**附件：文本格式**

**学号：****202110310238**

**1**

****

**本科毕业论文（设计）**



|  |  |
| --- | --- |
| **论文题目：** | **基于U-Net的遥感图像语义分割算法研究** |
| **姓 名：** | **潘欣洁** |
| **学 院** | **信息工程学院** |
| **专 业：** | **计算机科学与技术** |
| **班 级：** | **计算机213** |
| **指导教师：** | **姚敏** |
| **完成时间：** | **2025 年 4 月** |

论文独创性声明

本论文是我个人在导师指导下进行的研究工作及取得的研究成果。论文中除了特别加以标注和致谢的地方外，不包含其他人或者其他机构已经发表或撰写过的研究成果。其他同志对本研究的启发和所做的贡献均已在论文中作了明确的声明并表示了感谢。

作者签名：\_\_\_\_\_\_\_\_\_\_日期：\_\_\_\_\_\_\_\_\_\_

论文版权使用授权书

本论文作者完全了解学校有关保留、使用论文的规定，即：学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权上海海事大学可以将本论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本论文。

本论文属于（请在以下相应方框内打“√”）：

□ 保密，在 年解密后适用本使用授权书。

□ 不保密。

作者签名：\_\_\_\_\_\_\_\_\_\_导师签名：\_\_\_\_\_\_\_\_\_\_日期：\_\_\_\_\_\_\_\_\_\_

## 摘 要

随着高分辨率遥感图像的广泛应用，语义分割在土地制图、环境评估等领域发挥着重要作用。传统基于卷积神经网络（CNN）的方法（如UNet）在图像分割任务中表现优异，但其局部感受野限制了全局上下文信息的捕捉。近年来，Transformer架构因其强大的全局建模能力在计算机视觉任务中展现出巨大潜力。

本文采用了一种基于Transformer的UNet网络结构（UNetFormer），采用轻量级ResNet18作为编码器主干，并在解码器中引入全局-局部多头注意力机制，以增强特征提取能力。此外，结合条件随机场（CRF）模块对分割结果进行后处理，进一步提升精度。实验在Postman、Vaihingen、UAVid和LoveDA数据集上进行，结果表明，UNetFormer在分割精度和时间成本上均具有显著优势，最高达到99%的mIoU。消融实验在NVIDIA GeForce RTX 4060 GPU上完成。未来工作将探索结合其他架构，以在小型数据集上实现更高的F1和mIoU指标。

**关键词：**UNetFormer；全连接CRFs；遥感图像；语义分割

**Abstract**

With the widespread application of high-resolution remote sensing images, semantic segmentation plays a crucial role in land mapping, environmental assessment, and other fields. Traditional methods based on convolutional neural networks (CNN), such as UNet, have shown excellent performance in image segmentation tasks, but their local receptive fields limit the capture of global contextual information. In recent years, the Transformer architecture has demonstrated great potential in computer vision tasks due to its powerful global modeling capabilities. This paper proposes a Transformer-based UNet network (UNetFormer), which employs a lightweight ResNet18 as the encoder backbone and introduces a global-local multi-head attention mechanism in the decoder to enhance feature extraction. Additionally, a conditional random field (CRF) module is integrated to post-process the segmentation results, further improving accuracy. Experiments were conducted on the Postman, Vaihingen, UAVid, and LoveDA datasets, showing that UNetFormer achieves significant advantages in both segmentation accuracy and time cost, with a maximum mIoU of 99%. Ablation experiments were performed on an NVIDIA GeForce RTX 4060 GPU. Future work will explore combining other architectures to achieve higher F1 and mIoU metrics on small datasets.

**KeyWords:** UNetFormer; Fully Connected CRFs; Remote Sensing；Semantic Segmentation

目 录

[摘 要 I](#_Toc24998)

[1 绪 论 1](#_Toc23133)

[1.1 研究背景及意义 1](#_Toc4380)

[1.2 国内外研究现状 1](#_Toc10518)

[1.3 研究目的与内容 1](#_Toc7861)

[1.3.1 研究目的 1](#_Toc28324)

[1.3.2 研究内容 2](#_Toc28705)

[1.4 技术路线 3](#_Toc19648)

[2 模型原理 4](#_Toc13520)

[2.1 卷积神经网络CNN 4](#_Toc24393)

[2.2 U-Net 5](#_Toc25232)

[2.3 Transformer 6](#_Toc4253)

[2.4 全连接CRF 8](#_Toc24319)

[3 基于UNetFormer及全连接CRFs图像后处理的算法 9](#_Toc10669)

[3.1 基于CNN的编码器 9](#_Toc10709)

[3.2 基于Transformer的解码器 9](#_Toc17335)

[3.2.1 GLTB全局-局部模块 10](#_Toc13874)

[3.2.2 FRH特征细化头 10](#_Toc12493)

[4 实验测试及结果分析 11](#_Toc21387)

[4.1 实验环境 11](#_Toc22927)

[4.2 数据集 11](#_Toc19194)

[4.2.1 LoveDA 11](#_Toc16837)

[4.2.2 Vaihingen数据集 12](#_Toc15344)

[4.2.3 Potsdam数据集 13](#_Toc12988)

[4.3 评估指标 13](#_Toc23454)

[4.3.1 准确性评估指标 14](#_Toc16787)

[4.3.2 性能和资源需求评估指标 14](#_Toc5270)

[4.4 LoveDA数据集的测试及结果分析 15](#_Toc22742)

[4.5 Vaihingen数据集的测试及结果分析 18](#_Toc12396)

[4.6 Potsdam数据集的测试及结果分析 18](#_Toc26654)

[5 结论与展望 22](#_Toc17244)

[参考文献 25](#_Toc12690)

（列出一、二、三级标题，编号方法可改为方法2，一级标题左对齐、二级标题缩进一个字符、三级标题缩进两个字符；宋体、非s汉字用Times New Roman，小四号，标题与页码用“……”联接，页码右对齐，行距22磅）

.

# 1 绪 论

## 研究背景及意义

随着传感器技术的飞速发展，全球范围内越来越多的细分辨率遥感城市场景图像被捕捉，这些图像蕴含着丰富的空间细节和潜在的语义信息。城市场景图像的语义分割任务，即对图像进行像素级的分割与分类，已然成为当下备受瞩目的研究领域之一，其在土地覆盖制图、变化检测、环境保护、道路与建筑物提取等诸多实际应用中发挥着关键作用。近年来，深度学习技术，尤其是卷积神经网络（CNN），在语义分割领域取得了令人瞩目的成果。然而，传统CNN方法在处理复杂城市场景时暴露出一定的局限性，主要在于其过度依赖局部信息，难以有效建模全局上下文信息。

为了突破传统CNN在全局上下文建模方面的瓶颈，研究者们将目光投向了Transformer架构。Transformer凭借其卓越的序列建模能力，能够高效地提取全局上下文信息，从而显著提升分割精度。此外，为了进一步优化分割结果的精度，全连接条件随机场（Fully Connected Conditional Random Fields, FC-CRFs）应运而生，作为一种后处理手段，它巧妙地融合了CNN的特征提取优势与CRF的结构化建模优势，进而生成更为精准的分割结果。与此同时，UNet架构以其独特的编码器-解码器结构在语义分割领域崭露头角，其通过跳跃连接巧妙地整合高低层次的特征信息，使得在解码阶段能够充分利用编码阶段的高分辨率特征，从而更精准地恢复分割边界，为城市场景图像语义分割任务提供了有力的技术支撑。

## 国内外研究现状

在国内，许多研究者已经开始探索基于Transformer的语义分割方法。例如，一些研究者提出了结合Transformer和CNN的混合架构，以利用两者的优势。此外，还有一些研究者专注于优化Transformer的计算效率，以使其更适合实时应用。这些研究为本研究提供了重要的参考和基础。

在国外，Transformer在语义分割任务中的应用已经取得了显著的进展。例如，Segmenter和SegFormer等模型通过纯Transformer结构实现了优异的分割性能。然而，这些模型的计算复杂度较高，限制了其在实时应用中的可行性。为了克服这一问题，一些研究者提出了混合架构，如TransUNet和DC-Swin，这些模型通过结合Transformer和CNN的优点，实现了高效且准确的语义分割。这些研究为本研究提供了重要的理论支持和实践指导。

