X}'\right)\right)$ 和 $vec\left(g\left(\mathbf{X}'\right)\right) \in \mathbb{R}^{D'H'W'C \times 1}$。对偶函数 $f(\cdot,\cdot)$ 被提出用于计算一个组特征图中所有位置和所有通道之间的依赖关系。如文献[18]中所述，向量点乘可能是最简单的一种对偶函数，即如下运算：

$$f\left(vec\left(\theta\left(\mathbf{X}'\right)\right), vec\left(\phi\left(\mathbf{X}'\right)\right)\right) = vec\left(\theta\left(\mathbf{X}'\right)\right) vec\left(\phi\left(\mathbf{X}'\right)\right)^T. \tag{5.9}$$

之后，本章定义了如下函数来获取差异注意力模块的输出 $\mathbf{Y} \in \mathbb{R}^{D \times H' \times W' \times C}$。

$$\mathbf{Y} = concatenate\left(\mathbf{Z}\right) * \mathbf{W}_y + \mathbf{X}', \tag{5.10}$$

其中，“*concatenate*” 表示所有组沿着 D' 维度拼接。$\mathbf{W}_y$ 表示 $1 \times 1 \times 1$ 的分组卷积层的权重，$\mathbf{Y}$ 被重新排列以确保它的形状大小和输入的特征图 $\mathbf{X}$ 一致。

差异注意力模块中，本章设置组的数量 $G = 16$ 来使分组注意力模块能够捕获一个组内特征图任何位置和任何通道之间的依赖关系，其中每组包含 4 个小的特征立方体。因为每一个被分割后的小特征立方体都有表示一个等级差异的潜力，所以差异注意力模块能够学习多级别差异之间的相似性，并且增强对双时序遥感图像之间多级别变化的判别能力。除此之外，本章将会在实验中提供更多关于差异注意力模块中参数的分析，如小特征立方体的数量 D 和分组的数量 G。

第三节 实验结果与分析

在本小节，将会介绍本章实验中所使用的公开数据集和评测指标。除此之外，还展示了与其他变化检测方法的详细比较结果和本章模型的消融实验，来验证所提出模型和各个模块的有效性。

5.3.1 数据集

在本章的实验中，采用了大规模建筑物损毁变化检测数据集 xBD[7]来进行多级别变化检测和建筑物分割任务。这个数据集提供了 19 种不同自然灾害的成

对灾前和灾后遥感图像，包括地震、野火、海啸和火山爆发等。xBD 数据集包含 22068 张遥感图像（即 11034 对灾前和灾后的图像）和 850736 个建筑物的标注，总面积为 45362 平方公里。遥感图像的像素分辨率为 1024×1024，空间分辨率为 0.3 米/像素。同时，此数据集还提供了联合损毁等级（Joint Damage Scale）来创建一个统一的评估尺度，用于根据遥感图像对多种灾害类型造成的建筑物损毁进行评估，损毁程度包括无损毁、轻微损毁、严重损毁和完全损毁。xBD 数据集被划分为训练集、测试集、Holdout 和 Tier3 四个部分。本章选取包含 2799 对遥感图像的训练集和包含 933 对遥感图像的测试集，用于本章实验的训练和测试。

5.3.2 评测指标

为了验证本章所提出的差异感知注意力网络在多级别变化检测任务和建筑物分割任务上的表现，本节采用了文献[7]中提出的 xView2 比赛的评测指标。其定义如下：

$$S_{xView2} = 0.3F1_{loc} + 0.7F1_{damage}, \tag{5.11}$$

$$F1_{damage} = \frac{n}{\frac{1}{F1_{cls1}} + \cdots + \frac{1}{F1_{clsn}}}, \tag{5.12}$$

其中，$F1_{loc}$ 表示建筑物分割的 F1 值，用于获取每一处像素的预测值和灾前遥感图像真值之间的一致性。$F1_{damage}$ 是变化检测的 F1 值，其统计了灾后遥感图像中每个多边形内像素真值和预测值之间的一致性。$F1_{cls1} \ldots F1_{clsn}$ 表示 n 个损毁等级的变化检测 F1 值。值的注意的是，评测指标 S_{xView2} 使用分割任务 F1 值和多级别变化检测 F1 值调和均值的加权平均值，会惩罚对具有大量建筑多边形类别的过拟合。同时由于 xBD 数据集中不同损毁等级的分布严重不平衡，这使得该评测指标比较有挑战性。

5.3.3 实现细节

本章所提出的差异感知注意力网络采用了随机梯度下降（SGD）优化器，批大小（Batch Size）设置为 8。初始学习率的值被设置为 0.01，权重衰减系数（Weight Decay Coefficients）和动量系数（Momentum）分别被设置为 5×10^{-4} 和 0.9。本章的模型被训练了 150 个 Epoch，并在训练中采用 ‘poly’ 学习率策略，该策略通过乘 $(1 - \frac{iter}{maxiter})^{power}$ 来逐渐降低学习率，其中 $power = 3$ 。本章使用深

度学习框架 PyTorch[223] 来实现所提出的差异感知注意力网络，同时本节的实验是在 4 个有 24GB 显存的 NVIDIA RTX TITAN GPU 上实现的。在训练过程中，本节采用了包括随机旋转、尺度调节、水平翻转和高斯模糊的数据增强方法来提高模型的泛化能力。最后将 xBD 数据集中的图像随机裁剪到一个用于训练的固定大小 512×512。

5.3.4 与现有方法的对比

本小节将本章所提出的差异感知注意力网络与其他方法在 xBD 数据集[7]上进行了建筑物分割和多级别变化检测任务的比较。这些方法包括：Baseline[7]、Siamese-UNet(ResNext50)[123]、Siamese-UNet(DPN92)[123]、Dual-HRNet[244]、Dual-Temporal Fusion[138]和 RescueNet[245]。Gupta 等[7]提出的 Baseline 介绍了一种经过调整的 U-Net[46]结构进行建筑物的分割，同时利用在 ImageNet[228]上预训练的 ResNet-50[137]模型进行灾害损毁等级的分类任务。在变化检测任务中，Siamese-UNet 是一个应用广泛的架构，其基于灾前的遥感图像训练 U-Net 进行建筑物的分割，同时在灾后遥感图像上使用 Siamese-UNet 并利用分割模型中的共享权重进行变化检测。基于两个不同的主干网络 ResNext50[213]和 DPN92[54]，本小节实现了两种 Siamese-UNet 结构的模型。Dual-HRNet[244] 包含两个高分辨率网络结构（High-Resolution Network，HRNet）和一个基于双时序图像的用于将两个任务融合的模块。Dual-Temporal Fusion[138]采用包含一个特征金字塔网络模块[140]的 Mask R-CNN 模型[139]和一个语义分割分支来完成这两项任务。RescueNet[245]采用多尺度时序特征来识别建筑物中的变化，其可以同时分割建筑物并对建筑物损毁等级进行分类。为了公平客观的进行比较，本小节重新执行了这些方法，并在同一个数据集上开展了实验。

5.3.4.1 定量结果比较

在 xBD 数据集[7]上与其他方法的定量比较结果如 表 5.1中所示。本章所提出的差异感知注意力网络（Difference-Aware Attention Network，D2ANet）在建筑物分割任务中的 F1 值是 84.78%，相较于第二优的方法 RescueNet[245]提高了 0.69%。对于变化检测任务，本章方法 D2ANet 相较于 Dual-Temporal Fusion[138]，在变化检测 F1 值上提高了 3.29%。D2ANet 在四个损毁等级上都取得了最佳的性能，例如，轻微损毁的 F1 值为 53.13%，严重损毁的 F1 值为 72.95%，相比于 Dual-Temporal Fusion[138]分别提高了 2.36% 和 4.24%。此外，本小节也测试了

表 5.1 本章所提出的差异感知注意力网络（Difference-Aware Attention Network，D2ANet）与其他方法在 xBD 数据集[7]上的定量比较（%）。‘总体 F1’ 代表总体的 F1 值，即公式 (5.11) 中的 S_{xView2}。‘分割 F1’ 和 ‘变化 F1’ 分别指建筑物分割和变化检测的 F1 值；‘无 F1’、‘轻微 F1’、‘严重 F1’ 和 ‘完全 F1’ 分别指无损毁、轻微损毁、严重损毁和完全损毁建筑的变化检测 F1 值。

方法	总体 F1	分割 F1	变化 F1	无 F1	轻微 F1	严重 F1	完全 F1
Baseline[7]	28.41	80.48	6.09	65.79	7.08	2.16	26.40
Siamese-UNet(ResNext50)[123]	67.33	79.81	61.97	76.86	45.39	64.17	71.86
Siamese-UNet(DPN92)[123]	69.18	83.56	63.02	81.55	43.90	66.21	75.01
Dual-HRNet[244]	71.35	83.61	66.09	86.43	48.66	69.11	71.80
Dual-Temporal Fusion[138]	72.52	82.76	68.13	86.29	50.77	68.71	77.71
RescueNet[245]	70.23	84.09	63.94	86.09	45.72	62.76	76.15
D2ANet(ResNet-50)	**73.40**	**84.72**	**68.54**	**89.64**	49.79	**69.82**	**78.16**
D2ANet(ResNet-101)	**75.43**	**84.78**	**71.42**	**90.89**	**53.13**	**72.95**	**80.14**

表 5.2 本章所提出的差异感知注意力网络 D2ANet 与其他方法的参数量和运行时间分析。‘推理时间’ 表示在一张固定大小图像上推理的时间。

方法	参数数量（M）	推理时间（s/img）
Baseline[7]	44.2	0.039
Siamese-UNet(ResNext50)[123]	69.1	0.068
Siamese-UNet(DPN92)[123]	94.8	0.437
Dual-HRNet[244]	59.5	0.055
Dual-Temporal Fusion[138]	43.9	0.035
RescueNet[245]	44.3	0.043
D2ANet(ResNet-50)	43.6	0.024
D2ANet(ResNet-101)	62.6	0.063

使用 ResNet-50 作为主干网络的 D2ANet 的实验结果，其结果同样优于其他方法。相较于 Dual-Temporal Fusion[138]，使用 ResNet-50 的 D2ANet 的总体 F1 得分提高了 0.88%，在建筑物分割任务上 F1 得分提高了 1.96%。值得注意的是，xView2 比赛中的一些参与者取得了比较好的结果，但是这些参与者通常集成了多个语义分割模型并将两个任务分开进行训练，这样的做法实现起来比较复杂同时效率也比较低。本章方法利用单模型网络高效地获得了比目前广泛使用的 Siamese-UNet 模型更好的性能。D2ANet 在建筑物分割和多级别变化检测任务中更好的性能使其更适用于灾害评估。

本小节也分析了本章方法 D2ANet 和其他变化检测方法的参数量和运行时间，如 表 5.2中所示。为了公平的进行比较，所有方法的比较实验是在有 4 个 NVIDIA RTX TITAN GPU 的工作站上进行的，并采用推理一张 1024×1024 大

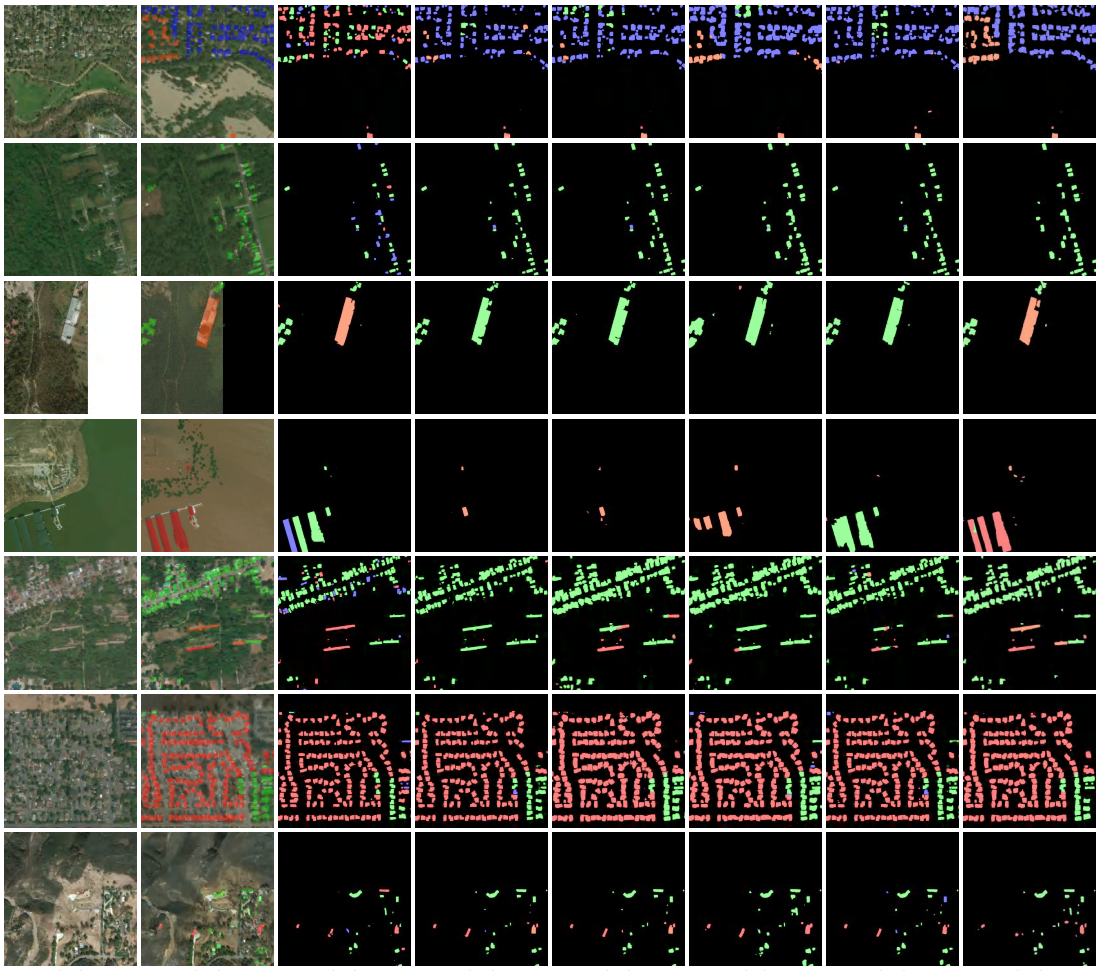


图 5.5 本章所提出的差异感知注意力网络 D2ANet 与其他方法在 xBD 数据集[7]上的定性比较结果。第一行到第七行是不同自然灾害的可视化结果，包括哈维飓风、佛罗伦萨飓风、地震、洪水、海啸、野火和火灾。(a) 灾害前遥感图像；(b) 包含多级别变化真值的灾害后遥感图像。其中的四种颜色代表四种变化等级，绿色、蓝色、橙红色和红色分别表示无损毁、轻微损毁、严重损毁和完全损毁。(c)-(h) Baseline[7]、Siamese-UNet(ResNext50)[123]、Siamese-UNet(DPN92)[123]、Dual-HRNet[244]、Dual-Temporal Fusion[138]、和本章 D2ANet 的可视化结果。

小的图片来比较运行时间。因为 Baseline[7]和 Dual-Temporal Fusion[138]都使用 ResNet-50 作为主干网络，为了公平比较，本小节也列出了使用 ResNet-50 作为主干网络的 D2ANet 的结果。可以看到，使用 ResNet-50 作为主干网络的 D2ANet 的参数量为 43.6M，这个参数量是最少的并且比 Dual-Temporal Fusion[138] 少了 0.3 M。然而，使用 ResNet-50 作为主干网络的 D2ANet 的总体 F1 得分比其他方法都好。除此之外，本章方法 D2ANet 推理一张图像所用的计算时间也更

