# 基于Transformer的遥感影像弱监督语义分割

魏梦菲[1,2]，袁和金[1,2]

（1. 华北电力大学 计算机系；
2. 复杂能源系统智能计算教育部工程研究中心，河北 保定 071003）

**摘　要：**针对遥感影像语义分割任务场景复杂、标注成本高的问题，提出一种基于Transformer的端到端图像级标签弱监督语义分割网络。首先，通过多类别标记编码模块，提高类别激活映射图的精确度和细粒度；其次，通过亲和伪标签生成模块进一步完成亲和关系表征的细化，生成高精度亲和伪标签作为分割监督信息，从而提高弱监督网络的能力；同时，设计混合标签数据增强模块强化遥感数据构成；最后，给出融合亲和损失的混合损失函数，强化网络的学习性能。在ISAID数据集上的实验结果表明，该模型在使用图像级标签下分割结果的mIoU达到38.836%，较对照网络表现出更好的鲁棒性和可靠性，在遥感影像弱监督语义分割领域具有较高的应用价值。

**关键词：**遥感影像；弱监督语义分割；图像级标签；Transformer
**DOI：**10. 11907/rjdk. 231624　　开放科学（资源服务）标识码（OSID）：
**中图分类号：**TP751；TP183　　**文献标识码：**A　　**文章编号：**1672-7800（2024）009-0200-09

# Transformer Based Weakly-Supervised Semantic Segmentation of Remote Sensing Images

WEI Mengfei[1,2], YUAN Hejin[1,2]

（1. *Department of Computer, North China Power University*;
2. *Engineering Research Center of Intelligent Computing for Complex Energy Systems, Ministry of Enducation, Baoding* 071003, *China*）

**Abstract:** A Transformer based end-to-end image level weakly supervised semantic segmentation network is proposed to address the complex scene and high annotation cost of remote sensing image semantic segmentation tasks. The network first improves the accuracy and granularity of the class activation map through a multi class label encoding module; Then, the affinity pseudo label generation module is used to further refine the representation of affinity relationships, generating high-precision affinity pseudo labels as segmentation supervision information, thereby improving the ability of weakly supervised networks; Simultaneously designing a mixed label data augmentation module to enhance the composition of remote sensing data; Finally, a mixed loss function with fusion affinity loss is provided to enhance the learning performance of the network. The experimental results on the ISAID dataset show that the model achieves an mIoU of 38.836% in segmentation results using image level labels, demonstrating better robustness and reliability compared to the control network. It has high application value in weakly supervised semantic segmentation of remote sensing images.

**Key Words:** remote sensing image; weakly-supervised semantic segmentation; image-level labels; Transformer

## 0　引言

高分辨遥感影像具有丰富的地物要素信息，被广泛应用于环境监测[1]、测绘[2]、城市规划[3]等领域，是对地观测的主要数据形式。语义分割是计算机视觉领域的重要研究课题，是场景理解的关键环节。遥感影像的语义分割是指对遥感影像中每个像素点分配正确的语义标签，以分割出图像中的不同语义目标。分割模型通常需要大量数据标注，精细的人工标注成本高且耗时久。为解决这一问题，研究者着手设计弱标注的语义分割网络，利用较少的标注数据和大量未标注数据进行图像语义分割，为实际应

**收稿日期：**2023-06-15
**扫描二维码阅读全文：**

**作者简介：**魏梦菲（1998-），女，华北电力大学计算机系硕士研究生，研究方向为遥感图像处理；袁和金（1977-），男，博士，华北电力大学计算机系副教授、硕士生导师，研究方向为图形与图像处理、计算机视觉和模式识别。

用提供了高效便捷的解决方案。

## 1　相关研究

大多数早期神经网络采用全监督的方法提高分割性能。例如，Badrinarayanan 等[4]提出一种深层卷积编码器—解码器结构 SegNet，有效提高了分割准确率；徐昭洪等[5]通过改进 U-Net 的预编码器网络，同时基于空洞卷积的级联并行模块捕获多尺度的高级语义特征，提升了高分遥感中建筑物的分割精度；Niu[6]提出 HMANet，通过多注意力从空间、通道和类别的角度自适应地捕获全局相关性。上述方法均依赖于像素级标签，消耗了大量人工成本。