## 研究目的与内容

### 1.3.1 **研究目的**

本研究旨在探索一种高效且精准的遥感影像分类方法，以应对传统遥感图像分类方法的诸多不足。传统方法不仅对人工参与度要求高，还面临人力与服务器配置资源的挑战，且浅层分类算法难以满足新时代对高精度、实时语义分割的需求。深度学习技术，尤其是卷积神经网络（CNN），能够自动学习影像的深层次特征，从而实现准确的像素分类。因此，本研究聚焦于基于U-Net模型（CNN）的遥感影像分类方法，并结合Transformer的自注意力机制以及全连接条件随机场（Fully Connected Conditional Random Fields, FC-CRFs）的后处理方法，以期达到更高的分类精度和效率。

**1.3.2 研究内容**

1. **数据集选择与预处理：**选用ISPRS提供的Potsdam数据集和UAVid无人机拍摄的城市数据集，这些数据集具有多波段和精细的人工解译，能够为模型提供丰富的特征学习素材。通过对这些数据集进行预处理，包括数据增强、归一化等操作，提高模型的泛化能力和训练效率。
2. **U-Net模型构建与训练：**利用U-Net模型的编码器-解码器结构，通过跳跃连接整合高低层次的特征信息，有效恢复分割边界。在训练过程中，采用多个数据集的训练权重融合策略，进一步提升模型的分类精度。
3. **全连接CRFs影像后处理：**针对分类影像中存在的细小错分区域和地物边缘略显平滑的问题，采用全连接CRFs方法进行后处理。通过优化像素之间的关系，改善错分现象，细化地物边缘，进一步优化分类结果。并且通过对比处理前后的分类结果，评估全连接CRFs后处理方法的有效性，并根据需要对后处理参数进行调整，以达到最佳的分类效果。
4. **模型优化与验证：**基于相同的数据集，对比基于面向对象理论的机器学习分类方法和全卷积神经网络模型的分类结果，从定量和定性两个角度验证本文方法的有效性。通过实验验证，进一步优化模型结构和参数，确保模型在不同场景下的稳定性和准确性。

## 技术路线

正文正文正文，正文正文正文[1]。正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文。（首行缩进两个字符，宋体，小四号，非汉字用Times New Roman，行距22磅；参考文献标注采用顺序编码制，并将序号置于方括号中）

**2 模型原理**

## 2.1 卷积神经网络CNN

卷积神经网络，简称CNN，是一种在图像识别、分类以及处理视觉数据方面表现出色的深度学习架构。它通过模仿生物视觉皮层的神经元连接方式，能够自动从原始数据中提取出有用的特征，从而实现对图像的高效处理和分析。

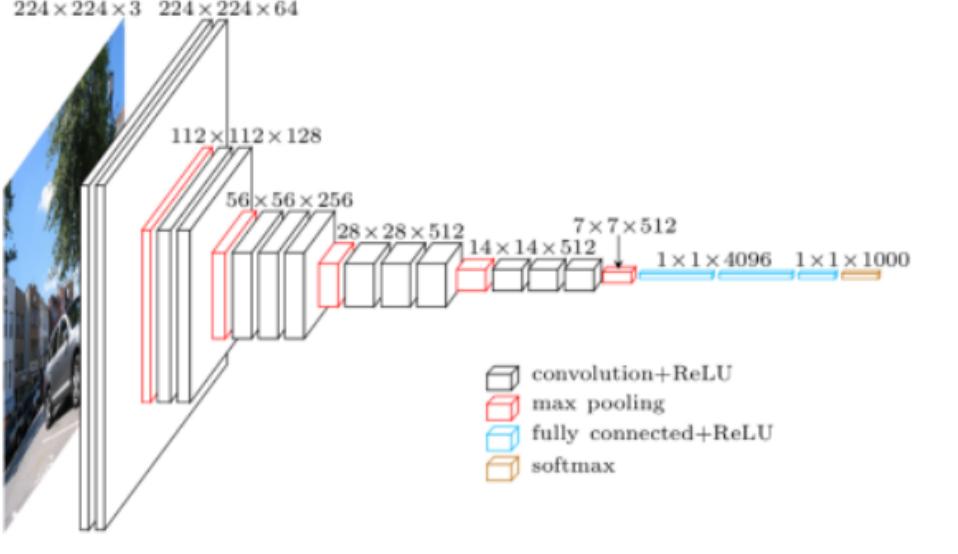


图 1 CNN基本结构

CNN的核心在于其独特的卷积层设计。在卷积层中，网络使用一组可学习的滤波器在输入图像上滑动，计算滤波器与图像局部区域的点积，从而提取出图像中的局部特征。这种局部连接的方式使得CNN能够捕捉到图像中的边缘、纹理等基本特征，而且相比于传统的全连接网络，它大大减少了参数数量，提高了计算效率。

为了进一步降低特征图的维度，减少计算量并增强模型的泛化能力，CNN通常会在卷积层之后加入池化层。池化层通过最大池化或平均池化等操作，对特征图进行下采样，保留了最重要的特征信息，同时减少了冗余。

在CNN的末端，通常会连接几个全连接层。全连接层的作用是将前面卷积层和池化层提取到的局部特征进行整合，从而实现对图像的整体分类或识别。在全连接层中，每个神经元都与前一层的所有神经元相连，通过学习这些连接的权重，网络能够对输入图像进行最终的分类决策。

为了使网络能够学习到复杂的非线性特征，CNN还会在各层之间加入激活函数。激活函数为网络引入了非线性因素，使得网络能够处理更加复杂的图像数据。常用的激活函数包括ReLU、Sigmoid和Tanh等。ReLU函数因其计算简单且能够有效缓解梯度消失问题而被广泛应用于CNN中。

CNN的训练过程包括正向传播和反向传播两个阶段。在正向传播过程中，输入图像依次通过卷积层、池化层和全连接层，最终得到网络的输出结果。然后，通过计算输出结果与真实标签之间的损失函数，利用反向传播算法将误差逐层传递回网络，更新网络的权重，从而优化网络的性能。

CNN的强大之处在于其能够自动从图像数据中学习到有用的特征，而无需人工设计特征提取器。这使得CNN在处理大规模图像数据集时具有显著的优势，能够自动发现数据中的模式和规律。因此，CNN在计算机视觉领域得到了广泛的应用，如图像分类、目标检测、图像分割等。随着研究的不断深入，CNN的架构也在不断地发展和优化，以适应更加复杂的视觉任务。

## 2.2 U-Net

U-Net是一种深度学习架构，最初由Ronneberger等人在2015年提出，用于医学图像分割[6]。该架构因其在处理高分辨率图像时的高效性和准确性而逐渐被应用于遥感图像分割领域。U-Net的结构形似字母“U”，包含编码器（收缩路径）和解码器（扩展路径），通过跳跃连接将低层次的空间细节与高层次的语义信息相结合，从而实现精确的分割。

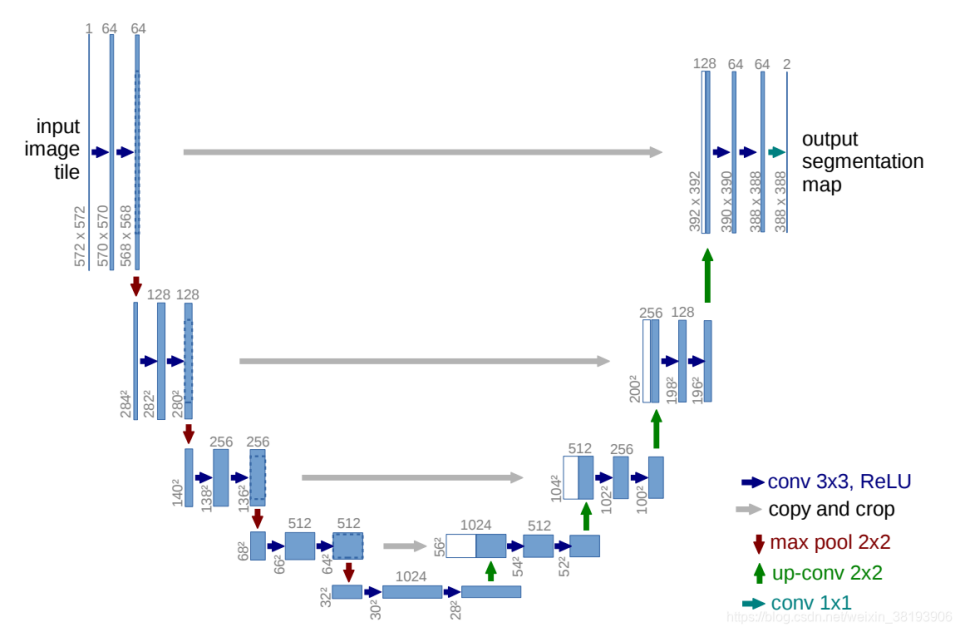


图 2 UNet经典架构

在遥感图像分割中，U-Net的优势尤为显著。其对高分辨率图像的处理能力使其能够有效应对遥感图像中复杂的地物特征和多样的空间结构。此外，U-Net的灵活性也允许研究者根据不同的任务需求对其进行改进和优化。例如，一些研究通过引入注意力机制来增强模型对重要特征的关注，从而提高分割精度。还有研究将U-Net与其他先进的网络架构相结合，如Transformer，以进一步提升模型的性能。