表 5.3 双时序聚合（Dual-Temporal Aggregation Module，DTA）模块和差异注意力（Difference-Attention Module，DA）模块的消融实验（%）。编号 1 是基于主干网络 ResNet-101 的分割模型。通过添加双时序聚合模块 DTA 和差异注意力模块 DA 来验证这两个模块的有效性（编号 2 和编号 3）。编号 4 是本章提出的完整版本的 D2ANet。

编号	DTA	DA	总体 F1	分割 F1	变化 F1	无 F1	轻微 F1	严重 F1	完全 F1
1			71.47	83.78	66.20	86.54	48.84	69.19	71.70
2	✔		73.35	83.86	68.85	90.02	50.67	70.19	76.88
3		✔	74.53	**84.96**	70.06	90.53	51.09	71.35	**80.27**
4	✔	✔	**75.43**	84.78	**71.42**	**90.89**	**53.13**	**72.95**	80.14

表 5.4 对差异注意力（Difference-Attention Module，DA）模块中不同特征立方体数量 D 的消融实验（%）。

编号	立方体数 D	总体 F1	分割 F1	变化 F1	无 F1	轻微 F1	严重 F1	完全 F1
1	16	71.42	84.47	65.83	89.82	45.61	66.00	79.69
2	64	**75.43**	**84.78**	**71.42**	**90.89**	**53.13**	**72.95**	**80.14**
3	256	67.26	83.97	60.10	84.59	40.26	61.44	73.42

少。当采用 ResNet-101 作为主干网络时，与使用 ResNet-50 作为主干网络的结果相比，总体 F1 得分提高了 2.03%，如 表 5.1所示。主干网络是 ResNet-101 的 D2ANet 的参数数量仍然小于 Siamese-UNet (ResNext50)[123]和和 Siamese-UNet (DPN92)[123]。Siamese-UNet 架构使用了两个网络用于建筑物分割和变化检测的任务，这引入了更多的参数。

5.3.4.2 可视化结果比较

本章方法 D2ANet 和其他变化检测方法的可视化比较结果如 图 5.5所示，其中包含了 7 个不同灾害的样例，分别是：哈维飓风、佛罗伦萨飓风、地震、洪水、海啸、野火和火灾。值得注意的是，本章方法得到的建筑物分割结果与真值更加一致。此外，D2ANet 在多级别变化检测任务上的结果也更加准确。例如，在 图 5.5中第一行的结果中，有几座建筑物被飓风破坏，本章 D2ANet 的可视化结果更接近真值。但是其他方法无法正确识别建筑物损毁的类型，尤其是 Baseline[7]和 Dual-Temporal Fusion[138]。此外，诸如 Siamese-UNet(ResNext50)[123] 和 Siamese-UNet(DPN92)[123] 等方法无法识别出完整的建筑物。在 xBD 数据集[7]上的可视化结果表明，本章提出的差异感知注意力网络对于从双时序遥感图像中进行建筑物分割和多级别变化检测任务具有优秀的性能。

表 5.5 对差异注意力（Difference-Attention Module，DA）模块中不同分组数量 G 的消融实验（%）。

编号	分组数 G	总体 F1	分割 F1	变化 F1	无 F1	轻微 F1	严重 F1	完全 F1
1	1	70.66	84.44	64.77	88.26	45.06	69.68	72.00
2	8	71.26	82.11	66.61	89.35	47.14	69.27	75.68
3	16	**75.43**	**84.78**	**71.42**	**90.89**	**53.13**	**72.95**	**80.14**
4	32	72.86	83.85	68.15	89.27	49.07	68.82	79.50

表 5.6 对双时序聚合（Dual-Temporal Aggregation Module，DTA）模块采用不同配置的消融实验，其能够用于捕获双时序遥感图像之间的全局变化模式（%）。‘无时序差异’表示不应用时序差异，即 $\mathbf{X}_c$ 是灾前特征 $\mathbf{X}'_b$ 和灾后特征 $\mathbf{X}'_a$ 的拼接。‘无通道卷积’表示不采用 3×3 的通道卷积，即时序差异 $\mathbf{X}_d = \mathbf{X}'_a - \mathbf{X}'_b$。‘无 SE 模块’表示不在 DTA 模块中应用 SE 模块。

模块配置	总体 F1	分割 F1	变化 F1	无 F1	轻微 F1	严重 F1	完全 F1
无时序差异	71.31	84.40	65.70	87.15	46.15	67.72	77.08
无通道卷积	72.38	84.71	67.09	88.63	48.70	66.53	78.26
无 SE 模块	74.23	84.59	69.79	89.18	51.66	71.53	78.33
D2ANet	**75.43**	**84.78**	**71.42**	**90.89**	**53.13**	**72.95**	**80.14**

5.3.5 消融实验

5.3.5.1 模块的有效性

为了建模全局变化模式以及捕获多级别变化之间的局部相关性，本章在差异感知注意力网络中提出了双时序聚合（Dual-Temporal Aggregation Module，DTA）模块和差异注意力（Difference-Attention Module，DA）模块，验证这两个模块有效性的实验结果如 表 5.3所示。编号 1 是基于主干网络 ResNet-101 的编码器-解码器架构下的分割模型，在添加了 DTA 和 DA 模块后，建筑物分割任务的 F1 值从 83.78% 提升到了 83.86% 和 84.96%。变化检测的 F1 值分别提高了 2.65% 和 3.86%，后者验证了 DA 模块可以增强对灾前和灾后遥感图像之间多级别变化的判别能力。结合两个模块后，总体 F1 值提高了 3.96%，变化检测的 F1 值提高了 5.22%。值得注意的是，添加 DA 模块比完整版本的 D2ANet 获得了更好的建筑分割 F1 值和完全损毁 F1 值，但总体 F1 值和变化检测的 F1 值较低。由于利用这两个模块有助于本章提出的模型有效地学习全局和局部的多级别变化信息，完整的 D2ANet 在识别不同的变化方面具有较大的优势。

5.3.5.2 差异注意力模块中不同配置的影响

在差异注意力模块 DA 中，有两个重要的参数：特征立方体数量 D 和分组数量 G。本小节分析了不同参数配置对实验结果的影响，对参数 D 的分析结果列在了 表 5.4中。可以看到，$D=64$ 是差异注意力模块的最佳配置，而更多或更少的特征立方体数量会阻碍模型性能的提高。例如，当特征立方体数量 $D=64$ 时，总体 F1 值为 75.43%，相较于模型配置 $D=16$ 和 $D=256$ 分别提高了 4.01% 和 8.17%。小特征立方体是通过划分特征图获得的，每个特征立方体被认为只包含双时序图像之间一种等级的变化。如果小特征立方体的数量较少，则每个特征立方体可能包含了多个等级的变化，这样会阻碍模型学习多级别变化之间的相似性。而当划分出较多的特征立方体时，每一个特征立方体会包含较少的像素，这样可能会限制特征的学习。由上述分析可得，表 5.4中的实验结果是符合预期的。

分组操作是差异注意力模块 DA 中的另一个重要策略。如 表 5.5中所展示的：当分组数量 $G=16$ 时，本章的模型获得了最佳的总体 F1 值，相比于其他模型配置 $G=8$ 和 $G=32$ 分别提高了 4.17% 和 2.57%。除此之外，建筑物分割的 F1 值和变化检测的 F1 值在分组数 $G=16$ 时也达到了最好。这个结果符合预期，因为 DA 模块考虑了不同变化之间像素点的亲和性。如果分组数比较少，每个组会包含较多的小特征立方体，从而会限制模型的优化。而当应用较多的分组时，其将会限制模型获取多级别变化之间丰富的依赖关系。如果不使用分组的操作，即令分组数 $G=1$，表 5.5中的结果是最差的。

根据上述的实验，本章将差异注意力模块中特征立方体的数量设置为 $D=64$，将分组的数量设置为 $G=16$。

5.3.5.3 双时序聚合模块中不同配置的影响

在双时序聚合模块 DTA 中，本章设计了多种操作来探索变化敏感的通道，同时捕获全局变化模式，这些操作包括：时序差异、通道卷积和 SE 模块[44]。如表 5.6中所示，本小节提供了 DTA 模块在不同配置下的实验结果，以验证上述三种操作的有效性。当不采用时序差异，即 $\mathbf{X}_c$ 只是灾前特征 $\mathbf{X}'_b$ 和灾后特征 $\mathbf{X}'_a$ 的拼接时，总体 F1 值比本章模型 D2ANet 低了 4.12%。如果不应用 3×3 的通道卷积层，即时序差异 $\mathbf{X}_d=\mathbf{X}'_a-\mathbf{X}'_b$，总体 F1 值比 D2ANet 低了 3.05%。此外，应用 SE 模块[44]将总体 F1 值增加了 1.20%。以上结果表明，双时序聚合模块中

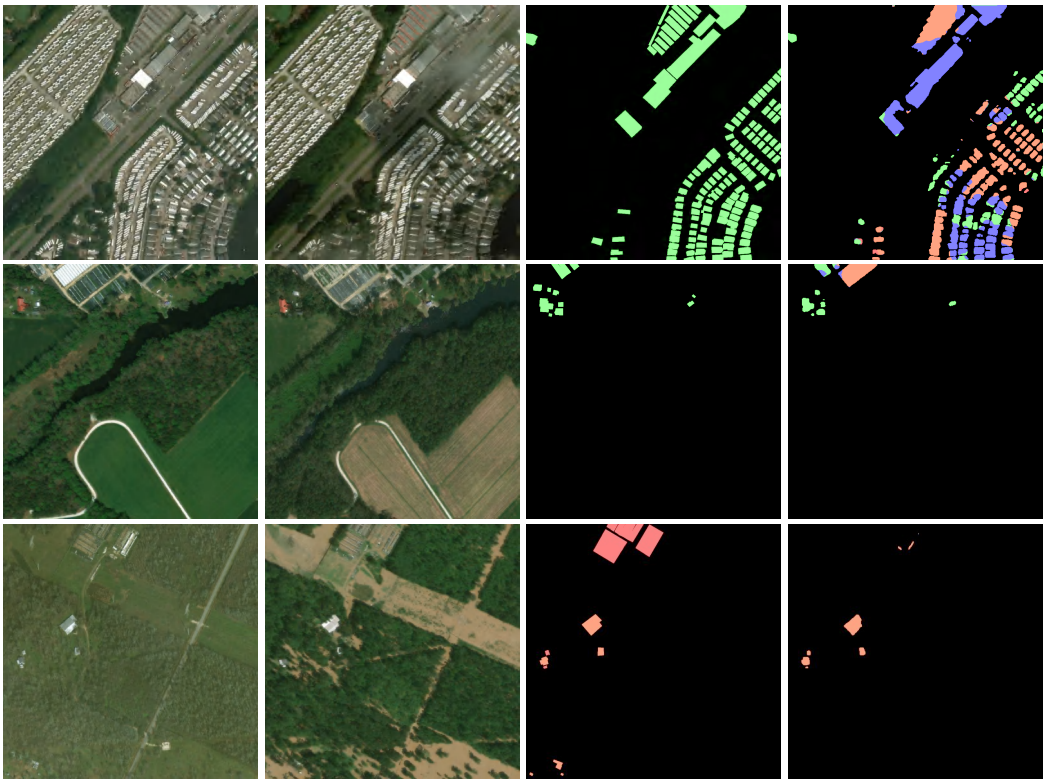


图 5.6 本章方法差异感知注意力网络 D2ANet 的一些失败案例的可视化结果。第一行到第三行分别是 xBD[7]数据集中的三个样例。(a) 灾害前遥感图像；(b) 灾害后遥感图像；(c) 变化检测任务的真值，其中绿色、蓝色、橙红色和红色分别表示无损毁、轻微损毁、严重损毁和完全损毁。(d) D2ANet 的变化检测结果。

的几种不同操作可以提高变化检测的结果，也进一步验证了双时序聚合模块设计的合理性和有效性。

5.3.6 失败案例分析

在上述的实验中，本章所提出的差异感知注意力网络 D2ANet 在 xBD[7]数据集上取得了更好的变化检测性能，但是其仍然存在一些失败的案例。在图 5.6中，本小节展示了三个失败的样例。对于图 5.6第一行，在双时序遥感图像之间存在光照差异和云层遮挡的情况，D2ANet 无法识别出被云层遮挡区域的变化。此外，D2ANet 将图 5.6第二行中的蔬菜大棚识别为建筑。因为建筑物的颜色和纹理与周围的农田很相似，D2ANet 没能识别出图 5.6第三行中的建筑物。在下一步的工作中，将考虑结合其他类型的遥感图像，例如高光谱图像等，来检测这些困难区域的受损建筑物。

第四节 本章小结

本章提出了一种差异感知注意力网络（Difference-Aware Attention Network，D2ANet），基于双时序遥感图像同时进行建筑物分割和多级别变化检测的任务。首先应用卷积神经网络作为编码器来提取双时序图像的特征，将经过编码器的灾前特征输入一个解码器来完成建筑物分割的任务。对于多级别变化检测的任务，本章提出了差异感知注意力块，其包含一个双时序聚合（Dual-Temporal Aggregation Module，DTA）模块和差异注意力（Difference-Attention Module，DA）模块。灾前和灾后图像的特征在不同通道上可能呈现出不同的信息，双时序聚合模块基于成对的图像特征来激活变化敏感的通道，同时学习全局变化信息。此外，设计了差异注意力模块来获取一个特征立方体组内任意位置和任意通道之间的依赖关系，其中的小特征立方体有表示多级别差异的潜力。大量的实验验证了本章所提出的差异感知注意力网络在变化检测任务中的优越性，同时本章还进行了大量的消融实验，以验证所提出的两个模块的有效性。

第六章　基于全局差异与局部注意力的双时序图像变化检测

本章选取时间序列图像中的双时序高分辨率遥感图像进行变化检测的任务。为了克服基于卷积神经网络的方法缺乏建模时间序列图像长程依赖关系能力的问题，本章结合 Transformer 提出了基于全局差异与局部注意力的时间序列图像变化检测模型。本章的章节安排为：第一节是对本章研究背景、研究内容和创新点的介绍；第二节介绍了所提出的基于全局差异与局部注意力的变化检测模型；第三节给出了本章所提出的变化检测模型与其他方法的对比实验及相关分析；第四节是对本章内容进行的总结。

第一节　引言

对地表覆盖的变化检测是监测全球和区域环境的一项强大的技术。遥感图像能够提供多时序以及大尺度的信息，得益于这个优势，变化检测已经在多个领域得到了广泛的应用[100-103]。灾害检测可以看为是变化检测的一个子任务，其是指当自然灾害发生时及时有效地对损毁的位置和严重程度进行评估，这对于快速开展灾害响应活动和人道主义援助来说非常重要。基于双时序高分辨率遥感图像的变化检测可以高效地满足灾害损毁评估的需求，近年来受到了越来越多的关注。

传统的变化检测研究通常基于多个时序的图像，采用感知像素之间差异的方法[105,117]，然而这些为特定数据而设计的变化检测方法在处理不同区域的图像时性能较差[114]。近年来，基于卷积神经网络（CNN）的模型被广泛地应用于变化检测任务[134,136]。孪生神经网络（Siamese Neural Network）是基于两个 CNN 构建的耦合结构，其以两个图像样本作为输入，并输出它们的特征表征以比较两张图像的相似度，这一特点使得孪生神经网络在变化检测任务中非常有效。一系列研究利用孪生神经网络来检测地表覆盖的变化[124-126]，由于需要使用两个网络来处理不同时序的图像，这些方法会引入更多的参数量。此外，由于卷积运算的局部性，基于卷积神经网络的模型缺乏建模长程依赖关系的能力。为了克服这一限制，一些研究[5,147]针对不同的图像识别任务为卷积神经网络构建了自注意力机制。Wang 等人[18]提出了非局部操作（Non-local）来对视频分