为降低成本，研究者们开始引入半监督、弱监督和无监督的方法。现有主流半监督方法使用生成对抗网络（Generative Adversarial Network，Gan）进行具有少量像素级标签的全监督分割[7]。例如，Sun 等[8]提出边界感知的半监督语义分割网络 BAS$^4$Net，通过从未标记图像中推断伪标签来完成超高分辨率遥感影像的分割；He 等[9]设计了一种基于一致性正则化和混合扰动范式的半监督学习方法。然而，上述方法在本质上仍然依赖于像素级标签。为进一步减轻像素级标签成本，有学者提出使用弱监督方法。根据不同标签类型可将弱监督方法分为涂鸦、边框和图像级标签 3 类，具体来说，涂鸦使用有限数量的像素提供有关目标位置和类别的信息；边框标签利用几何框捕获相应的目标位置和类别信息；与上述两种方法不同，图像级标签仅提供图像中所包含的类别信息，不提供位置和形状信息，标注简单、成本低，但也更具挑战性。本文即采用图像级标签的弱监督方法研究遥感影像的语义分割问题。

现有多数主流图像级标签网络利用类别激活映射图（Class Activation Mapping，CAM）作为初始伪标签训练分割模型，但鉴于分类网络的固有缺陷，容易造成激活区域不足的问题。Transformer 的注意力机制可对全局特征进行学习，从而关注到更为完整的目标区域。然而在语义分割任务中直接引入 Vision Transformer（ViT）对全局特征进行交互时产生的 CAM 存在过度平滑的现象。为此，本文提出一种基于 Transformer 的端到端图像级弱监督语义分割网络，充分利用像素间的亲和关系，采用 Transformer 设计多类别标记编码模块，使用亲和伪标签生成模块获得激活区域更精确的监督信息；使用混合标签数据增强丰富数据集；给出融合亲和损失的混合损失函数监督网络学习；设计比较实验和消融实验验证所提网络的鲁棒性。研究结果表明，该网络可降低标注成本、简化训练流程，为 ViT 在语义分割领域的应用提供了新思路。

## 2　网络模型构建

基于 Transformer 的端到端图像级标签的弱监督语义分割网络架构如图 1 所示。该网络首先通过多类别标记编码模块实现对遥感影像的编码，得到图像的初始伪标签和亲和关系图；然后通过亲和伪标签生成模块得到用于监督分割模块的亲和伪标签；最后通过分割模块实现最终的语义分割任务，同时运用混合标签数据增强方法增强数据集，以强化网络训练。在网络整体架构中选取目标为“plane”的输入图片为例，示意数据在网络中的流通过程。

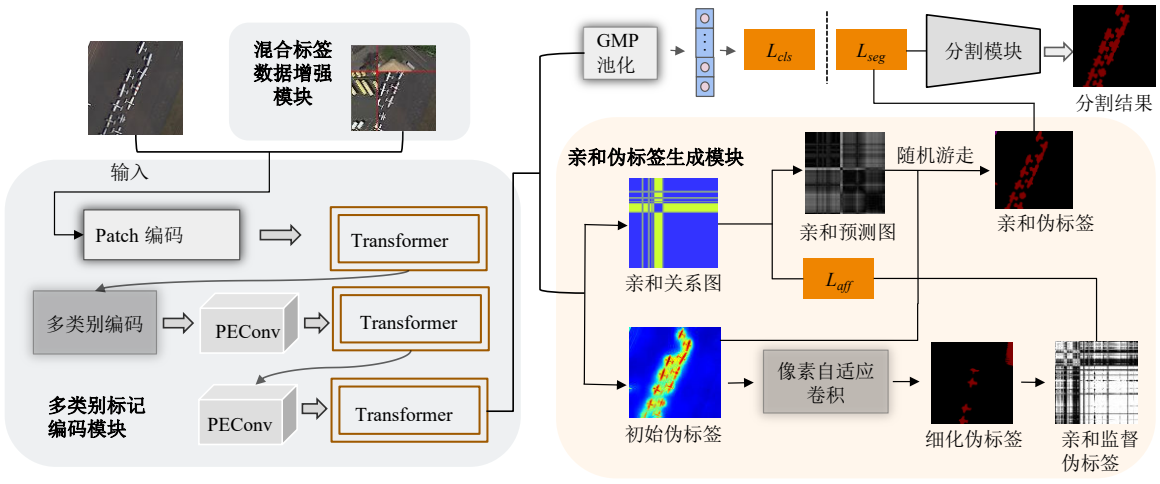



**Fig. 1　Network architecture**

**图 1　网络架构**

### 2.1　多类别标记编码模块

遥感影像中相同类别的目标通常排列密集，具有较大范围的变化尺度和较为明显的纹理颜色差异。为更好地适应遥感影像特点，受 MCTformer 启发，本文设计了多类别标记编码模块来获取包含亲和信息的亲和关系图[10]。模块结构如图 2 所示。