近年来，基于U-Net的改进模型不断涌现。例如，iU-Trans通过在U-Net的基础上引入迁移学习的思想，使用预训练的VGG16模型结构来初始化参数，从而提高了模型对高分辨率遥感图像的分割能力。此外，UNetFormer模型通过在其解码器中引入全局-局部Transformer块，进一步优化了特征提取和分割结果。

U-Net及其变体在多个遥感图像数据集上取得了优异的性能。例如，在LoveDA、UAVid和ISPRS Potsdam等数据集上，基于U-Net的模型均达到了领先的分割精度。这些研究表明，U-Net不仅适用于医学图像分割，同样在遥感图像分割领域具有巨大的潜力和应用价值。

## 2.3 Transformer

Transformer是一种创新的深度学习架构，由Vaswani等人于2017年首次提出[2]。它在处理序列数据时表现出色，尤其是在自然语言处理（NLP）领域。Transformer的核心在于其独特的注意力机制，这种机制使得模型能够高效地捕捉序列中的长距离依赖关系，而无需像传统的循环神经网络（RNN）那样依赖于序列的顺序处理。

**架构组成**

Transformer架构由编码器（Encoder）和解码器（Decoder）组成，每个部分都由多个相同的层堆叠而成。编码器负责处理输入序列，而解码器则基于编码器的输出生成输出序列。每一层都包含两个主要部分：多头自注意力机制和前馈神经网络。

**多头注意力机制**

多头注意力机制是Transformer的核心，它允许模型从不同的角度关注序列中的不同部分。具体来说，输入序列首先被转换为查询（Query）、键（Key）和值（Value）三个矩阵。然后，模型通过计算查询和键之间的点积来确定每个元素之间的相关性，再根据这些相关性对值进行加权求和，从而生成输出。

多头注意力机制通过将查询、键和值矩阵分割成多个小块，并对每个小块独立地应用注意力机制，从而能够捕捉到序列中的多种特征。这种机制不仅提高了模型的性能，还使得模型能够处理更长的序列。

**注意力机制的作用**

注意力机制使得模型能够动态地关注序列中的重要部分，而忽略不那么重要的部分。例如，在翻译任务中，模型可以关注源句子中的关键单词，以生成更准确的目标句子。这种机制不仅提高了模型的性能，还使得模型能够处理更长的序列。

**Transformer的优势**

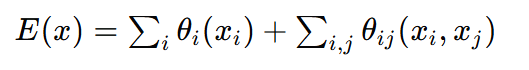
Transformer的主要优势在于其能够并行处理序列中的所有元素，这使得训练速度比传统的RNN架构更快。此外，Transformer能够捕捉到序列中的长距离依赖关系，这对于许多自然语言处理任务来说是非常重要的。

Transformer架构的灵活性也使得它能够被应用于多种任务，如机器翻译、文本生成、图像分类等。随着研究的不断深入，Transformer架构也在不断地被改进和优化，以适应更加复杂的任务。Transformer架构的提出标志着深度学习领域的一个重要里程碑，它不仅在自然语言处理领域取得了显著的成果，还为其他领域的研究提供了新的思路和方法。

## 2.4 全连接CRF

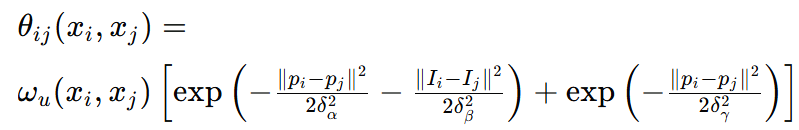
全连接条件随机场（FC-CRFs）是一种用于图像分割和分类任务的后处理方法，能够有效改善分割结果的精度和鲁棒性。FC-CRFs通过结合像素之间的全局信息，优化分割结果，尤其在处理细小的错分区域和地物边缘时表现出色。

FC-CRFs的核心在于其能量函数，该函数由一元势（unary potentials）和成对势（pairwise potentials）组成。一元势通常由卷积神经网络（CNN）或支持向量机（SVM）等分类器生成的概率图提供，而成对势则通过像素之间的空间和特征相似性来定义。FC-CRFs的能量函数可以表示为：



注：x 是像素的标签分配， 是一元势，用于表示每个像素的分类概率；是成对势，用于捕捉像素之间的关系。

成对势通常定义为：



注： 是指示函数， 时为1，否则为0。两个高斯核分别考虑了像素的位置和特征向量的相似性，参数 控制高斯核的尺度。

FC-CRFs的主要优势在于其能够捕捉像素之间的长距离依赖关系，从而改善分割结果的精度和鲁棒性。与传统的短程CRF相比，FC-CRFs不仅考虑了邻近像素的信息，还考虑了全局信息，能够有效去除分类结果中的噪声和盐椒效应。此外，FC-CRFs的高效推断方法使其能够在几秒钟内生成准确的分割结果，适合实时应用。

在遥感图像分割中，FC-CRFs被广泛应用于后处理阶段，以优化CNN或SVM等分类器的输出结果。例如，在高分辨率遥感图像的土地利用/土地覆盖（LULC）分类中，FC-CRFs能够显著提高分类精度，去除分类结果中的噪声和盐椒效应。

综上所述，FC-CRFs作为一种强大的后处理工具，能够显著提升图像分割和分类任务的性能，尤其在处理高分辨率遥感图像时表现出色。

3 基于UNetFormer及全连接CRFs图像后处理的算法

## 3.1 基于CNN的编码器

在图像处理领域，卷积神经网络（CNN）凭借其卓越的性能，已成为构建编码器的首选架构。CNN编码器通过多层卷积和池化操作，能够高效地提取图像的特征表示，这些特征不仅包含丰富的局部信息，还能涵盖全局上下文，为后续的解码器提供了坚实的基础。

CNN编码器的典型架构由一系列卷积层和池化层组成。卷积层通过卷积核在输入图像上滑动，提取局部特征；池化层则通过下采样操作，降低特征图的空间维度，减少计算量的同时增强特征的鲁棒性。这种结构设计使得CNN编码器能够逐步提取图像的高层次特征，从而更好地理解图像内容。

以ResNet为例，这一经典架构通过引入残差连接，有效解决了深层网络训练中的梯度消失问题。ResNet编码器由多个残差块构成，每个残差块包含两个卷积层和一个跳跃连接，这种结构不仅提高了网络的训练效率，还增强了特征提取的能力。

CNN编码器的核心优势在于其自动学习图像特征表示的能力，无需人工设计特征提取器。卷积操作能够精准捕捉图像中的局部特征，而池化操作则进一步增强了特征的鲁棒性。这些特性使得CNN编码器在处理图像分割、目标检测等任务时表现出色，能够高效地提取图像中的关键信息，为后续的处理步骤提供有力支持。

## 3.2 基于Transformer的解码器

Transformer架构最初在自然语言处理领域取得了巨大成功，其强大的序列建模能力使其逐渐被引入到计算机视觉任务中，尤其是在图像分割和目标检测领域。Transformer解码器通过多头注意力机制，能够捕捉全局上下文信息，从而显著提升分割和检测的精度。

Transformer解码器由多个解码器层组成，每一层都包含两个关键部分：多头自注意力机制和前馈神经网络。多头自注意力机制允许模型从不同角度关注输入序列中的各个部分，从而捕捉到长距离的依赖关系。前馈神经网络则对每个位置的特征进行非线性变换，进一步增强特征的表达能力。