类任务中任意位置之间的长程依赖关系进行建模。为序列到序列的预测任务而设计的 Transformer，在学习长程依赖关系方面展现了非常出色的能力。最近，许多计算机视觉任务中的研究应用 Transformer 进行探索，并取得了较好的效果[159,161,246-247]。

本章提出了一种基于全局差异与局部注意力的变化检测模型（Global Difference and Local Attention，GDLA），基于双时序遥感图像同时进行建筑分割和多级别变化检测。结合卷积神经网络擅长学习局部细节特征和 Transformer 能够建模长程依赖关系的优势，本章采用混合卷积神经网络和 Transformer 的架构作为编码器。同时采用渐进式上采样结构作为解码器，输出建筑分割和多级别变化检测的结果。一张遥感图像通常会覆盖较大的区域，其中包含的建筑在灾害发生后会受到不同程度的损毁。由于小尺寸的图像块会限制模型学习不同变化之间长程依赖关系的能力，本章提出了全局差异模块（Global Difference，GD）来缓解这个问题并探索全局变化模式。同时设计了局部门控注意力模块（Local Gated Attention，LGA），来学习局部变化差异并增强对双时序遥感图像间多级别变化的判别能力。在大规模建筑物损毁变化检测数据集 xBD 上的大量实验表明，本章所提出的方法 GDLA 应对变化检测任务是有效的。

本章研究工作的主要贡献包括：

- 提出了基于全局差异与局部注意力的网络模型，基于双时序遥感图像同时进行建筑物分割和多级别变化检测任务，其在大规模建筑物损毁变化检测数据集 xBD 上取得了较好的性能。
- 设计了一种全局差异模块，来学习全局变化模式并提高对双时序图像之间变化的整体认识。
- 提出了局部门控注意力模块，通过探索多级别变化之间的局部依赖性来提高对不同变化的判别能力。

第二节 基于全局差异与局部注意力的变化检测

考虑到现实中的复杂环境和多种多样的自然灾害，灾害发生前后的图像之间通常存在着多种级别的变化。为了学习多级别变化之间的关系，本章提出了一种基于全局差异与局部注意力的变化检测模型（Global Difference and Local Attention，GDLA），利用双时序高分辨率遥感图像同时进行建筑分割和多级变化检测两个任务，其网络结构如图 6.1所示。其中本章提出了一种全局差异模块

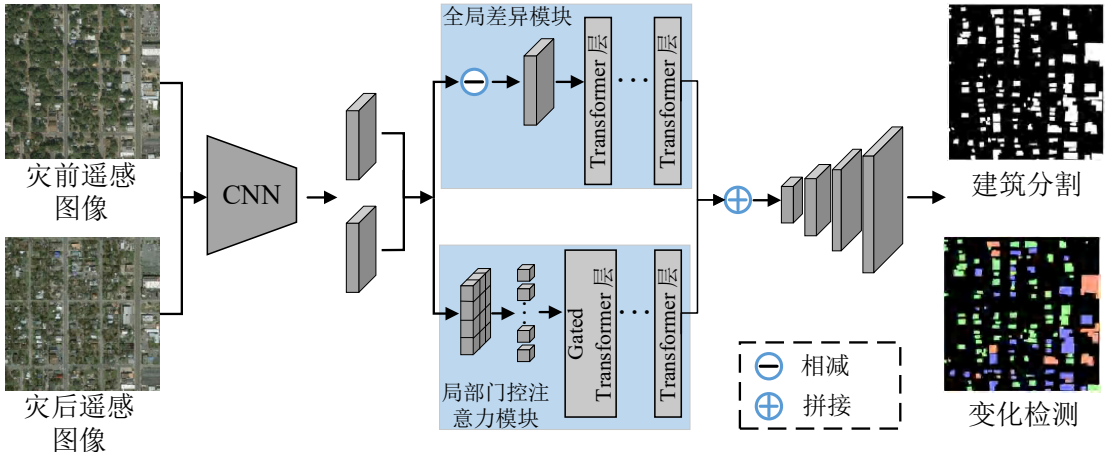


图 6.1 本章所提出的基于全局差异与局部注意力（Global Difference and Local Attention，GDLA）的变化检测方法总体架构图。双时序遥感图像被输入到混合卷积神经网络和 Transformer 架构的编码器来提取特征，之后利用渐进式上采样解码器来同时完成建筑物分割和变化检测。对于编码器中的 Transformer，本章提出了两个模块：全局差异模块和局部门控注意力模块，来分别获取全局和局部的变化模式。

（Global Difference，GD）来学习全局变化模式。同时进一步设计了一种局部门控注意力模块（Local Gated Attention，LGA）来探索图像中多级别变化之间的局部依赖关系。

6.2.1 网络结构

本章结合卷积神经网络擅长学习图像的局部细节特征和 Transformer 能够建模长程依赖关系的优势，采用混合卷积神经网络和 Transformer 的结构作为编码器，将卷积神经网络输出的降采样后的特征作为 Transformer 的输入。对于输入模型的双时序高分辨率遥感图像，即一对灾前和灾后的图像，首先利用 ResNet-50[137]作为特征提取器来生成一对特征图。基于生成的双时序特征图，本章提出了一个全局差异模块和一个局部门控注意力模块，来分别学习双时序图像之间的全局和局部变化模式。

本章采用渐进式上采样结构作为解码器，其中包含四个解码器块来对特征图进行上采样从而达到输入图像的分辨率。每个解码器块依次包含一个 3×3 卷积层、一个批归一化层、一个 ReLU（Rectified Linear Unit，整流线性单元）层和一个 $2 \times /4 \times$ 的上采样算子。在最后一个解码器块后有两个分支：分别是建筑物分割分支和多级别变化检测分支。每个分支都包含一个 3×3 卷积层，分别输出通道数量为 1 和 5 的预测结果。全局差异模块和局部门控注意力模块的输出

特征在融合后被输入到解码器中，进而输出建筑物分割和变化检测的结果。本章在建筑物分割任务中采用了 Combo 损失函数[242]，在变化检测任务中采用加权交叉熵损失函数（Cross-Entropy）。Combo 损失函数的定义为：

$$L_{Combo} = \lambda_{c1} L_{Dice} + \lambda_{c2} L_{Focal}, \tag{6.1}$$

其中，λ_{c1} 和 λ_{c2} 是平衡系数。Combo 损失函数是 Dice 损失[243]和 Focal 损失[52]的加权和，其定义如下：

$$L_{Dice} = 1 - \frac{2\sum_{i=1}^{N}(y_i \hat{y}_i)}{\sum_{i=1}^{N} y_i^2 + \sum_{i=1}^{N} \hat{y}_i{}^2}, \tag{6.2}$$

$$L_{Focal} = -\frac{1}{N}\sum_{i=1}^{N} \alpha y_i (1-\hat{y}_i)^{\gamma} \log(\hat{y}_i) - (1-\alpha)\hat{y}_i{}^{\gamma}(1-y_i)\log(1-\hat{y}_i), \tag{6.3}$$

其中，N 为特征图中像素的数量。y_i 是指位置 i 处像素为背景或建筑物的真值，$\hat{y}_i$ 是本章方法对建筑物分割任务的预测概率。α 是一个权重因子，γ 是一个可调节参数，用于处理类别不均衡的问题。

加权交叉熵损失的定义为：

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{n} w_c y_i^c \log(\hat{y_i^c}), \tag{6.4}$$

其中，n 为变化的等级数量，w_c 为不同级别变化的缩放权重。y_i^c 是指位置 i 处像素的变化等级为 c 的真值，$\hat{y_i^c}$ 为本章方法预测像素 i 处的变化等级为 c 的概率。

本章方法的总体损失函数被定义为：

$$L_{GDLA} = \lambda_1 L_{Combo} + \lambda_2 L_{CE}, \tag{6.5}$$

其中，λ_1 和 λ_2 是用于平衡两个损失函数的常数。

6.2.2 全局差异模块

利用图像块来训练 Transformer 能够有效的减少运算量，加快训练速度。然而，只利用图像块对于遥感图像的多级别变化检测是不充分的。一张高分辨率遥感图像会覆盖比较大的区域，其中包含灾害发生后不同变化程度的建筑。图像块的尺寸相比于整幅图像要小很多，这会限制模型学习不同变化之间长程依

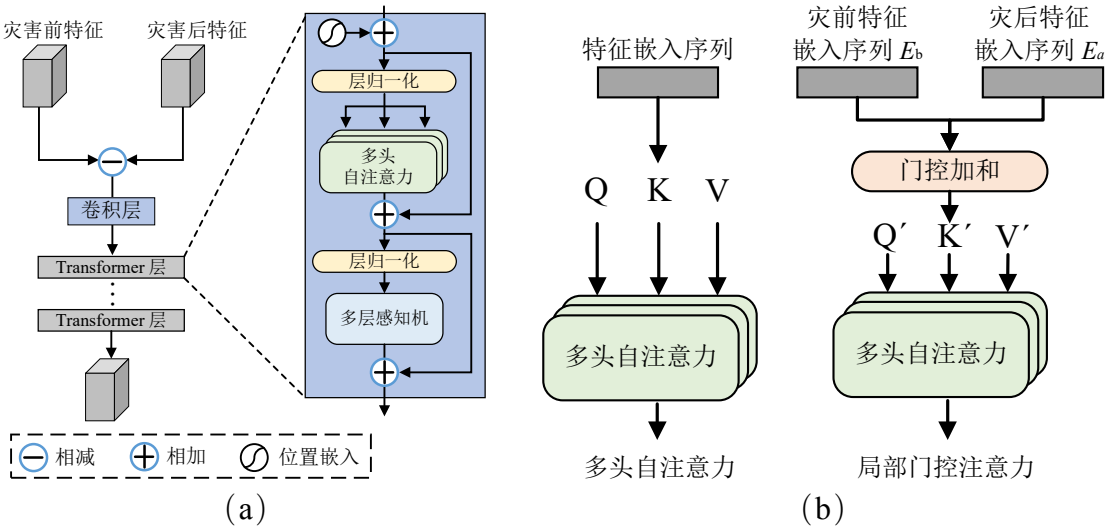


图 6.2 （a）全局差异模块（Global Difference，GD）结构示意图。（b）原本的多头自注意力（Multi-Head Self-Attention）和本章所提出的局部门控注意力（Local Gated Attention）的结构对比。利用门控加和操作，灾害前后图像块的特征可以自适应的进行融合。

赖关系的能力。为了学习全局变化信息并提高对图像中所有像素的整体理解，本章设计了一种全局差异模块（Global Difference，GD）。

全局差异模块的结构如图 6.2（a）所示。令 $\mathbf{X}_b$、$\mathbf{X}_a \in \mathbb{R}^{H\times W\times C}$ 表示卷积神经网络输出的灾前和灾后图像的一对特征，其被用来计算特征的差异。

$$\mathbf{X}_g = \mathbf{X}_a - \mathbf{X}_b, \tag{6.6}$$

其中，$\mathbf{X}_g$ 是指双时序特征间的全局差异。为了提高运算的效率，$\mathbf{X}_g$ 被输入到一个卷积层来减少特征图的空间维度和通道数量，其输出特征图为 $\mathbf{X}'_g \in \mathbb{R}^{H'\times W'\times C'}$。然后将 $\mathbf{X}'_g$ 输入到两个 Transformer 层中，每层 Transformer 包含层归一化（Layer Normalization，LN）、多头自注意力（Multi-Head Self-Attention，MSA）和多层感知机（Multi-Layer Perceptron，MLP）等运算操作。

由于 Transformer 层的输入需要是一个特征序列，因此将 $\mathbf{X}'_g$ 展平并输入到一个可以被训练的线性层从而获得特征嵌入序列 $e_g \in \mathbb{R}^{L\times C_g}$，其中 C_g 是隐藏通道的维数，L 是序列的长度。同时本章学习了特定的位置嵌入序列 p_g 来编码缺失的图像空间信息，并将其与特征嵌入序列 e_g 相加来形成最终的输入序列 $E_g = e_g + p_g$。这样的操作能够保留图像的空间信息，即使 Transformer 的自注意力本质是无序的。

在单个自注意力模块中，输入的特征序列 $E_g \in \mathbb{R}^{L\times C_g}$ 被线性变换为三个部

分，即查询（Query）$Q \in \mathbb{R}^{L \times d_k}$，键（Key）$K \in \mathbb{R}^{L \times d_k}$，以及值（Value）$V \in \mathbb{R}^{L \times d_v}$，其中 d_k、d_v 分别是查询 Query（键 Key）和值 Value 的维数。线性变换的定义为：

$$\text{Query } Q = E_g \mathbf{w}_Q, \quad \text{Key } K = E_g \mathbf{w}_K, \quad \text{Value } V = E_g \mathbf{w}_V, \tag{6.7}$$

其中，$\mathbf{w}_Q$、$\mathbf{w}_K$ 和 $\mathbf{w}_V$ 分别是三个线性变换函数的权重矩阵。Q、K、V 都源于输入特征序列 E_g 本身，Q 和 K 是用来计算注意力权重矩阵的特征向量，而 V 的引入是为了保留输入的特征序列。

然后将尺度点乘注意力（Scaled Dot-Product Attention）应用于 Q、K 和 V，其定义为：

$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V, \tag{6.8}$$

其中，'softmax' 是指 softmax 函数，$\sqrt{d_k}$ 是一个缩放因子。$\text{softmax}(QK^T/\sqrt{d_k})$ 被用于计算注意力权重矩阵，两个不同的特征向量 Q 和 K 的引入能够提高注意力权重矩阵的泛化能力，然后利用注意力权重矩阵将 V 映射到一个新的空间中。

多头自注意力是包含多个独立自注意力模块的拓展，其将查询 Query、键 Key 和值 Value 进行多次拆分，并行计算公式 (6.8) 中的注意力函数，之后将所有头的输出进行拼接并映射到最终的输出。多头自注意力使得模型能够联合不同位置的不同表征子空间的信息[146]，其定义为：

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1,\ldots,\text{head}_h)\mathbf{W}^O, \tag{6.9}$$

$$\text{head}_i = \text{Attention}(Q\mathbf{W}_i^Q, K\mathbf{W}_i^K, V\mathbf{W}_i^V), \tag{6.10}$$

其中，h 为多头自注意力中的多头数量。$\mathbf{W}_i^Q$、$\mathbf{W}_i^K$ 和 $\mathbf{W}_i^V$ 是第 i 个头的三个线性变换矩阵，$\mathbf{W}^O$ 为将拼接结果映射到最终输出的转换矩阵。

多头自注意力模块的输出在经过层归一化后，被输入到多层感知机中进行特征转换。此外，如图 6.2（a）所示，在多头自注意力和多层感知机的输出中都应用到了残差连接[137]。为了更清晰的描述 Transformer 的内部结构，本章将第 l 层 Transformer 的输出以公式的形式定义为：

$$Z_l = \text{MLP}(\text{LN}(Z_l')) + Z_l', \tag{6.11}$$

其中，

$$Z_l' = \text{MSA}(\text{LN}(Z_{l-1})) + Z_{l-1}, \tag{6.12}$$

$\{Z_1, Z_2, \ldots, Z_l\}$ 表示 Transformer 层的特征。LN$(\cdot)$ 是指层归一化操作，其被用于实现更快的收敛和更加稳定的训练。