在该模块中，首先对初始标记进行处理，将输入图像分成小的补丁后转换为一维序列向量，每个序列可作为一个补丁标记，将生成的补丁标记作为初始标记送入由多头注意力等构建的第 1 个 Transformer 编码块进行编码；其次

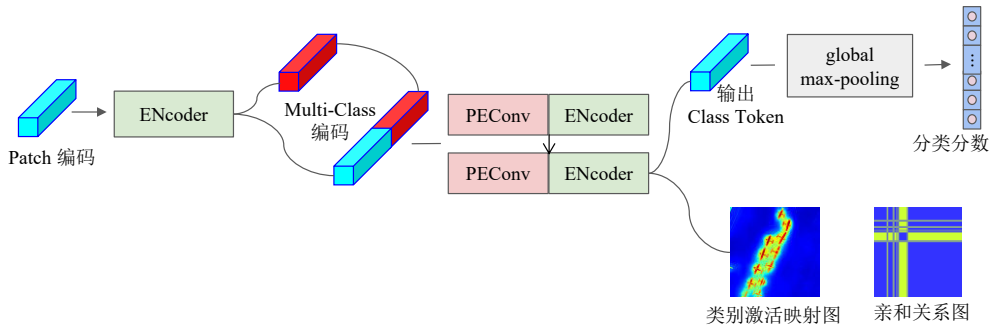



**Fig. 2 Structure of multi-class token encoding module**

**图 2 多类别标记编码模块结构**

将多类别标记与经过第 1 个编码块编码后的输出向量融合以构建新的标记向量；再次将新的标记向量连续送入两个编码块进行编码以生成注意力图，完成补丁标记与类别标记的特征提取；最后从注意力关系中提取亲和关系图，供网络后续模块进一步处理。

在该模块中，图像被切分为 4×4=16 个补丁，映射为一维的补丁标记序列 $S_p \in R^{N\times N\times D}$，其中 $N\times N$=4×4=16 为补丁的维度，$D$=16×16=256 为标记的维度。将多个类别标记向量与经过 1 个 Transformer 编码块编码后的补丁标记向量连接，生成多类别标记向量 $S_c \in R^{(N\times N+C)\times D}$，其中 $C$ 表示类别数。初始标记编码结构如图 3 所示。

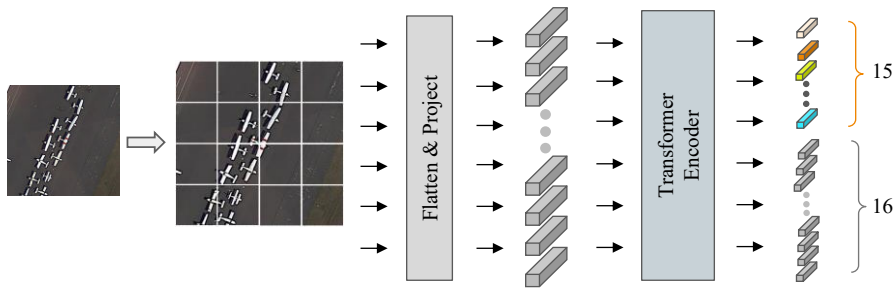



**Fig. 3 Initial tag encoding structure**

**图 3 初始标记编码结构**

以 ViT 为代表的网络采用固定大小的位置编码显式地完成标记向量中位置信息的嵌入，而本文所提网络设计隐式编码方式完成位置信息的嵌入，使得网络在适应图像标记平移等价性的同时实现具备隐式位置信息编码能力的编码结构的搭建。如图 4 所示，在构建完成多类别标记后，首先将标记向量转换为二维空间向量；然后通过卷积核大小为 $k(k\geq 3)$，$(k-1)/2$ 的零填充泛卷积操作完成位置信息的编码。由于多类别标记中没有相关位置信息，本文设计其不参与 PEConv 的操作，在位置信息卷积完成后再将其聚合，最后将聚合后的向量转换为标准的标记向量。