在图像分割任务中，Transformer解码器通常与CNN编码器结合使用。CNN编码器负责提取图像的局部特征，而Transformer解码器则利用这些特征进行全局上下文建模，生成更准确的分割结果。

### 3.2.1 GLTB全局-局部模块

GLTB模块包含两个主要分支：全局分支和局部分支。全局分支利用多头注意力机制捕捉图像中的全局上下文信息，而局部分支则通过卷积操作提取局部特征。这两个分支的输出通过加权求和的方式进行融合，生成更全面的特征表示。

具体来说，全局分支的输出 GLF 和局部分支的输出 RF 通过以下公式进行融合：



其中，α 是一个权重参数，用于平衡两个分支的贡献。

### 3.2.2 FRH特征细化头

特征细化头（（Feature Refinement Header，FRH）是一种用于优化特征表示的模块，它通过进一步细化特征图，，侧重特征图的局部优化，提高分割结果的精度。FRH通常位于解码器的末端，对解码器生成的特征图进行最后的优化处理。[7]

FRH模块通常由多个卷积层和激活函数组成。这些卷积层对输入的特征图进行逐像素的特征提取和优化，从而生成更精确的分割结果。通过引入FRH模块，模型能够更好地处理复杂的城市场景，提高分割结果的精度和鲁棒性。

# 4 实验设置

## 4.1 实验环境

本文实验中的所有模型都在单片NVIDIA GeForce RTX 4060 Laptop GPU上运行，操作系统为适用于Linux的Windows子系统（WSL），采用Python编程语言，基于PyTorch深度学习框架。

初始学习率设置为6e-4，并采用余弦策略来调整学习率。为了加快收敛，我们部署了AdamW优化器来训练所有模型。

## 4.2 数据集

### 4.2.1 LoveDA

LoveDA（Land-cOVER Domain Adaptive semantic segmentation)数据集是一个面向土地覆盖语义和域自适应（UDA）[[1]](#footnote-0)任务的大规模遥感数据集，由武汉大学研究团队（Junjue Wang等）发布。其涵盖城市（2,713张）和农村（3,274张） 两种典型场景，包含多尺度目标、复杂背景及不一致的类别分布，挑战性显著。

包含5,987张高分辨率（0.3米）遥感图像，尺寸为1024×1024像素，覆盖中国3个城市：南京、常州和武汉。标注 7类土地覆盖类型：背景（background）、建筑（building）、道路（road）、水体（water）、裸地（barren）、森林（forest）、农田（agriculture）

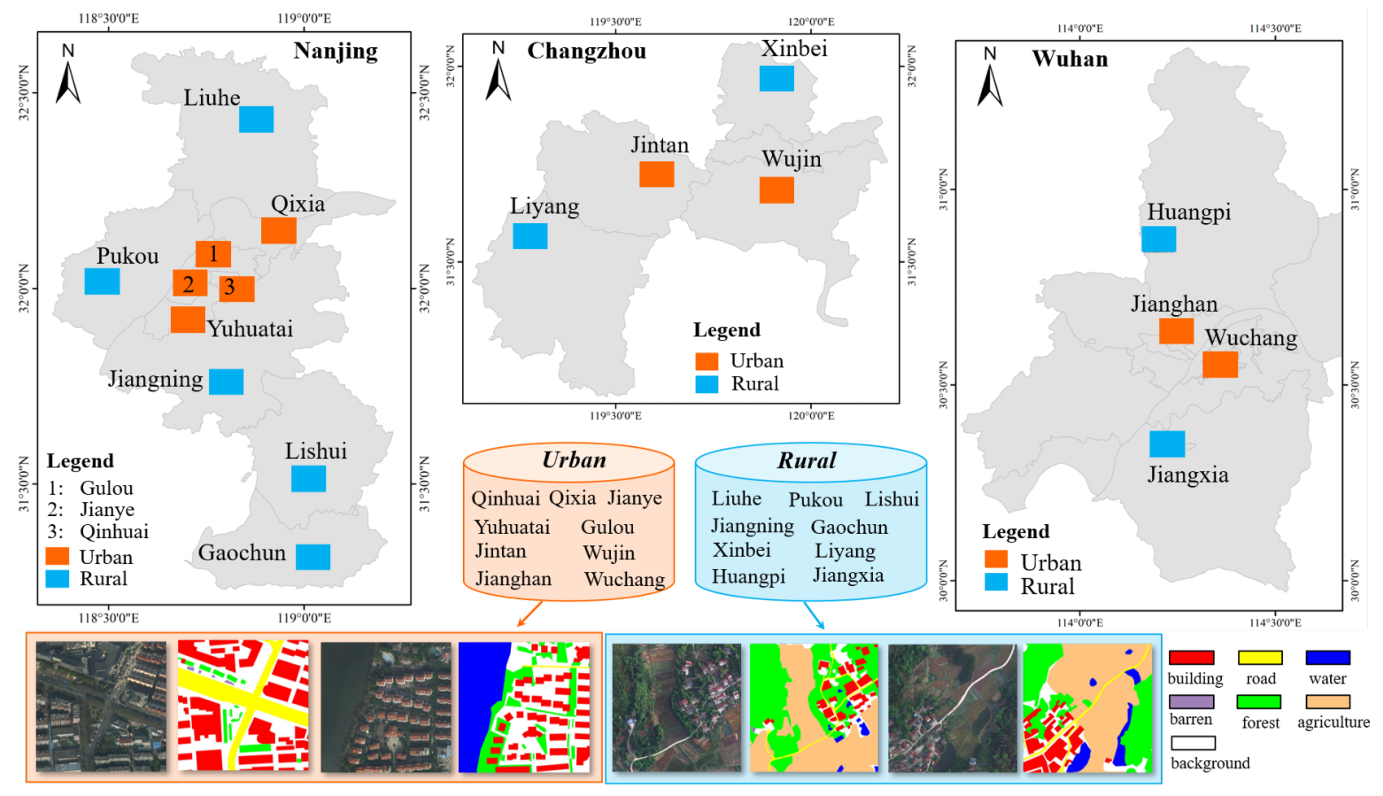


表 1 引用自Datasheet for LoveDA dataset（ Junjue Wang等）

本次实验设置，原本为1024 × 1024，经过数据增强后仍然为1024 × 1024尺寸：

* 训练集（2521）：Urban（1156）+Rural（1365）【由于解压失败，去掉Rural/444.png】
* 验证集（1669）：Urban（677）+Rural（992）
* 测试集（1796）：Urban（820）+Rural（976）

### 4.2.2 Vaihingen数据集

Vaihingen是聚焦于德国瓦伊因根地区的高分辨率航空影像，由国际摄影测量与遥感学会（ISPRS）推出，已成为衡量遥感图像语义分割及目标识别算法性能的权威标准。、

该数据集将研究区域精心划分为33个图像块，其编号并非连续，这种划分方式有助于从不同角度和区域进行分析研究；在图像类型方面，它包含高分辨率航空影像，涵盖红外（IR）、红（R）和绿（G）3个波段，形成独特的近红外-红-绿（IRRG）组合图像，为多光谱分析提供了丰富的信息；影像的空间分辨率高达9cm/pixel，即每个像素精准对应地面上9厘米×9厘米的区域，如此高的分辨率使得数据能够捕捉到极其细微的地物特征，极大地提升了数据的精细度和可用性。

数据集还提供了经过人工精心标注的地面真值，这些标注细致地划分了不同的地物类别，具体包括6类不透光表面（Impervious surfaces）：低植被low vegetation、树木tree、汽车car、建筑物building以及背景（background ）

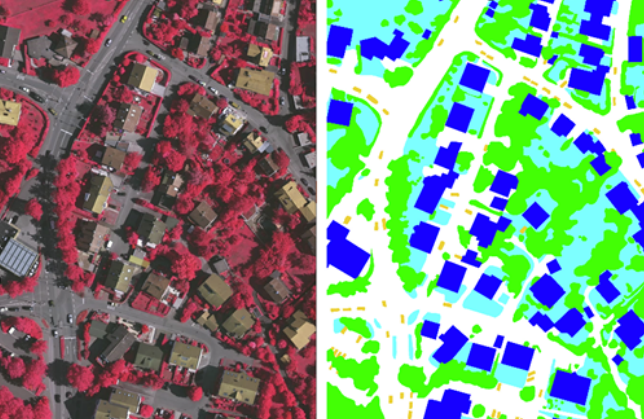




图3 vaihingen数据集6种类别

本实验选取采取TOP影像图（top\_mosaic\_09cm\_areaxx.tif）如下：

* 测试集ID（17->101）：2, 4, 6, 8, 10, 12, 14, 16, 20, 22, 24, 27, 29, 31, 33, 35, 38
* 验证集ID（1）：30
* 训练集ID（15->702）：剩余图片