6.2.3 局部门控注意力模块

自然灾害的发生通常会对建筑物造成不同程度的损毁，评估损毁的严重性是开展灾害响应和人道主义援助的先决条件，而学习双时序图像之间的局部依赖性能够帮助识别不同程度的建筑物损毁。本章设计了一种局部门控注意力模块（Local Gated Attention，LGA）来学习双时序遥感图像之间多级别变化的局部依赖性。

对于成对的灾害前后特征 $\mathbf{X}_b$、$\mathbf{X}_a \in \mathbb{R}^{H\times W\times C}$，首先将其分别划分为 $\frac{HW}{4^2}$ 个特征块。对于获得的两个时序的两组特征块，将其展平后输入到两个线性层中来获得特征嵌入序列。两组特征嵌入序列分别与相应的位置嵌入序列相加，以生成灾前特征序列 E_b 和灾后特征序列 E_a。之后 E_b 和 E_a 被输入到一层 Gated Transformer 和一层 Transformer 中，如图图 6.1所示。

在 Gated Transformer 层中，参考门控轴向注意力[248]，本章提出了一种局部门控注意力来替代原本的多头自注意力，如图 6.2（b）所示。输入的特征序列 E_b 和 E_a 通过线性变换，生成灾前特征序列的 Q_b、K_b、V_b 和灾后特征序列的 Q_a、K_a、V_a。局部门控注意力模块利用门控加和操作来融合 Q_b、K_b、V_b 和 Q_a、K_a、V_a，之后将它们输入到多头自注意力中。与公式 (6.8) 不同，本章的局部门控注意力被定义为：

$$\mathrm{LGA}(Q', K', V') = \mathrm{softmax}(\frac{Q'K'^T}{\sqrt{d_k}})V', \tag{6.13}$$

其中，$Q' = G_b^q Q_b + G_a^q Q_a$，$K' = G_b^k K_b + G_a^k K_a$，$V' = G_b^v V_b + G_a^v V_a$。$G_b^q$、$G_a^q$、$G_b^k$、$G_a^k$、$G_b^v$、$G_a^v$ 均为可以学习的参数，这些参数构建了一种门控机制，来控制学习到的灾害前后特征块的相应特征序列。如果一个时序的特征序列对识别多级别变化更有帮助，门控机制会为其分配较高的权重，同时为另一个时序的特征序列分配比较低的权重。利用这一门控机制，局部门控注意力模块能够更加关注灾害前后特征块中对变化敏感的特征，进而将变化敏感的特征序列输入到多头自注意力中。通过自注意力机制，能够学习相邻变化敏感特征块之间的局部依赖性，进而获取双时序遥感图像之间的局部变化差异，并对所识别出的多级别变化进一步细化。

第三节 实验结果与分析

本小节将介绍本章实验中所使用的公开数据集和评测指标。同时与其他的变化检测方法进行了详细的定量和定性的对比，还进行了大量的消融实验来验证所提出模型和各个模块在变化检测任务中的有效性。

6.3.1 数据集与评测指标

与章 5.3.1一样，本章也采用了 xBD 数据集[7]，其是目前最大的建筑物损毁变化检测数据集。本小节选取 xBD 数据集中包含 2799 对遥感图像的训练集用于本章实验的训练，同时选取包含 933 对图像的测试集用来测试。

本章采用了文献[7]中介绍的 xView2 比赛的评测指标，来验证本章所提出的基于全局差异与局部注意力的网络在建筑物分割和多级别变化检测两个任务上的性能表现。评测指标 S_{xView2} 的详细介绍见章 5.3.2。

6.3.2 实现细节

在本章所提出的基于全局差异与局部注意力的变化检测方法中，设置批大小（Batch Size）为 16，并采用随机梯度下降（SGD）优化器。权重衰减系数（Weight Decay Coefficients）和动量系数（Momentum）分别被设置为 5×10^{-4} 和 0.9。设置初始学习率设置为 0.02，训练的 Epoch 为 150 个。此外，本章在训练中采用了 “poly” 学习率策略，通过乘以 $(1 - \frac{iter}{maxiter})^{power}$ 来降低学习率，其中 $power = 3$。本章中的方法使用了深度学习框架 PyTorch[223]来实现，所有的实验是在一台具有 4 个 16GB 显存的 NVIDIA Tesla V100 GPU 的服务器上执行的。在训练期间，本章应用了随机缩放、水平翻转和高斯模糊等数据增强技术来提高本模型的泛化能力。最终用于训练的图像大小为 512×512。

6.3.3 与现有方法的对比

本小节介绍了本章所提出的基于全局差异与局部注意力的变化检测模型（GDLA）与其他方法在 xBD 数据集[7]上的结果比较，包括 Baseline[7]、Siamese-UNet(ResNext50)[123]、Siamese-UNet(DPN92)[123]、Dual-HRNet[244]、Dual-Temporal Fusion[138]和 RescueNet[245]。为了公平客观的进行比较，本章重新实现了这些方法，并在同一数据集上开展了实验。

表 6.1 本章所提出的基于全局差异与局部注意力的变化检测模型（GDLA）与其他方法在 xBD 数据集[7]上的定量比较（%）。‘总体 F1’ 代表总体的 F1 值，即公式 (5.11) 中的评测指标 S_{xView2}。‘分割 F1’ 和 ‘变化 F1’ 分别指建筑物分割和变化检测的 F1 值；‘无 F1’、‘轻微 F1’、‘严重 F1’ 和 ‘完全 F1’ 分别指无损毁、轻微损毁、严重损毁和完全损毁的变化检测 F1 值。

方法	总体 F1	分割 F1	变化 F1	无 F1	轻微 F1	严重 F1	完全 F1
Baseline[7]	28.41	80.48	6.09	65.79	7.08	2.16	26.40
Siamese-UNet(ResNext50)[123]	67.33	79.81	61.97	76.86	45.39	64.17	71.86
Siamese-UNet(DPN92)[123]	69.18	83.56	63.02	81.55	43.90	66.21	75.01
Dual-HRNet[244]	71.35	83.61	66.09	86.43	48.66	69.11	71.80
Dual-Temporal Fusion[138]	72.52	82.76	68.13	86.29	50.77	68.71	**77.71**
RescueNet[245]	70.23	84.09	63.94	86.09	45.72	62.76	76.15
GDLA	**73.65**	**84.14**	**69.16**	**88.42**	**51.21**	**71.50**	76.84

表 6.2 本章所提出的基于全局差异与局部注意力的变化检测模型（GDLA）与其他方法在网络参数量、浮点运算数（Floating Point Operations，FLOPs）和运行时间方面的比较分析。‘参数数量’ 是指网络模型的参数量，浮点运算数（FLOPs）的输入图像尺寸为 512×512，‘推理时间’ 表示一张图像的推理时间。

方法	参数数量（M）	浮点运算数（G）	推理时间（s/张）
Baseline[7]	44.2	224.9	0.039
Siamese-UNet(ResNext50)[123]	69.1	142.0	0.068
Siamese-UNet(DPN92)[123]	94.8	243.6	0.437
Dual-HRNet[244]	59.5	91.3	0.055
Dual-Temporal Fusion[138]	43.9	201.0	0.035
RescueNet[245]	40.4	182.9	0.043
GDLA	33.5	71.7	0.022

6.3.3.1 定量结果比较

表 6.1中展示的是本章方法 GDLA 与其他方法的定量化比较结果。可以看出，本章所提出的基于全局差异与局部注意力的变化检测模型在变化检测 F1 值上取得了 69.16%，相比 Dual-Temporal Fusion[138] 高了 1.03%。此外，本章方法 GDLA 在三个损毁等级上都取得了最好的性能，例如在无损毁 F1 值上达到了 88.42%，在轻微损毁 F1 值上达到了 51.21%，分别比方法 Dual-HRNet[244]高了 1.99% 和 2.55%。对于建筑分割任务的 F1 值上，GDLA 优于 Dual-HRNet[244] 0.53%，同时比 Dual-Temporal Fusion[138]高了 1.38%。值得注意的是，本章方法 GDLA 在变化检测任务中高效地获得了比目前广泛使用的孪生网络更好的结果，这使其更适合应用于自然灾害评估中。

本小节也分析了本章方法 GDLA 与其他变化检测方法的网络参数量、浮点

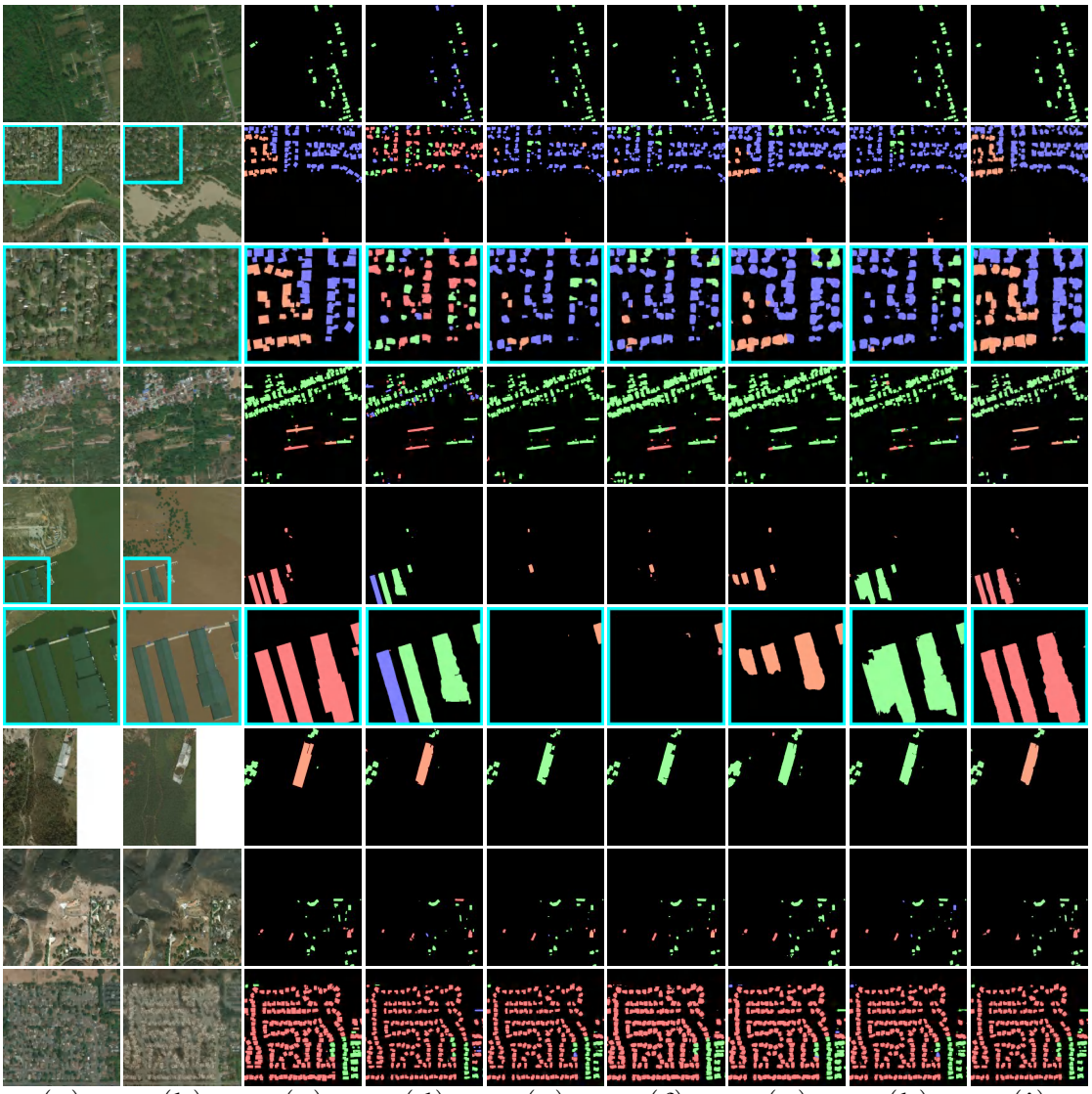


图 6.3 本章所提出的基于全局差异与局部注意力的变化检测模型（GDLA）与其他方法在 xBD 数据集[7]上的可视化比较。第一行到第九行分别展示了不同自然灾害的结果，包括飓风-佛罗伦萨、飓风-哈维、飓风-哈维的放大图、海啸、洪水、洪水的放大图、地震、火灾和野火。有青色外框的图片为放大图。（a）灾前遥感图像；（b）灾后遥感图像；（c）多级别变化的真值，其中的四种颜色代表四种损毁等级：绿色、蓝色、橙红色和红色分别表示无损毁、轻微损毁、严重损毁和完全损毁；（d）-（i）Baseline[7]、Siamese-UNet(ResNext50)[123]、Siamese-UNet(DPN92)[123]、Dual-HRNet[244]、Dual-Temporal Fusion[138]和本章方法 GDLA 的可视化结果。

运算数（Floating Point Operations，FLOPs）以及运行时间，如表 6.2中所示。其中浮点运算数 FLOPs 可以衡量一个网络模型的复杂度，单位为 G，1G FLOPs = 10^9FLOPs。为了公平的进行比较，表 6.2中列出了对大小为 512×512 输入

图像的浮点运算数，以及对一张 1024 × 1024 图像的推理时间，同时所有方法的实验都是在具有 4 个 NVIDIA Tesla V100 GPU 的服务器上执行的。从表中可以看出，本章方法 GDLA 的参数量和浮点运算数都相对较少，参数量比 Dual-Temporal Fusion[138]少了 10.4M，比 Dual-HRNet[244]少了 26.0M。值得注意的是，Siamese-UNet 的参数量是所有方法中最多的，因为其采用了两个网络来进行建筑分割任务和变化检测任务。此外，GDLA 相比于其他方法也获得了最快的推理时间。综上所述，本章方法 GDLA 的参数量更少、计算复杂度更小而且推理速度更快，同时又能获得更好的结果，因此更有利于实际中的变化检测应用。

根据上述实验能够看到，孪生网络引入了比较多的参数量，但是对于建筑分割和多级别变化检测的结果却相对较差，为了探究其中可能的原因，本小节开展了一系列的实验。如果将本章方法中混合卷积神经网络和 Transformer 结构的编码器替换到 Siamese-UNet 中，即将两个任务分成两个阶段来完成，可以获得结果：总体 F1 值为 72.78%、建筑物分割 F1 值为 83.65%、变化检测 F1 值为 68.12%，这个结果相比本章方法 GDLA 要低。如果将 Siamese-UNet 的结构改为单阶段，即利用孪生网络提取灾害前后图像的特征，双时序图像的特征在融合后，利用一个解码器同时输出建筑物分割和多级别变化检测的结果。Siamese-UNet(ResNext50) 的结果：总体 F1 值为 69.06%、建筑分割 F1 值为 82.91%、变化检测 F1 值为 63.13%。Siamese-UNet(DPN92) 的结果：总体 F1 值为 70.08%、建筑分割 F1 值为 83.60%、变化检测 F1 值为 64.28%。从这些结果中可以看出，相比于两个阶段的结构，单阶段的网络结构（即将两个任务同时完成）获得了更好的性能。从而可以分析出，双阶段孪生网络结果较低的原因是将两个任务分别独自进行训练，使模型无法从多任务学习中受益。由于建筑分割和多级别变化检测两个任务中包含的知识是相似的而且能够共享，多任务学习可以有效提高这两个任务的性能[249]。