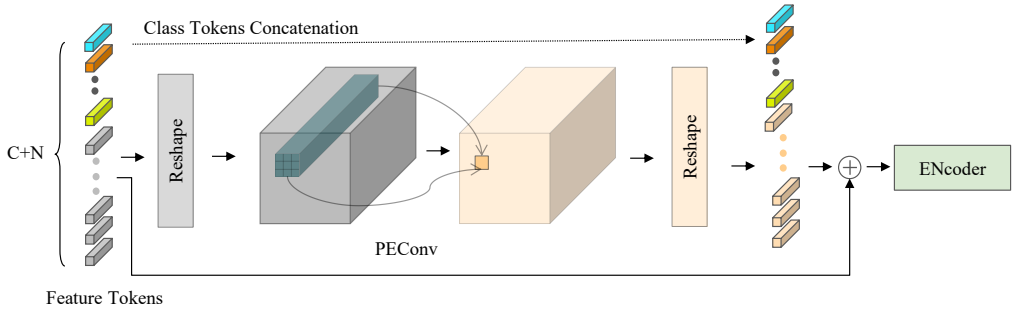



**Fig. 4 Position encoding convolutional structure**

**图 4 位置编码卷积结构**

完成位置信息编码后的标记向量首先经过线性转换得到 $Q \in R^{(N\times N+C)\times D}$、$K \in R^{(N\times N+C)\times D}$ 和 $V \in R^{(N\times N+C)\times D}$，送入多头注意力模块进行学习得到 $Q$ 与 $K$ 之间的关注度；然后通过一个归一化操作和多层感知器（Multilayer Perceptron，MLP）完成整个编码结构的编码工作。重复上述操作，共通过两层编码结构完成最终编码。编码整体结构如图 5 所示。

此外，取注意力图中类别对应补丁关系的部分以获得类别对补丁的注意力分数。考虑到不同层次会捕获不同尺度的图像信息，在经过多级 Transformer 块降低分辨率的

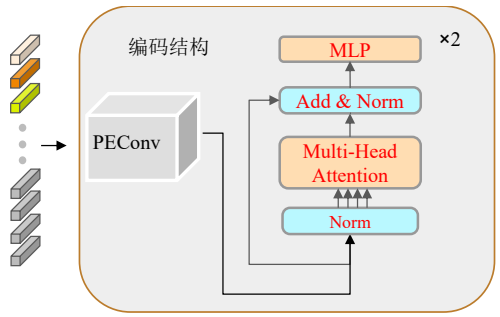



Fig. 5　Encoding structure

图5　编码整体结构

同时增加通道数，解码器混合各个不同分辨率的特征，上采样统一分辨率和通道。此时，特征图同时具备丰富的物体细节和较强的表征能力。将不同编码层输出的注意力图聚合，得到最终的类别补丁关系注意力图。表示为：

$$\hat{A}=\frac{1}{k}\sum_{i}^{k}\hat{A}^{i} \tag{1}$$

式中：$\hat{A}$表示注意力图，$k$表示注意力图聚合的编码层数，$\hat{A}^{i}$表示第$i$层编码层生成的注意力图。

将融合后的类别补丁关系注意力图沿两个不同维度执行min-max标准化(min-max normalization)操作，得到多类别目标定位图，将其作为种子，进一步生成类别激活映射图。如图6所示，从多类别目标定位图中提取补丁间的注意力特征图作为亲和关系图，图中[$P_1$:$P_4$, $P_1$:$P_4$]为图像标记的注意力图，每个补丁向量均表征其以及包括本身在内的所有补丁标记的相关度，利用其作为亲和力图完成网络后续工作。