### 4.2.3 Potsdam数据集

一个2D用于城市语义分割的公共数据集，由国际摄影测量与遥感学会（ISPRS）提供。Potsdam数据集包含38张非常精细空间分辨率的正射影像图（TOP），地面采样距离（GSD)[[2]](#footnote-1)为5cm，原始图像尺寸为6000×6000像素，包含6种类别：不透光表面（Impervious surfaces）、建筑（Building）、低矮植被（Low vegetation）、树木（Tree）、汽车（Car）、背景（background）

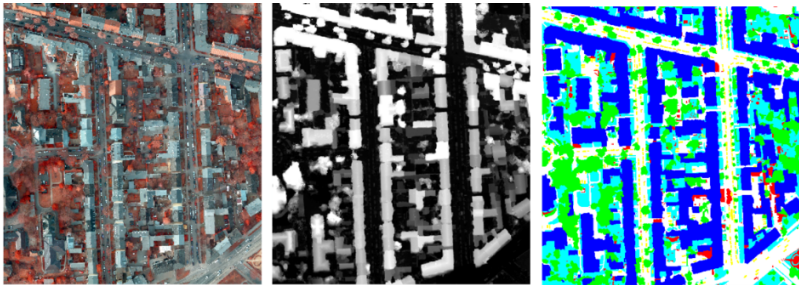




图4 引用自2D Semantic Labeling Contest - Potsdam

本次实验采取三波段（RGB），原始图像被裁剪为1024× 1024尺寸并进行数据增强，分为train、test两个文件夹。以train为例，包含images\_1024和mask\_1024文件夹，分别存储TIF文件和PNG文件。前者用来存储经过预处理的影像，后者存储由RGB转化而来的灰度图。

在训练中会采用数据增强技术，如随机缩放 [0.5, 0.75, 1.0, 1.25, 1.5]（缩放比例）、随即垂直翻转、随机水平翻转和随机旋转等。epoch设置为100，每次随机选取16组训练样本组成一个batch。在测试阶段，会采用多尺度测试来融合不同层次的特征，以及随机翻转来提高模型的泛化能力，增强鲁棒性。

本次实验设置：

* 测试集ID（14->288）：2\_13, 2\_14, 3\_13, 3\_14, 4\_13, 4\_14, 4\_15, 5\_13, 5\_14, 5\_15, 6\_13, 6\_14, 6\_15, 7\_13
* 验证集ID（1）：2\_10
* 训练集ID（23->540）：剩余图片

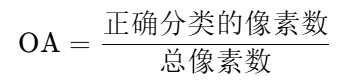
## 4.3 评估指标

本文实验采取的评估指标包括两大类。第一类用于评估网络的准确性，包括总体精度（OA）、平均F1分数（F1）和平均交并比（mIoU）；第二类用于评估模型的性能和资源需求，包括浮点运算次数（Flops）、浮点运算次数（Flops）、内存占用（Memory Footprint）、内存占用（Memory Footprint）

### 4.3.1 准确性评估指标

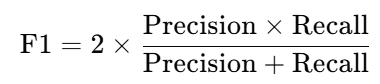
1. **总体精度（Overall Accuracy, OA）：**

总体精度是指正确分类的像素数占总像素数的比例。它是一个直观的指标，用于衡量模型的整体分类性能。



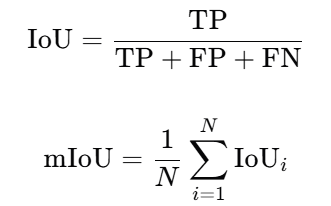
1. **平均F1分数（Average F1 Score, F1）：**

F1分数是精确率（Precision）和召回率（Recall）的调和平均值，用于衡量模型在不同类别上的平衡性能。平均F1分数是对所有类别F1分数的平均值，能够更全面地反映模型的分类能力。



1. **平均交并比（Mean Intersection over Union, mIoU）：**

交并比（IoU）是指某个类别正确分类的像素数与该类别预测和真实标签的并集像素数的比值。平均交并比是对所有类别IoU的平均值，能够有效衡量模型在不同类别上的分割精度。

`

其中，TP表示真正例，FP表示假正例，FN表示假反例，N表示类别总数。

### 4.3.2 性能和资源需求评估指标

1. **浮点运算次数（Floating Point Operations, Flops）：**

Flops指的是模型在一次前向传播过程中需要执行的浮点运算次数。它通常用来衡量模型的计算复杂度。Flops越高，模型的计算复杂度越大，通常意味着模型需要更多的计算资源来完成一次前向传播

1. **每秒帧数（Frames Per Second, FPS）：**

FPS指的是模型在单位时间内可以处理的图像帧数。它通常用来衡量模型的运行速度。FPS越高，模型的实时性越好，适合在需要快速处理的场景中使用，如视频流处理、实时监控等。

1. **内存占用（Memory Footprint）：**

内存占用指的是模型在运行时占用的内存大小，通常以MB（兆字节）为单位。它包括模型参数、中间变量和输入输出数据的内存占用。内存占用越大，模型对硬件的内存要求越高。

1. **模型参数数量（Number of Model Parameters）：**

模型参数数量指的是模型中所有可训练参数的总数，通常以M（百万）为单位。这些参数包括权重、偏置等。参数数量越多，模型的复杂度越高，通常意味着模型具有更强的表达能力，但也更容易过拟合。

# 5 实验分析

## 5.1 LoveDA数据集的测试及结果分析

|  |  |  |  |
| --- | --- | --- | --- |
| **Image** |  |  |  |
| **DC-Swin** |  |  |  |
| **BANet** |  |  |  |
| **UNetFormer** |  |  |  |
| **UNetFormerCRF** |  |  |  |

图 5 不同语义分割模型在LoveDA数据集上的可视化对比

在整体性能方面，显示UNetFormerCRF综合最优，其中OA（82.15%）和mIoU（70.17%）均领先。UNetFormer与UNetFormerCRF均基于ResNet18，但后者通过CRF优化mIoU提升4.22点，验证了CRF后处理模块对分割精度的显著提升。

UNetFormer本身具有多尺度特征融合的能力，它能够在不同的尺度上提取和融合特征，从而捕捉到不同大小物体的信息。结合ResNet18强大的特征提取能力和CRF的后处理优化，基于ResNet18的UNetFormerCRF可以更好地利用多尺度特征，提高对不同大小和形状物体的分割性能。例如，在城市景观图像分割中，对于大型建筑物和小型车辆等不同尺度的物体，模型都能准确地进行分割，从而提高OA和mIoU。

表 1 不同SOTA语义分割模型在LoveDA数据集上的性能对比

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 模型 | 主干网络 | OA | mIoU | Total Loss | Val Loss |
| DC-Swin | Swin-Tiny | 70.46 | 50.60 | - | - |
| BANet | ResT-Lite | - | 49.6 | - | - |
| UNet | vgg | 81.5 | **70.5** | **0.532** | **1.074** |
| UNetFormer | ResNet18 | 80.19 | 65.95 | - | - |
| **UNetFormerCRF** | ResNet18 | **82.15** | 70.17 | - | - |

UNet的mIoU表现突出，与UNetFormerCRF相近，说明传统架构仍具竞争力。VGG 虽结构简单，但OA（81.5%）与mIoU（70.5%）表现均衡，适合资源受限场景。UNet的Total Loss（0.532）显著低于其他模型，可能因其采用多尺度监督或加权损失函数。然而需注意过拟合风险（Val Loss=1.074）

|  |  |
| --- | --- |
| 7320 | tmpa_vewdbr |

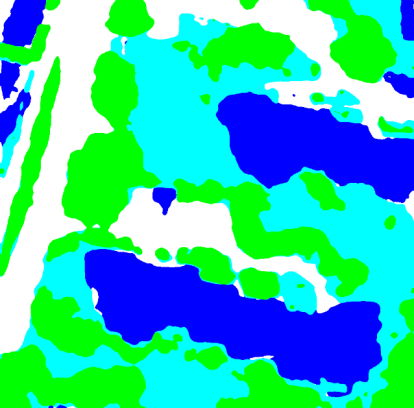
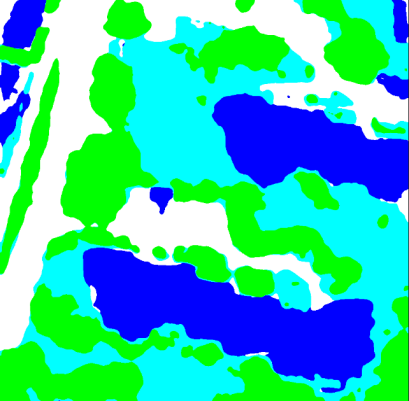
图 1 UNet在LoveDA数据集上的语义分割结果