6.3.3.2 可视化结果比较

本章方法 GDLA 与其他方法在 xBD 数据集上的可视化结果如 图 6.3所示。其中分析了来自七个不同自然灾害的示例，包括飓风-佛罗伦萨、飓风-哈维、海啸、洪水、地震、火灾和野火。为了对可视化结果进行更加清晰的比较，本章将飓风-哈维和洪水两个示例的部分区域进行了放大展示，图 6.3中有青色外框的图片即为相应区域的放大图。从可视化结果中可以看出，本章方法 GDLA 的

表 6.3 本章所提出的全局差异模块（Global Difference，GD）和局部门控注意力模块（Local Gated Attention，LGA）的消融实验（%）。编号 1 是以 ResNet-50 作为编码器，采用渐进式上采样结构作为解码器的基准网络；在基准网络的基础上，添加全局差异模块和局部门控注意力模块来验证这两个模块的有效性（编号 2 和编号 3）；编号 4 是本章方法 GDLA。'时间' 表示一张图像的推理时间（s/张）。

编号	GD	LGA	时间	总体 F1	分割 F1	变化 F1	无 F1	轻微 F1	严重 F1	完全 F1
1			0.015	69.38	79.89	64.88	84.86	46.11	68.49	73.65
2	✔		0.019	71.08	82.42	66.22	86.37	48.83	67.13	74.32
3		✔	0.018	71.67	**84.19**	66.31	86.97	47.93	68.53	74.79
4	✔	✔	0.022	**73.65**	84.14	**69.16**	**88.42**	**51.21**	**71.50**	**76.84**

建筑物分割结果更加准确，同时多级别变化检测的结果也与真值更加一致。例如，在图 6.3的第五行示例以及其第六行的放大图中，有几座建筑物被洪水摧毁。其他的方法无法分割出完整的建筑物（例如 Siamese-UNet(ResNext50)[123]、Siamese-UNet(DPN92)[123]和 Dual-HRNet[244]），或者不能准确地识别出变化等级（例如 Baseline[7]和 Dual-Temporal Fusion[138]），而本章方法 GDLA 获得了更好的结果，与其他方法相比更加接近真值。上述这些在 xBD 数据集上的可视化结果，进一步验证了本章方法在建筑分割和变化检测两个任务中的有效性。

6.3.4 消融实验

6.3.4.1 模块的有效性

在基于全局差异与局部注意力的变化检测模型（GDLA）中，本章提出了一个全局差异模块（Global Difference，GD）来学习全局变化模式，同时设计了一个局部门控注意力模块（Local Gated Attention，LGA）以获取多级别变化之间的局部依赖关系。如表 6.3中所列，本章对所提出的两个模块进行了消融实验来验证其有效性。编号 1 是以 ResNet-50 作为编码器，以渐进式上采样结构作为解码器的基准网络。将本章所提出的全局差异模块和局部门控注意力模块添加到基准网络后，变化检测 F1 值从 64.88% 分别提高到 66.32% 和 66.31%，建筑物分割 F1 值分别提高了 2.53% 和 4.30%。结合这两个模块后，相比于编号 1 的基准网络，变化检测 F1 值提高了 4.28%，总体 F1 值提高了 4.27%。上述这些实验结果验证了本章所提出的两个模块都可以提高对灾前和灾后图像中不同变化的判别能力。此外，本章也分析了在不同模块的配置下，网络模型对于一张 1024×1024 大小图像的推理时间。可以看到不同配置下模型的推理时间比较接近，编号 1 的基准网络模型较为简单，推理时间会更快。

表 6.4 对 Transformer 层数 l 的消融实验（%）。本章将全局差异模块和局部门控注意力模块中 Transformer 的层数设置为相同数量。

编号	层数 l	总体 F1	分割 F1	变化 F1	无 F1	轻微 F1	严重 F1	完全 F1
1	1	69.47	79.89	65.00	84.82	46.07	67.82	75.21
2	2	**73.65**	**84.14**	**69.16**	88.42	**51.21**	**71.50**	**76.84**
3	3	71.03	79.22	67.52	**90.31**	47.84	70.45	76.51

表 6.5 对 Transformer 中多头自注意力的多头（Multi-Head）数量 h 的消融实验（%）。本章将全局差异模块和局部门控注意力模块中多头自注意力的多头数量设置为相同。

编号	多头数量 h	总体 F1	分割 F1	变化 F1	无 F1	轻微 F1	严重 F1	完全 F1
1	2	71.26	81.27	66.97	87.64	48.58	70.01	74.37
2	4	**73.65**	**84.14**	**69.16**	88.42	**51.21**	**71.50**	76.84
3	8	72.20	84.12	67.09	**88.63**	48.70	66.53	**78.26**

表 6.6 对局部门控注意力模块中不同尺寸大小特征块的消融实验（%）。

编号	特征块尺寸	总体 F1	分割 F1	变化 F1	无 F1	轻微 F1	严重 F1	完全 F1
1	2×2	71.11	79.22	67.64	**89.27**	48.37	70.09	**76.94**
2	4×4	**73.65**	**84.14**	**69.16**	88.42	**51.21**	**71.50**	76.84
3	8×8	72.47	81.85	68.44	88.08	50.41	71.34	75.54

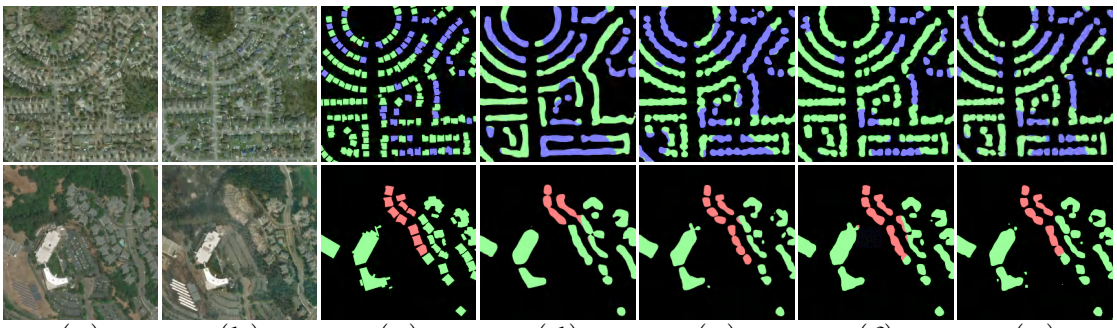
本章方法 GDLA 在不同模块配置下的可视化结果如 图 6.4所示。从中可以看出，基准网络对于建筑物分割和多级别变化的识别都相对较差。在基准网络基础上添加了全局差异模块后，能够更好地定位不同变化的位置，但是在一些局部区域不同变化等级的识别上会出现效果较差的情况。局部门控注意力模块能够帮助局部区域内不同变化等级的识别，在一些小区域变化等级的检测上性能较好，但是定位变化位置的能力比较差，会出现错误检测大面积区域变化等级的情况。结合这两个模块后，可以有效地缓解使用单个模块所出现的问题，获得更好的建筑分割和变化检测结果。

6.3.4.2 Transformer 中不同配置的影响

本章开展了多个消融实验来分析 Transformer 层中不同参数设置对实验结果的影响，包括 Transformer 的层数 l 和多头自注意力中的多头（Multi-Head）数量 h。对 Transformer 层数 l 的分析如表 6.4中所示，当 Transformer 层数 $l=2$ 时，总体 F1 值为 73.65%，比另外两个设置 $l=1$ 和 $l=3$ 分别提高了 4.18% 和 2.62%。表 6.5中是对多头自注意力的多头数量 h 的分析结果。可以看出，当多

表 6.7 对建筑物分割和变化检测两个任务损失函数组合的消融实验（%）。其中 'Combo Loss'[242]是 Dice 损失函数[243]和 Focal 损失函数[52]的加权和，'CE Loss' 是指交叉熵损失。

编号	建筑分割	变化检测	总体 F1	分割 F1	变化 F1
1	CE Loss	CE Loss	71.46	81.51	67.15
2	Combo Loss	Combo Loss	72.57	82.78	68.19
3	CE Loss	Combo Loss	71.99	82.11	67.66
4	Combo Loss	CE Loss	**73.65**	**84.14**	**69.16**



（a） （b） （c） （d） （e） （f） （g）

图 6.4 本章方法 GDLA 在不同模型配置下的可视化结果。第一二行分别是来自灾害飓风和野火的示例。（a）灾前遥感图像；（b）灾后遥感图像；（c）多级别变化的真值；（d）表 6.3中编号 1 基准网络的可视化结果；（e）表 6.3中编号 2 的可视化结果；（f）表 6.3中编号 3 的可视化结果；（g）表 6.3中编号 4，即本章方法 GDLA 的可视化结果。

头数量 $h=4$ 时结果是最好的，总体 F1 值相比于多头数量 $h=2$ 和 $h=8$ 分别提高了 2.39% 和 1.45%。此外，本章也对局部门控注意力模块中不同尺寸大小特征块进行了消融实验，结果如表 6.6所示。可以看出，4×4 是局部门控注意力模块中特征块尺寸的最佳设置，而特征块越小或者越大都会限制性能的提高。在 Transformer 中，特征嵌入序列的长度与特征块尺寸的平方成反比。减小特征块的尺寸，Transformer 层能够使用更长的特征序列来编码复杂的长程依赖关系，因而可以提高性能。然而太小的特征块可能会阻碍网络学习不同变化之间的相似性，进而影响变化检测的结果。

根据上述实验，本章设置 Transformer 层数 $l=2$，设置多头自注意力中的多头数量 $h=4$，并使用 4×4 作为本章方法中的特征块尺寸。

6.3.4.3 两个任务不同损失函数组合的影响

本章开展了消融实验来对建筑物分割和变化检测两个任务损失函数的选择进行分析，结果如表 6.7所示本章选择了目前被广泛使用的交叉熵损失函数和能够较好的处理类别不均衡问题的 Combo 损失函数[242]，其中 Combo 损失函数是 Dice 损失[243] 和 Focal 损失[52]的加权和。通过分析这两个损失函数对于两个任务的几种组合，可以得出，对建筑分割任务采用 Combo 损失函数，同时对变化检测任务采用加权交叉熵损失时，本章所提出的模型可以达到最优的性能。因此本章针对这两个具体任务选择了目前所用的损失函数。

第四节 本章小结

本章提出了一种基于全局差异与局部注意力的变化检测模型（Global Difference and Local Attention，GDLA），基于双时序高分辨率遥感图像同时进行建筑物分割和多级别变化检测两个任务。结合卷积神经网络擅长学习局部细节特征和 Transformer 可以建模长程依赖关系的优势, 本章采用混合卷积神经网络和 Transformer 的结构作为编码器，同时采用渐进式上采样结构作为解码器，来输出建筑分割和多级别变化检测的结果。一张遥感图像通常会覆盖较大的区域，其中包含的建筑在灾害发生后会受到不同程度的损毁。本章设计了一个全局差异模块（Global Difference，GD），来学习全局变化模式并提高对双时序图像之间变化的整体认识。同时提出了一个局部门控注意力模块（Local Gated Attention，LGA），来获取多级别变化之间的局部依赖性并增强对双时序图像之间不同变化的判别能力。在 xBD 数据集上的大量实验验证了本章方法的优越性，同时消融实验也证明了所提出的两个模块的有效性。

第七章 总结与展望

本章针对本文的主要研究成果和创新点进行了总结，并对未来可能的研究进行了展望。

第一节 本文工作总结

序列图像是一组按照特定时空域顺序排列的图像集合，其中不仅蕴含图像内部的语义信息，图像之间也存在着语义依赖关系（即序列图像的“序列性”）。然而近年来的研究大多忽略了序列图像中图像之间的语义关联信息，也不能准确地描述序列图像的复杂模式。本文以序列图像的识别与检测算法为研究目标，将序列图像展开为立体空间序列图像、平面空间序列图像和时间序列图像分别进行研究，致力于探索如何高效地挖掘图像内部及图像之间的深层次语义关联信息。面向不同维度分布的序列图像，本文从肺结节检测、道路提取和变化检测等三个具体的计算机视觉任务出发，通过改进视觉注意力机制，建立了具有更强的序列关联性表达能力的深度网络模型，并实现了更好的识别与检测性能。具体而言，本文的主要研究工作可以总结为：

（1）对于立体空间序列图像中的肺部 CT 图像，针对其立体空间维度较高的难点，本文提出了一种切片关联注意力网络进行肺结节的检测。受到医生临床诊断肺结节方式的启发，本文设计了一种切片分组非局部模块，将切片维度分组的思想引入自注意力机制中，来充分学习 CT 图像中的立体空间序列信息，同时有效的降低计算复杂度。三维区域候选网络对肺结节的检测具有较高的灵敏度，但其通常会带来大量的假阳性样例，本文设计了基于多尺度特征图的假阳性抑制模块来进一步优化检测的结果。此外，本文提出了新的肺结节检测数据集 PN9，与之前的数据集相比，其在数据规模、种类多样性、图像丰富度和检测困难程度上都有了较大的提高。通过在不同数据集上的大量实验，充分验证了本文所提出的切片关联注意力网络能够有效地提高肺结节检测的性能。

（2）对于平面空间序列图像中的道路遥感图像，如何缓解其他地物遮挡或复杂交通场景的影响，并利用道路的拓扑序列信息来保证所识别道路的连通性，是一个难点。本文提出了一种拓扑连通注意力网络，其能够直接从遥感影像中

提取出连通性较好的道路。考虑到道路在平面维度上连续分布并呈现出跨度大和细长的形状，本文设计了一种条形卷积模块，利用水平、垂直、左对角线和右对角线等四种不同方向的条形卷积来学习道路的长距离依赖信息，同时抑制不相关区域对特征学习的干扰。此外设计了一种连通性注意力模块来探索相邻像素之间的连通关系，其能够缓解建筑物或树木等对道路的遮挡问题，提高道路的拓扑正确性。通过在两个公开数据集上的大量实验，验证了本文所提出的拓扑连通注意力网络在保证道路连通性方面的有效性。

（3）针对时间序列图像中的双时序高分辨率遥感图像，本文提出了差异感知注意力网络来同时进行建筑物分割和多级别变化检测。灾前和灾后图像特征之间不同的通道可能表达出不同的信息，为了探索其中能够反映出差异性变化模式的通道，本文设计了一种双时序聚合模块，其能够同时学习全局变化信息。此外，考虑到图像中存在的不同级别的变化，本文进一步设计了一种差异注意力模块。通过将特征图划分重组，再利用自注意力来获取一个特征立方体组内任意位置和任意通道之间的依赖关系，其中每个组中的小特征立方体都有表示多级别变化的潜力。在大规模建筑物变化检测数据集上的大量实验表明，相比于其他变化检测方法，本文提出的差异感知注意力网络具有较大的优越性。

（4）基于时间序列图像中的双时序高分辨率遥感图像，本文进一步提出了基于全局差异与局部注意力的时间序列图像检测模型。结合卷积神经网络能够更好的提取图像低阶细节信息的特点和 Transformer 可以对长程依赖关系进行建模的能力，采用混合卷积神经网络和 Transformer 的架构作为编码器来提取图像特征。在视觉任务中的 Transformer 结构通常会采用图像块作为输入，但是图像块的尺寸要小于整幅图像，这会限制模型学习不同变化之间长程依赖关系的能力。本文设计了一种全局差异模块，来学习全局变化信息并提高对图像中所有像素的整体理解。此外，本文设计了一种局部门控注意力模块，来学习局部变化差异并增强对双时序图像间多级别变化的判别能力，其利用门控自注意力机制来学习相邻变化敏感特征块之间的局部依赖性。通过大量实验，验证了本文提出的基于全局差异与局部注意力的网络应对变化检测任务是有效的。