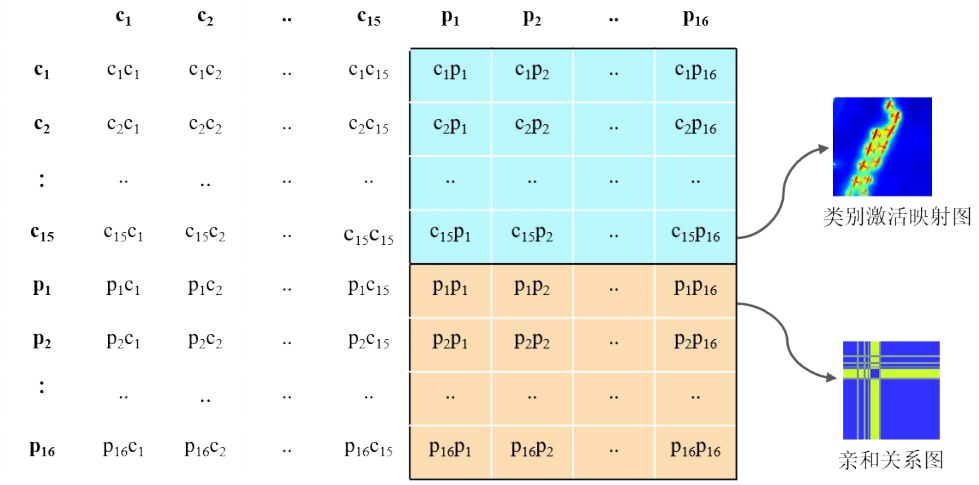



Fig. 6　Class activation mapping and affinity relationship diagram

图6　类别激活映射图与亲和关系图

## 2.2　亲和伪标签生成模块

为使用更加准确的标签监督语义分割任务，解决现有方法激活区域精度不够的问题，本文设计亲和伪标签生成模块，结构如图7所示。为了学习到良好的语义亲和关系，受Xu等[11]启发，使用像素自适应卷积的方法对初始伪标签提取局部RGB信息以生成细化伪标签；然后对细化伪标签进行处理得到亲和监督伪标签，使用亲和监督伪标签作为监督信息约束亲和关系图以生成语义精准的亲和预测图。包含语义的亲和预测图通过随机游走算法对初始伪标签中的高亲和区域进行激活，对低亲和区域进行抑制，从而细化语义边界，最终得到更加可靠的亲和伪标签用于分割任务。

在构建细化伪标签时，首先取图像中任意两个像素位置$(i,j)$和$(k,l)$，计算其像素级变化$T_{(ij,kl)}$。计算公式为：

$$T_{(ij,kl)}=\left[x_{(i-1,j),(k,l)}-x_{(i,j),(k,l)}\right]^{2}+\left[x_{(i,j+1),(k,l)}-x_{(i,j),(k,l)}\right]^{2} \tag{2}$$

然后计算位置间的局部信息核$k_{rgb}^{ij,kl}$。计算公式为：

$$k_{rgb}^{ij,kl}=-\left(\frac{\left|I_{ij}-I_{kl}\right|}{\alpha\sigma_{rgb}^{ij}}\right)^{2} \tag{3}$$

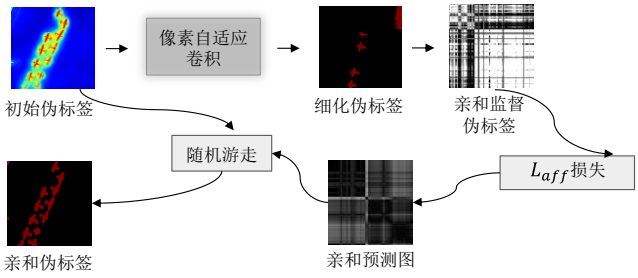



Fig. 7　Structure of affinity pseudo-label generation module

图7　亲和伪标签生成模块结构

式中：$I$表示像素位置点的RGB信息；$\sigma_{rgb}^{ij}$表示标准差；$\alpha$为起平滑作用的权重参数。

为了增强细化伪标签的局部一致性，首先通过观察局部信息核$k_{rgb}^{ij,kl}$对图像中变化较大的像素进行修正，同时通过归一化操作得到新的亲和核。计算公式为：

$$k_{rgb}^{ij,kl}=\frac{\exp\left(k_{rgb}^{ij,kl}\right)}{\sum_{(x,y)\in N(i,j)}\exp\left(k_{rgb}^{ij,kl}\right)}-\beta\frac{\exp\left[T_{(ij,kl)}\right]}{\sum_{(x,y)\in N(i,j)}\exp\left[T_{(ij,kl)}\right]} \tag{4}$$

式中：$N(i,j)$为像素位置点$(i,j)$的邻域点集合，通过扩张卷积获得；$\beta$为权重参数。

然后采用迭代更新策略更新像素标签 CAM，第 $t$ 次迭代得到 $A_t^{i,j,c}$。表示为：

$$A_t^{i,j,c} = \sum_{(k,l)\in N(i,j)} k_{rgb}^{ij,kl} A_{t-1}^{k,l,c} \tag{5}$$

多类别标记编码模块获得的亲和关系图在训练时不受外部约束，不能直接用于监督细化伪标签，需要构建亲和监督伪标签。具体来讲，构造一个与原图像同等大小的矩阵 $A_{ff}$，遍历细化伪标签，使用阈值将细化伪标签区分为前景与背景区域。提取大于前景阈值的区域所对应的激活值最大的语义类别并填入矩阵 $A_{ff}$，遍历其中像素是否具有相同语义，如果语义相同，在亲和监督伪标签中将对应位置设置为正，否则设置为负，生成可信赖度高的亲和监督伪标签。