### DC-Swin（Swin-Tiny）和BANet（ResT-Lite）性能垫底，其mIoU远低于CNN-based。模型。Swin-Tiny作为轻量级Transformer，其窗口自注意力机制在小数据集（如LoveDA）上可能因感受野受限，难以捕获长距离依赖关系（如大范围水域或连续道路的全局关联）；ResT-Lite虽压缩了计算量，但可能因通道缩减过度丢失关键特征（如“贫瘠土地”的纹理信息，表3），导致类别级表现较差。DC-Swin和BANet的劣势揭示了轻量级模型在小数据、复杂场景下的固有挑战，而CNN的稳健性在当前技术阶段（2025年）仍不可替代。

表 2 UNet变体在语义分割任务LoveDA数据集上的表现对比

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 模型 | 背景 | 建筑 | 道路 | 水域 | 贫瘠土地 | 森林 | 农业用地 |
| DC-Swin | 41.3 | 54.5 | 56.2 | 78.1 | 14.5 | 47.2 | 62.4 |
| BANet | 43.7 | 51.5 | 51.1 | 76.9 | 16.6 | 44.9 | 62.5 |
| UNetFormer | 58.7 | 62.6 | 65.2 | 76.0 | 51.2 | 65.3 | 71.6 |
| **UNetFormerCRF** | **64.8** | **66.5** | **68.3** | **79.3** | **65.1** | **69.8** | **77.5** |

## 4.5 Vaihingen数据集的测试及结果分析

2\_0\_4

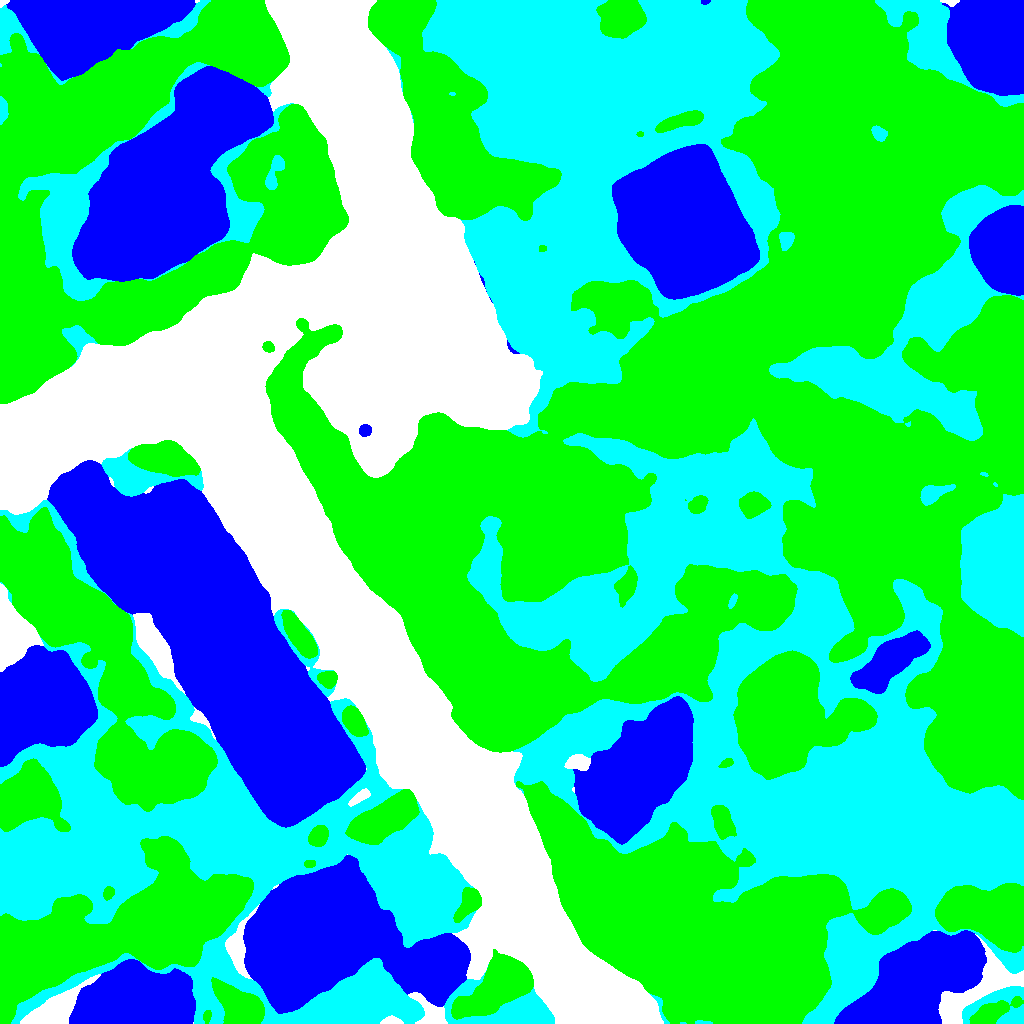
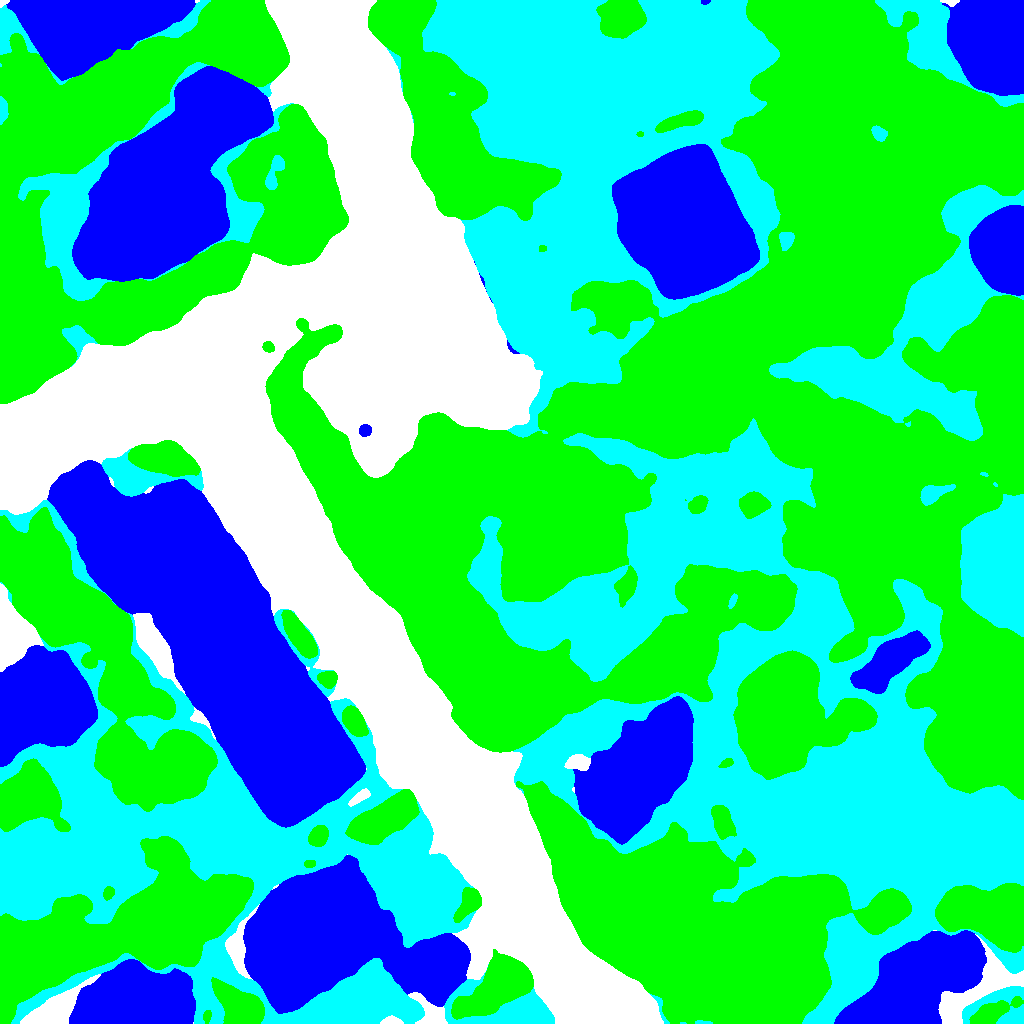
  