第二节　对未来工作的展望

本文的研究内容致力于探索具有更强序列关联性表达能力的网络模型来处理序列图像，立足于目前的研究成果，未来可进一步深入的研究工作包括以下

几点：

（1）**针对立体空间序列图像检测的轻量级网络模型研究**。目前基于立体空间序列图像中 CT 图像的肺结节检测研究，大多采用了三维卷积神经网络来提取 CT 图像的特征，进而生成三维的候选框。相比于二维卷积神经网络，利用三维卷积神经网络能够直接生成三维的肺结节检测结果，且检测性能也获得了较大的提高。但同时也带来了更多的网络参数量，导致其需要较长的时间和较多的 GPU 显存来进行训练，难以在临床诊断环境中快速部署和应用。在当前数据集规模越来越大且深度学习能够有效辅助医生诊断的情况下，设计一个参数量更少、能够快速识别肺结节且性能较好的网络模型是非常有必要的。因此，用于肺结节检测的参数量少且保证性能的轻量级网络，是未来一个重要且有挑战性的研究工作。

（2）**针对序列图像的无监督或弱监督识别与检测算法的研究**。相比于零散分布的自然图像，序列图像通常会有更高的维度和更大的数据规模，例如空间序列图像中的 CT 图像有更高的切片维度，而时间序列图像在时间维度上存在大量的图片。这导致序列图像的获取和标注也更加困难，会成倍增加所消耗的人力和物力。无监督或弱监督的识别与检测算法，不需要或者仅需要很少的标注数据，就可以使神经网络有效地进行学习。对于一些数据较难获取的任务，例如涉及到患者隐私的医学图像等，无监督或弱监督的算法是未来主要的研究方向，然而这些算法目前的性能与监督学习相比相差较多。因此，针对序列图像的无监督或弱监督识别与检测算法的研究，成为未来一个值得关注的研究方向。

（3）**针对长时间序列图像的识别与检测算法的研究**。本文针对时间序列图像中的双时序高分辨率遥感图像，进行了建筑物分割和多级别变化检测的研究。然而，所选取的双时序遥感图像，只有两个不同时期的两张图像，不能充分体现时间序列图像中的时间关联信息。针对长时间序列图像中的视频，可以开展基于视频的语义分割、目标检测以及目标跟踪等，这些任务都需要充分探索目标的时间维度依赖信息。因此，接下来的研究工作会致力于视频，对其中的深层次时间序列信息进行探索。

参考文献

[1] 王昌淼. 基于胸部影像的肺结节检测与分类关键技术研究[D]. 中国科学院大学(中国科学院深圳先进技术研究院), 2018.

[2] 张立强, 李洋, 侯正阳, 等. 深度学习与遥感数据分析[J]. 武汉大学学报 ● 信息科学版, 2020, 45(12): 1857-1864.

[3] 张哲. 基于视频序列的图像拼接关键技术研究[D]. 天津理工大学, 2021.

[4] 唐邓清. 无人机序列图像目标位姿估计方法及应用[D]. 国防科技大学, 2019.

[5] MEI J, CHENG M M, XU G, et al. SANet: A slice-aware network for pulmonary nodule detection[J]. IEEE transactions on pattern analysis and machine intelligence, 2022, 44(8): 4374-4387.

[6] MEI J, LI R J, GAO W, et al. CoANet: Connectivity Attention Network for Road Extraction From Satellite Imagery[J]. IEEE Transactions on Image Processing, 2021, 30: 8540-8552.

[7] GUPTA R, GOODMAN B, PATEL N, et al. Creating xBD: A dataset for assessing building damage from satellite imagery[C]//Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2019: 10-17.

[8] LI B, WU W, WANG Q, et al. Siamrpn++: Evolution of siamese visual tracking with very deep networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 4282-4291.

[9] LOWE D G. Object recognition from local scale-invariant features[C]//Proceedings of the seventh IEEE international conference on computer vision: vol. 2. 1999: 1150-1157.

[10] OJALA T, PIETIKAINEN M, MAENPAA T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns[J]. IEEE Transactions on pattern analysis and machine intelligence, 2002, 24(7): 971-987.

[11] PITAS I. Digital image processing algorithms and applications[M]. John Wiley & Sons, 2000.

[12] 姬晓鹏. 基于深度图像序列的人体动作识别方法研究[D]. 中国科学院大学 (中国科学院深圳先进技术研究院), 2018.

[13] 冯诚, 张聪炫, 陈震, 等. 基于光流与多尺度上下文的图像序列运动遮挡检测[J]. 自动化学报, 2021: 1001-1012.

[14] GEORGE J, SKARIA S, VARUN V, et al. Using YOLO based deep learning network for real time detection and localization of lung nodules from low dose CT scans[C]//

Medical Imaging 2018: Computer-Aided Diagnosis: vol. 10575. 2018: 347-355.

[15] XIE H, YANG D, SUN N, et al. Automated pulmonary nodule detection in CT images using deep convolutional neural networks[J]. Pattern Recognition, 2019, 85: 109-119.

[16] TREISMAN A M, GELADE G. A feature-integration theory of attention[J]. Cognitive psychology, 1980, 12(1): 97-136.

[17] STYLES E. The psychology of attention[M]. Psychology Press, 2006.

[18] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7794-7803.

[19] GAO S, ZHOU C, MA C, et al. Aiatrack: Attention in attention for transformer visual tracking[C]//European Conference on Computer Vision. 2022: 146-164.

[20] XIA Z, PAN X, SONG S, et al. Vision transformer with deformable attention[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 4794-4803.

[21] CHENG B, MISRA I, SCHWING A G, et al. Masked-attention mask transformer for universal image segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 1290-1299.

[22] 王军. 基于胸部 CT 的肺结节计算机自动检测与人工智能辅助诊断关键技术研究[D]. 浙江大学, 2019.

[23] MACMAHON H, NAIDICH D P, GOO J M, et al. Guidelines for management of incidental pulmonary nodules detected on CT images: from the Fleischner Society 2017 [J]. Radiology, 2017, 284(1): 228-243.

[24] FIELD J K, OUDKERK M, PEDERSEN J H, et al. Prospects for population screening and diagnosis of lung cancer[J]. The Lancet, 2013, 382(9893): 732-741.

[25] MESSAY T, HARDIE R C, ROGERS S K. A new computationally efficient CAD system for pulmonary nodule detection in CT imagery[J]. Medical image analysis, 2010, 14(3): 390-406.

[26] JACOBS C, van RIKXOORT E M, TWELLMANN T, et al. Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images[J]. Medical image analysis, 2014, 18(2): 374-384.

[27] DUGGAN N, BAE E, SHEN S, et al. A technique for lung nodule candidate detection in CT using global minimization methods[C]//International workshop on energy minimization methods in computer vision and pattern recognition. 2015: 478-491.

[28] LOPEZ TORRES E, FIORINA E, PENNAZIO F, et al. Large scale validation of the M5L lung CAD on heterogeneous CT datasets[J]. Medical physics, 2015, 42(4): 1477-1489.

[29] GUPTA A, MARTENS O, LE MOULLEC Y, et al. Methods for increased sensitivity and scope in automatic segmentation and detection of lung nodules in CT images[C]// 2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT). 2015: 375-380.

[30] WANG B, TIAN X, WANG Q, et al. Pulmonary nodule detection in CT images based on shape constraint CV model[J]. Medical physics, 2015, 42(3): 1241-1254.

[31] CHAN T F, VESE L A. Active contours without edges[J]. IEEE Transactions on image processing, 2001, 10(2): 266-277.

[32] SILVA A C, de PAIVA A C, NUNES R A, et al. 3D shape analysis to reduce false positives for lung nodule detection systems[J]. Medical & biological engineering & computing, 2017, 55(8): 1199-1213.

[33] KHORDEHCHI E A, AYATOLLAHI A, DALIRI M R. Automatic lung nodule detection based on statistical region merging and support vector machines[J]. Image Analysis & Stereology, 2017, 36(2): 65-78.

[34] FROZ B R, de CARVALHO FILHO A O, SILVA A C, et al. Lung nodule classification using artificial crawlers, directional texture and support vector machine[J]. Expert Systems with Applications, 2017, 69: 176-188.

[35] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]//International Conference on Neural Information Processing Systems. 2015: 91-99.

[36] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot multibox detector[C]// European conference on computer vision. 2016: 21-37.

[37] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[38] SETIO A A A, CIOMPI F, LITJENS G, et al. Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks[J]. IEEE transactions on medical imaging, 2016, 35(5): 1160-1169.

[39] FU L, MA J, REN Y, et al. Automatic detection of lung nodules: false positive reduction using convolution neural networks and handcrafted features[C]//Medical Imaging 2017: Computer-Aided Diagnosis: vol. 10134. 2017: 60-67.

[40] JIANG H, MA H, QIAN W, et al. An automatic detection system of lung nodule based on multigroup patch-based deep learning network[J]. IEEE journal of biomedical and health informatics, 2017, 22(4): 1227-1237.

[41] DING J, LI A, HU Z, et al. Accurate pulmonary nodule detection in computed tomog-

raphy images using deep convolutional neural networks[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. 2017: 559-567.

[42] DOU Q, CHEN H, YU L, et al. Multilevel contextual 3-D CNNs for false positive reduction in pulmonary nodule detection[J]. IEEE Transactions on Biomedical Engineering, 2016, 64(7): 1558-1567.

[43] LI Y, FAN Y. DeepSEED: 3D squeeze-and-excitation encoder-decoder convolutional neural networks for pulmonary nodule detection[C]//2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). 2020: 1866-1869.

[44] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[45] ZHU W, LIU C, FAN W, et al. Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification[C]//IEEE Winter Conference on Applications of Computer Vision. 2018: 673-681.

[46] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. 2015: 234-241.

[47] LIAO F, LIANG M, LI Z, et al. Evaluate the malignancy of pulmonary nodules using the 3-d deep leaky noisy-or network[J]. IEEE transactions on neural networks and learning systems, 2019, 30(11): 3484-3495.

[48] KIM B C, YOON J S, CHOI J S, et al. Multi-scale gradual integration CNN for false positive reduction in pulmonary nodule detection[J]. Neural Networks, 2019, 115: 1-10.

[49] OZDEMIR O, RUSSELL R L, BERLIN A A. A 3D probabilistic deep learning system for detection and diagnosis of lung cancer using low-dose CT scans[J]. IEEE transactions on medical imaging, 2019, 39(5): 1419-1429.

[50] HARSONO I W, LIAWATIMENA S, CENGGORO T W. Lung nodule detection and classification from thorax ct-scan using retinanet with transfer learning[J]. Journal of King Saud University-Computer and Information Sciences, 2022, 34(3): 567-577.

[51] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? a new model and the kinetics dataset[C]//proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6299-6308.

[52] 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.

[53] SONG T, CHEN J, LUO X, et al. CPM-Net: A 3D Center-Points Matching Network for Pulmonary Nodule Detection in CT Scans[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. 2020: 550-559.

[54] CHEN Y, LI J, XIAO H, et al. Dual path networks[C]//International Conference on Neural Information Processing Systems. 2017: 4467-4475.

[55] KHOSRAVAN N, BAGCI U. S4ND: single-shot single-scale lung nodule detection[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. 2018: 794-802.

[56] YAN X, PANG J, QI H, et al. Classification of lung nodule malignancy risk on computed tomography images using convolutional neural network: A comparison between 2d and 3d strategies[C]//Asian Conference on Computer Vision. 2016: 91-101.

[57] ALSHEHHI R, MARPU P R. Hierarchical graph-based segmentation for extracting road networks from high-resolution satellite images[J]. ISPRS journal of photogrammetry and remote sensing, 2017, 126: 245-260.

[58] BASTANI F, HE S, ABBAR S, et al. Roadtracer: Automatic extraction of road networks from aerial images[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4720-4728.

[59] BATRA A, SINGH S, PANG G, et al. Improved road connectivity by joint learning of orientation and segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2019: 10385-10393.

[60] KONG H, AUDIBERT J Y, PONCE J. General road detection from a single image[J]. IEEE Transactions on Image Processing, 2010, 19(8): 2211-2220.

[61] SONG W, KELLER J M, HAITHCOAT T L, et al. Automated geospatial conflation of vector road maps to high resolution imagery[J]. IEEE Transactions on Image Processing, 2008, 18(2): 388-400.

[62] AMO M, MARTÍNEZ F, TORRE M. Road extraction from aerial images using a region competition algorithm[J]. IEEE Transactions on Image Processing, 2006, 15(5): 1192-1201.

[63] HE Y, WANG H, ZHANG B. Color-based road detection in urban traffic scenes[J]. IEEE Transactions on intelligent transportation systems, 2004, 5(4): 309-318.

[64] ZHANG Q, COULOIGNER I. Benefit of the angular texture signature for the separation of parking lots and roads on high resolution multi-spectral imagery[J]. Pattern recognition letters, 2006, 27(9): 937-946.

[65] LAPTEV I, MAYER H, LINDEBERG T, et al. Automatic extraction of roads from aerial images based on scale space and snakes[J]. Machine Vision and Applications, 2000, 12(1): 23-31.

[66] CHAI D, FORSTNER W, LAFARGE F. Recovering line-networks in images by junction-point processes[C]//Proceedings of the IEEE conference on computer vision

and pattern recognition. 2013: 1894-1901.

[67] WEGNER J D, MONTOYA-ZEGARRA J A, SCHINDLER K. A higher-order CRF model for road network extraction[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2013: 1698-1705.

[68] 曾发明, 杨波, 吴德文, 等. 基于 Canny 边缘检测算子的矿区道路提取[J]. 国土资源遥感, 2013(4): 72-78.

[69] GAETANO R, ZERUBIA J, SCARPA G, et al. Morphological road segmentation in urban areas from high resolution satellite images[C]//2011 17th International Conference on Digital Signal Processing (DSP). 2011: 1-8.

[70] STOICA R, DESCOMBES X, ZERUBIA J. A Gibbs point process for road extraction from remotely sensed images[J]. International Journal of Computer Vision, 2004, 57(2): 121-136.

[71] 滕鑫鹏, 宋顺林, 詹永照. 一种改进的基于结构张量的高分辨率遥感图像道路提取算法[J]. 科技通报, 2013, 29(2): 175-177.

[72] LI M, STEIN A, BIJKER W, et al. Region-based urban road extraction from VHR satellite images using binary partition tree[J]. International journal of applied earth observation and geoinformation, 2016, 44: 217-225.

[73] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.

[74] BADRINARAYANAN V, KENDALL A, CIPOLLA R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(12): 2481-2495.

[75] CHEN L C, ZHU Y, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//European conference on computer vision. 2018: 801-818.

[76] WU H, ZHANG J, HUANG K, et al. Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation[J]. arXiv preprint arXiv:1903.11816, 2019.

[77] TAKIKAWA T, ACUNA D, JAMPANI V, et al. Gated-scnn: Gated shape cnns for semantic segmentation[C]//Proceedings of the IEEE international conference on computer vision. 2019: 5229-5238.

[78] ALSHEHHI R, MARPU P R, WOON W L, et al. Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks[J]. ISPRS journal of photogrammetry and remote sensing, 2017, 130: 139-149.

[79] ZHANG Z, LIU Q, WANG Y. Road extraction by deep residual u-net[J]. IEEE Geo-

science and Remote Sensing Letters, 2018, 15(5): 749-753.

[80] MÁTTYUS G, URTASUN R. Matching adversarial networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8024-8032.

[81] MNIH V, HINTON G E. Learning to detect roads in high-resolution aerial images[C]//European conference on computer vision. 2010: 210-223.

[82] CHENG G, WANG Y, XU S, et al. Automatic road detection and centerline extraction via cascaded end-to-end convolutional neural network[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(6): 3322-3337.

[83] PANBOONYUEN T, JITKAJORNWANICH K, LAWAWIROJWONG S, et al. Road segmentation of remotely-sensed images using deep convolutional neural networks with landscape metrics and conditional random fields[J]. Remote Sensing, 2017, 9(7): 680.

[84] MENDES C C T, FRÉMONT V, WOLF D F. Exploiting fully convolutional neural networks for fast road detection[C]//2016 IEEE International Conference on Robotics and Automation (ICRA). 2016: 3174-3179.

[85] CHAURASIA A, CULURCIELLO E. Linknet: Exploiting encoder representations for efficient semantic segmentation[C]//IEEE Visual Communications and Image Processing. 2017: 1-4.

[86] ZHOU L, ZHANG C, WU M. D-LinkNet: LinkNet With Pretrained Encoder and Dilated Convolution for High Resolution Satellite Imagery Road Extraction.[C]//Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2018: 182-186.

[87] WEGNER J D, MONTOYA-ZEGARRA J A, SCHINDLER K. Road networks as collections of minimum cost paths[J]. ISPRS journal of photogrammetry and remote sensing, 2015, 108: 128-137.

[88] MÁTTYUS G, LUO W, URTASUN R. Deeproadmapper: Extracting road topology from aerial images[C]//Proceedings of the IEEE international conference on computer vision. 2017: 3438-3446.

[89] MOSINSKA A, MARQUEZ-NEILA P, KOZIŃSKI M, et al. Beyond the pixel-wise loss for topology-aware delineation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 3136-3145.

[90] LIU Y, YAO J, LU X, et al. Roadnet: Learning to comprehensively analyze road networks in complex urban scenes from high-resolution remotely sensed images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2018, 57(4): 2043-2056.

[91] LI X, WANG Y, ZHANG L, et al. Topology-enhanced urban road extraction via a geographic feature-enhanced network[J]. IEEE Transactions on Geoscience and Remote

Sensing, 2020, 58(12): 8819-8830.

[92] CHOI Y W, JANG Y W, LEE H J, et al. Three-dimensional LiDAR data classifying to extract road point in urban area[J]. IEEE Geoscience and Remote Sensing Letters, 2008, 5(4): 725-729.

[93] YADAV M, SINGH A K, LOHANI B. Extraction of road surface from mobile LiDAR data of complex road environment[J]. International journal of remote sensing, 2017, 38(16): 4655-4682.

[94] LIANG J, HOMAYOUNFAR N, MA W C, et al. Convolutional recurrent network for road boundary extraction[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2019: 9512-9521.

[95] HU J, RAZDAN A, FEMIANI J C, et al. Road network extraction and intersection detection from aerial images by tracking road footprints[J]. IEEE Transactions on Geoscience and Remote Sensing, 2007, 45(12): 4144-4157.

[96] BIAGIONI J, ERIKSSON J. Map inference in the face of noise and disparity[C]//Proceedings of the 20th International Conference on Advances in Geographic Information Systems. 2012: 79-88.

[97] SHAN Z, WU H, SUN W, et al. COBWEB: A robust map update system using GPS trajectories[C]//Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. 2015: 927-937.

[98] YUAN J, CHERIYADAT A M. Image feature based GPS trace filtering for road network generation and road segmentation[J]. Machine Vision and Applications, 2016, 27(1): 1-12.

[99] SUN T, DI Z, CHE P, et al. Leveraging crowdsourced gps data for road extraction from aerial imagery[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2019: 7509-7518.

[100] MAHDAVI S, SALEHI B, HUANG W, et al. A PolSAR change detection index based on neighborhood Information for flood mapping[J]. Remote Sensing, 2019, 11(16): 1854.

[101] CHEN C F, SON N T, CHANG N B, et al. Multi-decadal mangrove forest change detection and prediction in Honduras, Central America, with Landsat imagery and a Markov chain model[J]. Remote Sensing, 2013, 5(12): 6408-6426.

[102] MARIN C, BOVOLO F, BRUZZONE L. Building change detection in multitemporal very high resolution SAR images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2014, 53(5): 2664-2682.

[103] SONG C, HUANG B, KE L, et al. Remote sensing of alpine lake water environment

changes on the Tibetan Plateau and surroundings: A review[J]. ISPRS journal of photogrammetry and remote sensing, 2014, 92: 26-37.

[104] LU X, YUAN Y, ZHENG X. Joint dictionary learning for multispectral change detection[J]. IEEE transactions on cybernetics, 2016, 47(4): 884-897.

[105] SINGH A. Review article digital change detection techniques using remotely-sensed data[J]. International journal of remote sensing, 1989, 10(6): 989-1003.

[106] QUARMBY N, CUSHNIE J. Monitoring urban land cover changes at the urban fringe from SPOT HRV imagery in south-east England[J]. International Journal of Remote Sensing, 1989, 10(6): 953-963.

[107] HOWARTH P J, WICKWARE G M. Procedures for change detection using Landsat digital data[J]. International Journal of Remote Sensing, 1981, 2(3): 277-291.

[108] LUDEKE A K, MAGGIO R C, REID L M. An analysis of anthropogenic deforestation using logistic regression and GIS[J]. Journal of Environmental Management, 1990, 31(3): 247-259.

[109] DU P, WANG X, CHEN D, et al. An improved change detection approach using tri-temporal logic-verified change vector analysis[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 161: 278-293.

[110] DENG J, WANG K, DENG Y, et al. PCA-based land-use change detection and analysis using multitemporal and multisensor satellite data[J]. International Journal of Remote Sensing, 2008, 29(16): 4823-4838.

[111] ZHONG J, WANG R. Multi-temporal remote sensing change detection based on independent component analysis[J]. International Journal of Remote Sensing, 2006, 27(10): 2055-2061.

[112] NIELSEN A A, CONRADSEN K, SIMPSON J J. Multivariate alteration detection (MAD) and MAF postprocessing in multispectral, bitemporal image data: New approaches to change detection studies[J]. Remote Sensing of Environment, 1998, 64(1): 1-19.

[113] KASETKASEM T, VARSHNEY P K. An image change detection algorithm based on Markov random field models[J]. IEEE Transactions on Geoscience and Remote Sensing, 2002, 40(8): 1815-1823.

[114] GAPPER J J, EL-ASKARY H, LINSTEAD E, et al. Coral Reef change Detection in Remote Pacific islands using support vector machine classifiers[J]. Remote Sensing, 2019, 11(13): 1525.

[115] ZHONG P, WANG R. A multiple conditional random fields ensemble model for urban area detection in remote sensing optical images[J]. IEEE Transactions on Geoscience

and Remote Sensing, 2007, 45(12): 3978-3988.

[116] NEMMOUR H, CHIBANI Y. Multiple support vector machines for land cover change detection: An application for mapping urban extensions[J]. ISPRS journal of photogrammetry and remote sensing, 2006, 61(2): 125-133.

[117] IM J, JENSEN J R. A change detection model based on neighborhood correlation image analysis and decision tree classification[J]. Remote Sensing of Environment, 2005, 99(3): 326-340.

[118] ZHANG Y, PENG D, HUANG X. Object-based change detection for VHR images based on multiscale uncertainty analysis[J]. IEEE Geoscience and Remote Sensing Letters, 2017, 15(1): 13-17.

[119] TAN K, ZHANG Y, WANG X, et al. Object-based change detection using multiple classifiers and multi-scale uncertainty analysis[J]. Remote Sensing, 2019, 11(3): 359.

[120] CAYE DAUDT R, LE SAUX B, BOULCH A, et al. Guided anisotropic diffusion and iterative learning for weakly supervised change detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019: 1-10.

[121] PAPADOMANOLAKI M, VERMA S, VAKALOPOULOU M, et al. Detecting urban changes with recurrent neural networks from multitemporal Sentinel-2 data[C]//IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium. 2019: 214-217.

[122] WU C, CHEN H, DU B, et al. Unsupervised Change Detection in Multitemporal VHR Images Based on Deep Kernel PCA Convolutional Mapping Network[J]. IEEE transactions on cybernetics, 2022, 52(11): 12084-12098.

[123] DAUDT R C, LE SAUX B, BOULCH A. Fully convolutional siamese networks for change detection[C]//2018 25th IEEE International Conference on Image Processing (ICIP). 2018: 4063-4067.

[124] LIU J, GONG M, QIN K, et al. A deep convolutional coupling network for change detection based on heterogeneous optical and radar images[J]. IEEE transactions on neural networks and learning systems, 2016, 29(3): 545-559.

[125] WANG M, TAN K, JIA X, et al. A Deep Siamese Network with Hybrid Convolutional Feature Extraction Module for Change Detection Based on Multi-sensor Remote Sensing Images[J]. Remote Sensing, 2020, 12(2): 205.

[126] CHEN H, WU C, DU B, et al. Change detection in multisource VHR images via deep siamese convolutional multiple-layers recurrent neural network[J]. IEEE Transactions on Geoscience and Remote Sensing, 2019, 58(4): 2848-2864.

[127] DU B, RU L, WU C, et al. Unsupervised deep slow feature analysis for change detection in multi-temporal remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2019, 57(12): 9976-9992.

[128] LIU Y, PANG C, ZHAN Z, et al. Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model[J]. IEEE Geoscience and Remote Sensing Letters, 2020, 18(5): 811-815.

[129] ZHANG C, YUE P, TAPETE D, et al. A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 166: 183-200.

[130] CHEN J, YUAN Z, PENG J, et al. DASNet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 14: 1194-1206.

[131] CHEN H, QI Z, SHI Z. Remote sensing image change detection with transformers[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1-14.

[132] XU J Z, LU W, LI Z, et al. Building damage detection in satellite imagery using convolutional neural networks[J]. arXiv preprint arXiv:1910.06444, 2019.

[133] ZHU X, LIANG J, HAUPTMANN A. Msnet: A multilevel instance segmentation network for natural disaster damage assessment in aerial videos[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021: 2023-2032.

[134] JI M, LIU L, BUCHROITHNER M. Identifying collapsed buildings using post-earthquake satellite imagery and convolutional neural networks: A case study of the 2010 Haiti earthquake[J]. Remote Sensing, 2018, 10(11): 1689.

[135] DUARTE D, NEX F, KERLE N, et al. Satellite image classification of building damages using airborne and satellite image samples in a deep learning approach[J]. ISPRS Annals Photogrammetry, 2018, 4(2): 1-8.

[136] RUDNER T G, RUßWURM M, FIL J, et al. Multi3net: segmenting flooded buildings via fusion of multiresolution, multisensor, and multitemporal satellite imagery[C]//Proceedings of the AAAI Conference on Artificial Intelligence: vol. 33: 01. 2019: 702-709.

[137] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[138] WEBER E, KANÉ H. Building Disaster Damage Assessment in Satellite Imagery with Multi-Temporal Fusion[C]//International Conference on Learning Representa-

tions Workshop. 2020: 1-7.

[139] HE K, GKIOXARI G, DOLLÁR P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.

[140] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.

[141] SHEN Y, ZHU S, YANG T, et al. Bdanet: Multiscale convolutional neural network with cross-directional attention for building damage assessment from satellite images [J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1-14.

[142] JADERBERG M, SIMONYAN K, ZISSERMAN A, et al. Spatial transformer networks [J]. Advances in neural information processing systems, 2015, 28.

[143] YU C, WANG J, PENG C, et al. Learning a discriminative feature network for semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 1857-1866.

[144] ROY A G, NAVAB N, WACHINGER C. Concurrent spatial and channel ‘squeeze & excitation’ in fully convolutional networks[C]//International conference on medical image computing and computer-assisted intervention. 2018: 421-429.

[145] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[146] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[147] HUANG Z, WANG X, HUANG L, et al. Ccnet: Criss-cross attention for semantic segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 603-612.

[148] ZHU Z, XU M, BAI S, et al. Asymmetric non-local neural networks for semantic segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 593-602.

[149] YIN M, YAO Z, CAO Y, et al. Disentangled non-local neural networks[C]//European Conference on Computer Vision. 2020: 191-207.

[150] WANG H, ZHU Y, GREEN B, et al. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation[C]//European Conference on Computer Vision. 2020: 108-126.

[151] ZHAO H, JIA J, KOLTUN V. Exploring self-attention for image recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 10076-10085.

[152] FU J, LIU J, TIAN H, et al. Dual attention network for scene segmentation[C]//

Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 3146-3154.

[153] KENTON J D M W C, TOUTANOVA L K. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of NAACL-HLT. 2019: 4171-4186.

[154] WANG Y, XU Z, WANG X, et al. End-to-end video instance segmentation with transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 8741-8750.

[155] WU J, JIANG Y, BAI S, et al. Seqformer: Sequential transformer for video instance segmentation[C]//European Conference on Computer Vision. 2022: 553-569.

[156] GU J, KWON H, WANG D, et al. Multi-scale high-resolution vision transformer for semantic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 12094-12103.

[157] 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. 2020.

[158] TOUVRON H, CORD M, DOUZE M, et al. Training data-efficient image transformers & distillation through attention[C]//International Conference on Machine Learning. 2021: 10347-10357.

[159] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]//European conference on computer vision. 2020: 213-229.

[160] ZHU X, SU W, LU L, et al. Deformable DETR: Deformable Transformers for End-to-End Object Detection[C]//International Conference on Learning Representations. 2020.

[161] ZHENG S, LU J, ZHAO H, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 6881-6890.

[162] VIOLA P, JONES M. Rapid object detection using a boosted cascade of simple features [C]//Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001: vol. 1. 2001: I-I.

[163] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]// 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05): vol. 1. 2005: 886-893.

[164] BAY H, TUYTELAARS T, GOOL L V. Surf: Speeded up robust features[C]// European conference on computer vision. 2006: 404-417.

[165] WANG X, HAN T X, YAN S. An HOG-LBP human detector with partial occlusion handling[C]//2009 IEEE 12th international conference on computer vision. 2009: 32-39.

[166] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.

[167] REDMON J, FARHADI A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

[168] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[169] ZHANG S, WEN L, BIAN X, et al. Single-shot refinement neural network for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4203-4212.

[170] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.

[171] UIJLINGS J R, VAN DE SANDE K E, GEVERS T, et al. Selective search for object recognition[J]. International journal of computer vision, 2013, 104(2): 154-171.

[172] HE K, ZHANG X, REN S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(9): 1904-1916.

[173] GIRSHICK R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[174] DAI J, LI Y, HE K, et al. R-fcn: Object detection via region-based fully convolutional networks[J]. Advances in neural information processing systems, 2016, 29.

[175] PAN X, SHI J, LUO P, et al. Spatial as deep: Spatial cnn for traffic scene understanding [C]//Proceedings of the AAAI Conference on Artificial Intelligence: vol. 32: 1. 2018.

[176] 钟昌源, 胡泽林, 李淼, 等. 基于分组注意力模块的实时农作物病害叶片语义分割模型.[J]. Transactions of the Chinese Society of Agricultural Engineering, 2021, 37(4).

[177] LIU L, WU F X, WANG Y P, et al. Multi-receptive-field CNN for semantic segmentation of medical images[J]. IEEE Journal of Biomedical and Health Informatics, 2020, 24(11): 3215-3225.

[178] PAPPAS T N, JAYANT N S. An adaptive clustering algorithm for image segmentation[C]//International Conference on Acoustics, Speech, and Signal Processing, 1989: 1667-1670.

[179] DUDA R O, HART P E. Use of the Hough transformation to detect lines and curves in

pictures[J]. Communications of the ACM, 1972, 15(1): 11-15.