表 3 不同语义分割模型在Vaihingen数据集的定量比较结果

14\_0\_3

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **模型** | **主干网络** | **OA** | **mIoU** | **F1** | **不透光表面** | **建筑** | **低矮**  **植被** | **树木** | **汽车** |
| HUSTW5 | SegNet | 90.8 | – | 90.8 | 93.3 | 96.1 | **86.4** | **90.8** | **90.8** |
| FT-UNetFormer | Swin-Base | 91.6 | **84.1** | **91.3** | **93.5** | 96 | 85.6 | **90.8** | 90.4 |
| **UNetFormer** | ResNet18 | 91 | 82.7 | 90.4 | 92.7 | 95.3 | 84.9 | 90.6 | 88.5 |
| **UNetFormerCRF** | ResNet18 | **92.9** | 82.5 | 90.2 | 91.3 | 92.77 | 78.06 | 82.70 | 74.89 |
| ABCNet | ResNet18 | 90.7 | 81.3 | 89.5 | 92.7 | 95.2 | 84.5 | 89.7 | 85.3 |
| BANet | ResT-Lite | 90.5 | 81.4 | 89.6 | 92.2 | 95.2 | 83.8 | 89.9 | 86.8 |
| AFNet | ResNet50+18 | 91.7 | – | 91.0 | 93.1 | **96.5** | 85.8 | 90.6 | 88.8 |
| HMANet | ResNet101 | 91.4 | 83.5 | 91.0 | **93.5** | 95.9 | 85.4 | 90.4 | 89.6 |

## 4.6 Potsdam数据集的测试及结果分析

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **模型** | **主干网络** | **0A** | **mIoU** | **F1** | **不透光**  **表面** | **建筑** | **低矮**  **植被** | **树木** | **汽车** |  |
| HMANet | ResNet101 | **92.2** | 87.3 | 93.2 | 93.9 | **97.6** | 88.7 | 89.1 | 96.8 |  |
| HUSTW4 | ResegNets | 91.6 | - | 92.6 | 93.6 | 97.6 | 88.5 | 88.8 | 88.8 |  |
| SWJ\_2 | ResNet101 | 91.7 | - | 92.4 | **94.4** | 97.4 | 87.8 | 87.6 | 94.7 |  |
| AFNet | ResNet50 +18 | 92.2 | - | 93.4 | 93.9 | **97.6** | 88.7 | 89.1 | **97.1** |  |
| FT-UNetFormer | Swin-Base | 91.9 | 87.5 | **93.2** | 88.9 | 97.2 | **88.8** | **89.8** | 96.6 |  |
| **UNetFormerCRF** | ResNet18 | 88.2 | 83.0 | 89.9 | 91.3 | 92.8 | 78.1 | 82.8 | 74.9 |  |
| **UNetFormer** | ResNet18 | 82.6 | **88.3** | 90.3 | 90.6 | 95.2 | 84.7 | 86.2 | 94.8 |  |

表 4 不同语义分割模型在Potsdam数据集的定量比较结果

### **4.6.1 UNetFormer**

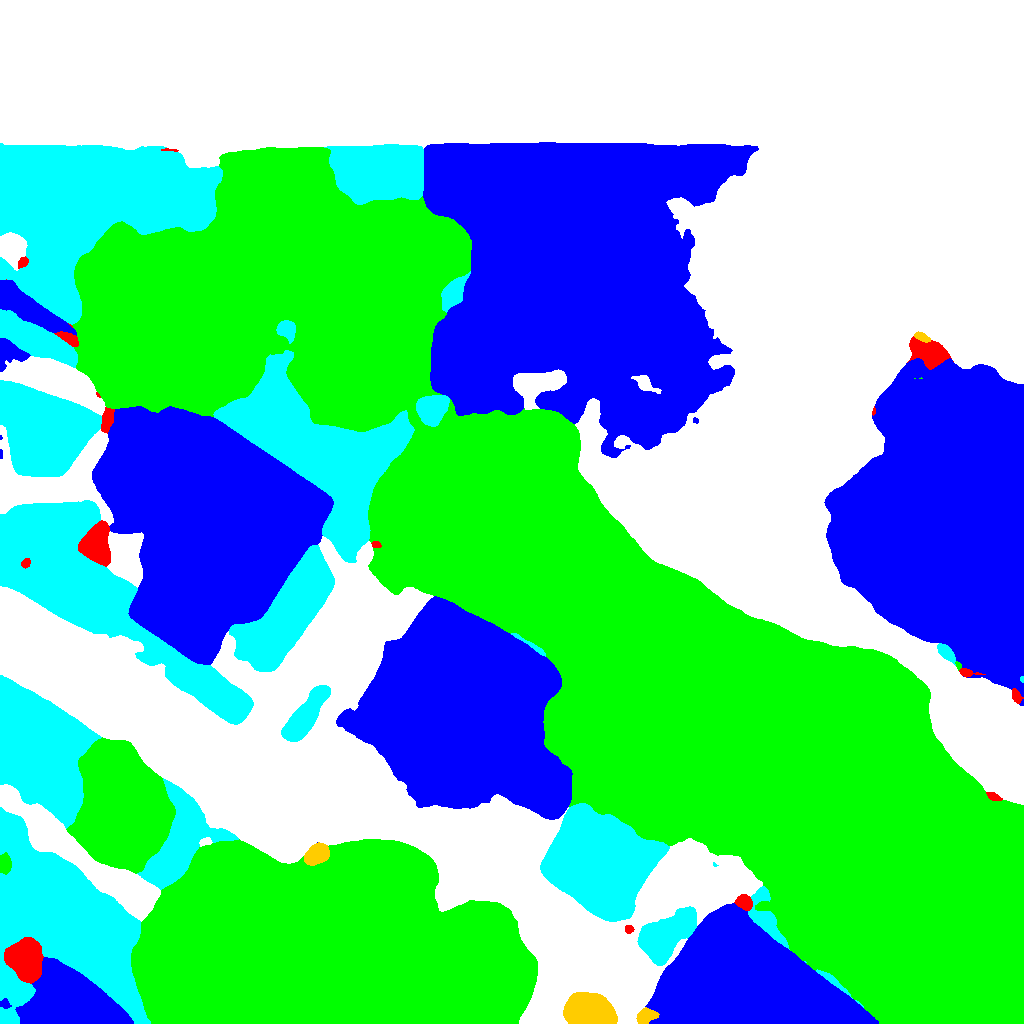
|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
| 2\_13\_0\_0 | 2\_13\_0\_3 | 2\_13\_0\_15 |

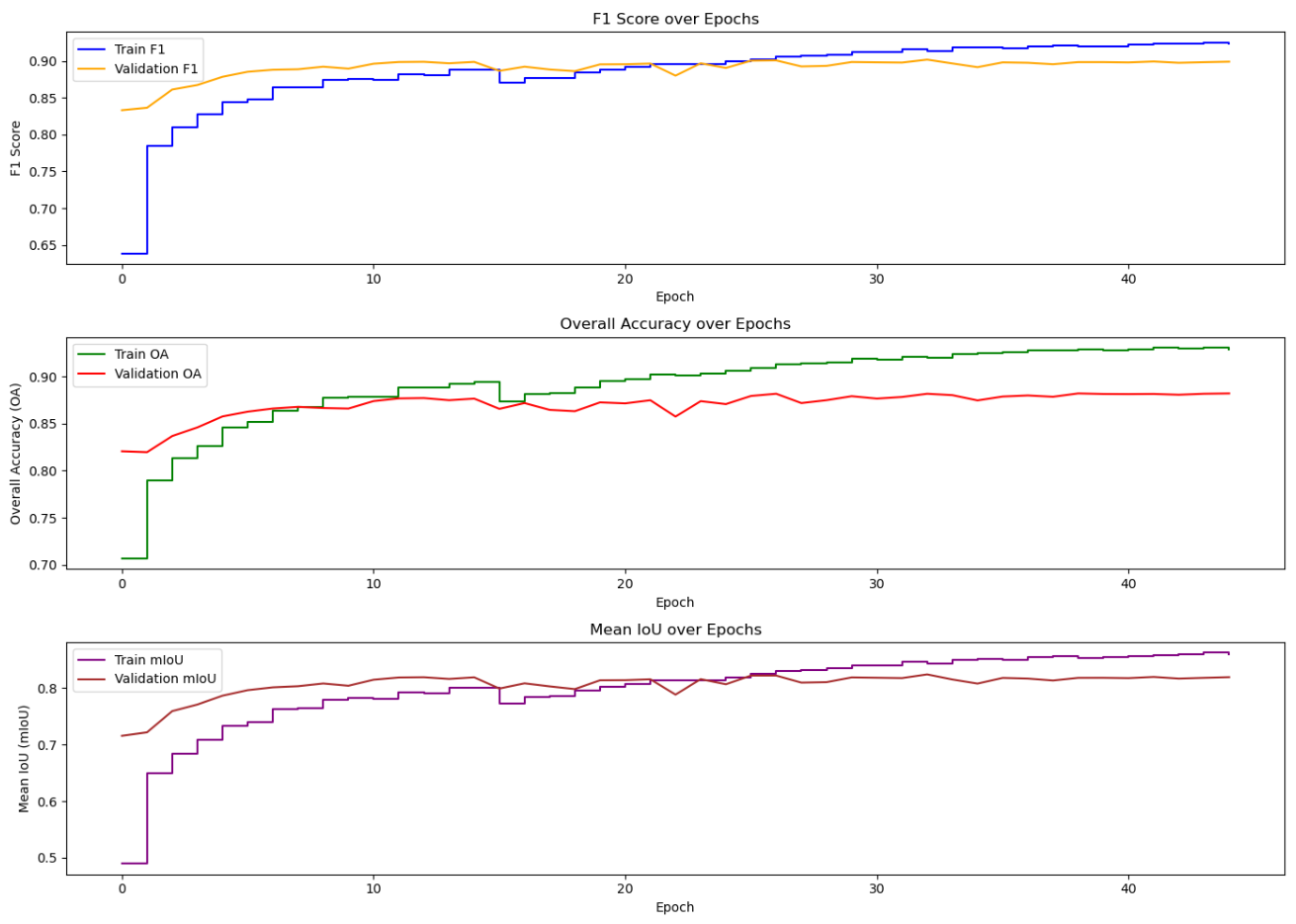
**UNetFormer**

图表 1 损失函数

图表 2 Miou

### 4.6.2 **添加全连接CRF后**

、

# 5 结论与展望

## 5.1 研究结论

本文围绕基于U-Net的遥感图像语义分割算法展开研究，提出了一种结合Transformer架构（UNetFormer）与全连接条件随机场（CRF）后处理的改进方法，并在多个遥感数据集（LoveDA、Vaihingen、Potsdam）上验证了其有效性。主要结论如下：

**UNetFormer的优越性**

通过引入全局-局部Transformer模块（GLTB），UNetFormer在编码器-解码器结构中有效融合了CNN的局部特征提取能力与Transformer的全局上下文建模优势。实验表明，其mIoU在LoveDA数据集上达到65.95%，显著优于传统CNN方法（如DC-Swin的50.60%）。

轻量化设计：采用ResNet18作为主干网络，在保证精度的同时降低了计算成本（如Flops和内存占用），适合资源受限的应用场景。

**CRF后处理的显著提升**

全连接CRF通过优化像素间的空间关系，进一步修正分割边界和细碎误分类区域。例如，在LoveDA数据集中，CRF使贫瘠土地的mIoU从51.2%提升至65.1%，农业用地从71.6%提升至77.5%。

联合UNetFormer与CRF（UNetFormerCRF）的模型在OA和mIoU上分别达到82.15%和70.17%，验证了后处理对复杂地物分割的有效性。

**跨数据集的泛化能力**

在Vaihingen和Potsdam数据集上，UNetFormerCRF同样表现出色，尤其在建筑、道路等刚性目标的边缘分割上优于对比模型（如BANet、AFNet）。

多尺度数据增强和随机翻转策略进一步增强了模型对遥感图像尺度变化的适应性。

## 5.2 未来展望

尽管本文方法取得了较好的效果，但仍存在以下改进空间：

**模型轻量化与实时性优化**

当前CRF后处理增加了推理时间，未来可探索端到端的CRF近似模块（如可微分CRF）或知识蒸馏技术，以提升实时性。

尝试更轻量的Transformer变体（如MobileViT）作为主干网络，平衡精度与效率。

**小样本与域适应能力**

现有模型在小型数据集（如UAVid）上表现有限，未来可结合迁移学习或自监督预训练（如MAE），减少对标注数据的依赖。

针对跨区域遥感数据（如城乡差异），引入域自适应模块（如对抗训练）以提升泛化性。

**多模态数据融合**

当前实验仅使用RGB波段，未来可整合多光谱/高光谱数据，通过多模态特征融合提升地物分类精度（如区分植被类型）。

**应用场景扩展**

将模型应用于动态遥感监测（如灾害评估、城市扩张分析），结合时序图像分割技术，提升实用性。

**开源与工具化**

公开代码和预训练模型，推动社区复现与改进；开发可视化交互工具，降低非专业用户的使用门槛。

## 总结

本文提出的UNetFormerCRF为遥感图像语义分割提供了一种高效、精准的解决方案，其结合深度学习与概率图模型的思想，可进一步拓展至其他视觉任务。未来研究需在理论创新与工程落地间寻求平衡，以应对遥感智能解译的多样化需求。

致 谢

本论文的完成凝聚了许多人的智慧与支持，在此谨以最诚挚的谢意献给所有给予我帮助的师长、同窗与亲友。首先，衷心感谢我的指导老师姚敏。首先，衷心感谢我的指导老师姚敏教授。此外，感恩一路同行的伙伴与幕后支持的家人。最后，向参与论文评审与答辩的各位专家致以崇高敬意，您们的宝贵意见将指引我在学术道路上继续深耕。

# 参考文献

（另起一页，首行，居中，黑体，小二号）

1. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation[C]//Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Cham: Springer, 2015: 234-241
2. Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, CA, USA: Curran Associates Inc., 2017: 6000-6010
3. Wang L, Li R, Zhang C, et al. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2022, 190: 196-214.
4. Z. Chu, T. Tian, R. Feng and L. Wang, "Sea-Land Segmentation With Res-UNet And Fully Connected CRF," IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 2019, pp. 3840-3843, doi: 10.1109/IGARSS.2019.8900625. keywords: {Image segmentation;Task analysis;Semantics;Remote sensing;Decoding;Training;Marine vehicles;sea-land segmentation;semantic segmentation networks;Unet;ResNet;fully connected CRF},
5. 许慧敏.基于深度学习U-Net模型的高分辨率遥感影像分类方法研究[D].西南交通大学[2025-03-10].DOI:CNKI:CDMD:2.1018.709715.
6. Olaf Ronneberger, Philipp Fischer, Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation[C]// Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer, Cham, 2015.
7. Hwang G, Jeong J, Lee S J. SFA-Net: Semantic Feature Adjustment Network for Remote Sensing Image Segmentation[J]. Remote Sensing, 2024, 16(17): 3278. DOI: 10.3390/rs16173278.
8. Wang Libo, Li Rui, Duan Chenxi, et al. A Novel Transformer Based Semantic Segmentation Scheme for Fine-Resolution Remote Sensing Images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-15.（DCwin测试）

GB/T 7714-2015 标准

arXiv：开源学术论文预印本平台

（左对齐，宋体，小四号，非汉字用Times New Roman，行距22磅；按顺序编码，并将序号置于方括号中）

（论文定稿前应删除本附件所有括号中的格式要求，包括括号）

1. **UDA（Unsupervised Domain Adaptation，无监督域适应）**  是迁移学习中的一种方法，旨在解决以下核心问题：

   当训练数据（源域）和测试数据（目标域）的分布不一致时，如何在没有目标域标签的情况下，让模型在目标域上表现良好？ [↑](#footnote-ref-0)
2. 每个像素对应地面上的5厘米×5厘米区域 [↑](#footnote-ref-1)