[180] OTSU N. A threshold selection method from gray-level histograms[J]. IEEE transactions on systems, man, and cybernetics, 1979, 9(1): 62-66.

[181] WONG A K, SAHOO P K. A gray-level threshold selection method based on maximum entropy principle[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1989, 19(4): 866-871.

[182] LIM Y W, LEE S U. On the color image segmentation algorithm based on the thresholding and the fuzzy c-means techniques[J]. Pattern recognition, 1990, 23(9): 935-952.

[183] CHEN S Y, LIN W C, CHEN C T. Split-and-merge image segmentation based on localized feature analysis and statistical tests[J]. CVGIP: Graphical Models and Image Processing, 1991, 53(5): 457-475.

[184] ADAMS R, BISCHOF L. Seeded region growing[J]. IEEE Transactions on pattern analysis and machine intelligence, 1994, 16(6): 641-647.

[185] GEMAN S. Gibbs distribution, and the Bayesian restoration of images[J]. IEEE Proc. Pattern Analysis and Machine Intelligence, 1984, 6: 774-778.

[186] KRUSKAL J B. On the shortest spanning subtree of a graph and the traveling salesman problem[J]. Proceedings of the American Mathematical society, 1956, 7(1): 48-50.

[187] GRADY L. Random walks for image segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2006, 28(11): 1768-1783.

[188] WEI S, ZHANG H, WANG C, et al. Multi-temporal SAR data large-scale crop mapping based on U-Net model[J]. Remote Sensing, 2019, 11(1): 68.

[189] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs[C]//International Conference on Learning Representations. 2015.

[190] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs [J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(4): 834-848.

[191] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation[J]. arXiv preprint arXiv:1706.05587, 2017.

[192] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2881-2890.

[193] LIN G, MILAN A, SHEN C, et al. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1925-1934.

[194] XIE E, WANG W, YU Z, et al. SegFormer: Simple and efficient design for semantic segmentation with transformers[J]. Advances in Neural Information Processing Systems, 2021, 34.

[195] SIEGEL R L, MILLER K D, JEMAL A. Cancer statistics, 2019[J]. CA: a cancer journal for clinicians, 2019, 69(1): 7-34.

[196] FERLAY J, SOERJOMATARAM I, DIKSHIT R, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012[J]. International journal of cancer, 2015, 136(5): E359-E386.

[197] INFANTE M, CAVUTO S, LUTMAN F R, et al. A randomized study of lung cancer screening with spiral computed tomography: three-year results from the DANTE trial [J]. American Journal of Respiratory and Critical Care Medicine, 2009, 180(5): 445-453.

[198] TEAM N L S T R. Reduced lung-cancer mortality with low-dose computed tomographic screening[J]. New England Journal of Medicine, 2011, 365(5): 395-409.

[199] SINGH S, GIERADA D S, PINSKY P, et al. Reader variability in identifying pulmonary nodules on chest radiographs from the national lung screening trial[J]. Journal of thoracic imaging, 2012, 27(4): 249.

[200] VAN GINNEKEN B, ARMATO III S G, de HOOP B, et al. Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the ANODE09 study[J]. Medical image analysis, 2010, 14(6): 707-722.

[201] SHIN H C, ORTON M R, COLLINS D J, et al. Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4D patient data[J]. IEEE transactions on pattern analysis and machine intelligence, 2012, 35(8): 1930-1943.

[202] ESTEVA A, KUPREL B, NOVOA R A, et al. Dermatologist-level classification of skin cancer with deep neural networks[J]. Nature, 2017, 542(7639): 115.

[203] GULSHAN V, PENG L, CORAM M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs [J]. Jama, 2016, 316(22): 2402-2410.

[204] LITJENS G, KOOI T, BEJNORDI B E, et al. A survey on deep learning in medical image analysis[J]. Medical image analysis, 2017, 42: 60-88.

[205] PENG X, SCHMID C. Multi-region two-stream R-CNN for action detection[C]// European conference on computer vision. 2016: 744-759.

[206] LUO X, SONG T, WANG G, et al. SCPM-Net: An anchor-free 3D lung nodule detection network using sphere representation and center points matching[J]. Medical Image Analysis, 2022, 75: 102287.

[207] ARMATO III S G, MCLENNAN G, BIDAUT L, et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans[J]. Medical physics, 2011, 38(2): 915-931.

[208] SETIO A A A, TRAVERSO A, DE BEL T, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge[J]. Medical image analysis, 2017, 42: 1-13.

[209] YUE K, SUN M, YUAN Y, et al. Compact generalized non-local network[C]// International Conference on Neural Information Processing Systems. 2018: 6510-6519.

[210] LIN T Y, ROYCHOWDHURY A, MAJI S. Bilinear cnn models for fine-grained visual recognition[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1449-1457.

[211] CHOLLET F. Xception: Deep learning with depthwise separable convolutions[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1251-1258.

[212] HOWARD A G, ZHU M, CHEN B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.

[213] XIE S, GIRSHICK R, DOLLÁR P, et al. Aggregated residual transformations for deep neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1492-1500.

[214] WU Y, HE K. Group normalization[C]//European conference on computer vision. 2018: 3-19.

[215] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]//International conference on machine learning. 2015: 448-456.

[216] ETTINGER D S, WOOD D E, AKERLEY W, et al. NCCN guidelines insights: non–small cell lung cancer, version 4.2016[J]. Journal of the National Comprehensive Cancer Network, 2016, 14(3): 255-264.

[217] DETTERBECK F C, MAZZONE P J, NAIDICH D P, et al. Screening for lung cancer: diagnosis and management of lung cancer: American College of Chest Physicians evidence-based clinical practice guidelines[J]. Chest, 2013, 143(5): e78S-e92S.

[218] MANOS D, SEELY J M, TAYLOR J, et al. The Lung Reporting and Data System (LU-RADS): a proposal for computed tomography screening[J]. Canadian Association of Radiologists Journal, 2014, 65(2): 121-134.

[219] KAZEROONI E A, AUSTIN J H, BLACK W C, et al. ACR–STR practice parameter for the performance and reporting of lung cancer screening thoracic computed tomography

(CT): 2014 (Resolution 4)[J]. Journal of thoracic imaging, 2014, 29(5): 310-316.

[220] MCNITT-GRAY M F, ARMATO III S G, MEYER C R, et al. The Lung Image Database Consortium (LIDC) data collection process for nodule detection and annotation[J]. Academic Radiology, 2007, 14(12): 1464-1474.

[221] CLARK K, VENDT B, SMITH K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository[J]. Journal of Digital Imaging, 2013, 26(6): 1045-1057.

[222] PEREIRA F R, MENOTTI D, de OLIVEIRA L F. A 3D Lung Nodule Candidate Detection by Grouping DCNN 2D Candidates.[C]//VISIGRAPP (4: VISAPP). 2019: 537-544.

[223] PASZKE A, GROSS S, MASSA F, et al. Pytorch: An imperative style, high-performance deep learning library[J]. Advances in neural information processing systems, 2019, 32: 8026-8037.

[224] TANG H, ZHANG C, XIE X. Nodulenet: Decoupled false positive reduction for pulmonary nodule detection and segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. 2019: 266-274.

[225] ETTEN A V. City-scale road extraction from satellite imagery v2: Road speeds and travel times[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2020: 1786-1795.

[226] BARZOHAR M, COOPER D B. Automatic finding of main roads in aerial images by using geometric-stochastic models and estimation[J]. IEEE transactions on pattern analysis and machine intelligence, 1996, 18(7): 707-721.

[227] MNIH V, HINTON G E. Learning to label aerial images from noisy data[C]//Proceedings of the 29th International conference on machine learning (ICML-12). 2012: 567-574.

[228] DENG J, DONG W, SOCHER R, et al. Imagenet: A large-scale hierarchical image database[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2009: 248-255.

[229] VAN ETTEN A, LINDENBAUM D, BACASTOW T M. Spacenet: A remote sensing dataset and challenge series[J]. arXiv preprint arXiv:1807.01232, 2018.

[230] DEMIR I, KOPERSKI K, LINDENBAUM D, et al. Deepglobe 2018: A challenge to parse the earth through satellite images[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2018: 172-181.

[231] SINGH S, BATRA A, PANG G, et al. Self-Supervised Feature Learning for Semantic Segmentation of Overhead Imagery.[C]//BMVC: vol. 1: 2. 2018: 4.

[232] TAN Y Q, GAO S H, LI X Y, et al. Vecroad: Point-based iterative graph exploration for road graphs extraction[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2020: 8910-8918.

[233] AHN J, KWAK S. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4981-4990.

[234] LI L, YAN J, WANG H, et al. Anomaly detection of time series with smoothness-inducing sequential variational auto-encoder[J]. IEEE transactions on neural networks and learning systems, 2021, 32(3): 1177-1191.

[235] KIM J Y, BU S J, CHO S B. Zero-day malware detection using transferred generative adversarial networks based on deep autoencoders[J]. Information Sciences, 2018, 460: 83-102.

[236] HAUSKRECHT M, BATAL I, VALKO M, et al. Outlier detection for patient monitoring and alerting[J]. Journal of biomedical informatics, 2013, 46(1): 47-55.

[237] GUEGUEN L, HAMID R. Large-scale damage detection using satellite imagery[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1321-1328.

[238] LIU Z, LI G, MERCIER G, et al. Change detection in heterogenous remote sensing images via homogeneous pixel transformation[J]. IEEE Transactions on Image Processing, 2017, 27(4): 1822-1834.

[239] FU J, LIU J, WANG Y, et al. Stacked deconvolutional network for semantic segmentation[J]. IEEE Transactions on Image Processing, 2019: 1-13.

[240] LIU Y, PANG C, ZHAN Z, et al. Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model[J]. IEEE Geoscience and Remote Sensing Letters, 2020, 18(5): 811-815.

[241] WU C, ZHANG F, XIA J, et al. Building Damage Detection Using U-Net with Attention Mechanism from Pre-and Post-Disaster Remote Sensing Datasets[J]. Remote Sensing, 2021, 13(5): 905.

[242] TAGHANAKI S A, ZHENG Y, ZHOU S K, et al. Combo loss: Handling input and output imbalance in multi-organ segmentation[J]. Computerized Medical Imaging and Graphics, 2019, 75: 24-33.

[243] MILLETARI F, NAVAB N, AHMADI S A. V-net: Fully convolutional neural networks for volumetric medical image segmentation[C]//2016 fourth international conference on 3D vision (3DV). 2016: 565-571.

[244] WANG J, SUN K, CHENG T, et al. Deep high-resolution representation learning for

visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2020, 43(10): 3349-3364.

[245] GUPTA R, SHAH M. Rescuenet: Joint building segmentation and damage assessment from satellite imagery[C]//2020 25th International Conference on Pattern Recognition (ICPR). 2021: 4405-4411.

[246] WANG W, TAN X, ZHANG P, et al. A CBAM based multiscale transformer fusion approach for remote sensing image change detection[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2022, 15: 6817-6825.

[247] LI Q, ZHONG R, DU X, et al. TransUNetCD: A Hybrid Transformer Network for Change Detection in Optical Remote-Sensing Images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-19.

[248] VALANARASU J M J, OZA P, HACIHALILOGLU I, et al. Medical transformer: Gated axial-attention for medical image segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. 2021: 36-46.

[249] ZHANG Y, YANG Q. A survey on multi-task learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2021: 1-20.

致谢

时光如梭，转眼间，博士研究生阶段即将结束。值此毕业之际，特别向帮助过我的人致以最诚挚的谢意！

首先感谢我的博士生导师程明明教授。感谢程老师在我博士期间在科研上以及职业规划上给予的帮助和指导，同时程老师平易近人的性格和严谨踏实的科研态度也深深激励着我，给我树立了科研工作者的榜样。

感谢硕士期间的师兄师姐以及博士期间的实验室同学，感谢大家给予的帮助和支持。

特别感谢我的父母、家人以及女朋友的支持和鼓励，在我最迷茫无助的时候，你们总会无条件的支持我，祝愿你们永远幸福快乐！

最后，感谢二十多年来一直不断努力的自己，是自己的坚持才能让我度过那些痛苦的日子、那些彻夜难眠的日子，也成就了今天的我。

梅杰

2022 年 10 月

于　南开大学-计算机学院

个人简历、在学期间发表的学术论文与研究成果

个人简历

梅杰，出生于 1993 年 05 月 07 日。在 2016 年 6 月本科毕业于北京师范大学地理信息系统专业并获得理学学士学位，在 2019 年 6 月硕士毕业于北京师范大学地图学与地理信息系统专业并获得理学硕士学位。于 2019 年 9 月至今在南开大学就读计算机科学与技术专业博士研究生。

博士期间发表的论文:

1. **Jie Mei**, Ming-Ming Cheng, Gang Xu, Lan-Ruo Wan, and Huan Zhang. SANet: A Slice-Aware Network for Pulmonary Nodule Detection [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, 44(8): 4374 – 4387. (中科院一区，CCF A 类期刊，影响因子 24.314.)
2. **Jie Mei**, Rou-Jing Li, Wang Gao, and Ming-Ming Cheng. CoANet: Connectivity Attention Network for Road Extraction from Satellite Imagery [J]. *IEEE Transactions on Image Processing*, 2021, 30: 8540 – 8552. (中科院一区, CCF A 类期刊，影响因子 11.041.)
3. **Jie Mei**, Yi-Bo Zheng, and Ming-Ming Cheng. D2ANet: Difference-Aware Attention Network for Multi-Level Change Detection from Satellite Imagery [J]. *Computational Visual Media*, 2022. (中科院二区，影响因子 4.127.)
4. **梅杰**, 程明明. 基于全局结构差异与局部注意力的变化检测 [J]. 中国科学: 信息科学, 2022. (中文核心，CCF A 类期刊.)
5. Yun Cao#, **Jie Mei#**, Yue-Bin Wang, Li-Qiang Zhang, Jun-Huan Peng, Bing Zhang, and Li-Hua Li. SLCRF: Subspace Learning With Conditional Random Field for Hyperspectral Image Classification [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2020, 59(5): 4203 – 4217. (中科院一区，影响因子 8.125.)

6. Yu-Huan Wu, Shang-Hua Gao, **Jie Mei**, Jun Xu, Deng-Ping Fan, Rong-Guo Zhang, and Ming-Ming Cheng. JCS: An Explainable COVID-19 Diagnosis System by Joint Classification and Segmentation [J]. *IEEE Transactions on Image Processing*, 2021, 30: 3113 – 3126. (中科院一区，CCF A 类期刊，影响因子 11.041.)

博士期间其它成果:

1. 程明明；梅杰. 基于切片感知的三维神经网络的肺结节探测方法及系统 [P]. 发明专利号：ZL202011529192.7，授权日：2022-05-17.
2. 程明明；梅杰；郑一博. 一种基于差异注意力神经网络的多级别变化检测方法、系统、介质及电子设备 [P]. 发明专利号：ZL202110083681.2，授权日：2022-04-12.

博士期间参与课题:

1. 项目名称：图像场景理解. 国家自然科学基金优秀青年科学基金项目，项目号: 61922046.
2. 项目名称：不确定环境下小样本目标识别研究. 国家自然科学基金面上项目，项目号: 62176130.
3. 项目名称：知识引导的自适应感知与结构理解. “新一代人工智能”重大项目，项目号: 2018AAA0100400.
4. 项目名称：场景语义智能识别与理解技术. 天津市新一代人工智能科技重大专项，项目号: 18ZXZNGX00110.
5. 项目名称：认知规律启发的弱监督图像场景理解. 天津市杰出青年科学基金项目，项目号: 17JCJQJC43700.