通过亲和伪标签模块生成的亲和伪标签含有丰富且有利于语义分割任务的特征信息，将其送入分割模块执行最后的分割任务，进一步完成整个基于端到端的 Transformer 网络工作，有效改善本文所提弱监督语义分割网络的分割性能。

### 2.3 混合标签数据增强模块

遥感影像具有类内方差大、类间方差小的特点，即在同一场景内通常具有较大程度的变动性，而不同场景之间又具有一定程度的相似性。因此，为了更好地训练网络的泛化能力，针对现有分类网络处理分割任务时的固有缺点，除传统的几何变换方法外，对网络输入的图像进行混合标签数据增强处理[12]，使网络输入分为单张原始数据集图像和跨图像融合的具有混合标签的增强图像两类。如图 8 所示，在 1 个批次中选取 4 张图像，随机裁剪 4 个图像的某些部分拼接成与原始图像相同大小的新的混合标签拼接图像。

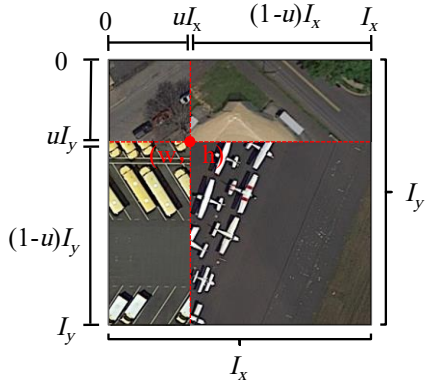



**Fig. 8 Schematic of hybrid label splicing**

**图 8 混合标签拼接示意**

记原始图像的宽和高分别为 $I_x$ 和 $I_y$，假定已知拼接后图像左上小图像的宽为 $w$，左下小图像的高为 $h$。为完成随机裁剪，裁剪区域左上角的坐标界定公式为：

$$x_k \sim u(0, I_x - w_k) \tag{6}$$

$$y_k \sim u(0, I_y - h_k) \tag{7}$$

式中：$k\in\{1,2,3,4\}$。

在裁剪生成混合标签拼接图像的同时，按照图像混合面积比例 $W_k$ 通过 4 个拼接图像的独热编码类标签生成新的混合类别标签。采用公式表示为：

$$W_k = \frac{w_k h_k}{I_x I_y} \tag{8}$$

$$L = \sum_{k\in\{1,2,3,4\}} W_k c_k \tag{9}$$

新的拼接图像由随机选择图像中的一部分区域从空间上拼接合成，丰富了样本数据的内容，促使网络激活各类别中具有较弱判别力的区域。网络在次要特征、部分特征或者其他经常被忽略的信息中学习，有效提高了其在弱监督领域对每个像素分类的准确性。

### 2.4 损失函数

损失函数可以帮助网络自动调整各环节的权重，学习更优的决策策略，从而更好地拟合数据。本文损失函数 *loss* 包括分类损失 $L_{cls}$、亲和损失 $L_{aff}$ 和分割损失 $L_{seg}$。表示为：

$$loss = \lambda_1 L_{cls} + \lambda_2 L_{aff} + \lambda_3 L_{seg} \tag{10}$$

式中：$\lambda_1$ 的作用为调节分类分支在整个网络损失函数中的比重；$\lambda_2$ 和 $\lambda_3$ 的作用分别为调节亲和伪标签分支和分割分支所占比重。亲和损失 $L_{aff}$ 由亲和伪标签监督亲和关系图产生；分割损失 $L_{seg}$ 采用多分类交叉熵损失函数，在分割模块中由亲和伪标签约束分割结果产生。

分类损失函数 $L_{cls}$ 包含单图分类损失 *loss_one* 和混合标签分类损失 *loss_batch*。表示为：

$$L_{cls} = loss\_one + loss\_batch \tag{11}$$

对多类别标记编码分支的输出进行全局最大池化以生成类别分数 *class*。在图像级弱监督分割任务中，唯一可知的只有图像中的类别信息。分类损失 *loss_one* 以图像级的真值标签 *label* 直接监督类别分数 *class*，用公式表示为：

$$loss\_one(class, label) = -\frac{1}{C} * \sum(label * logsigmiod(class) + (1 - label) * logsigmiod(-class)) \tag{12}$$

$$logsigmiod(x) = log(\frac{1}{1 + exp(-x)}) \tag{13}$$

类别分数 *class* 通过多类别标记编码分支的输出从而进行全局最大池化生成，图像混合面积比例 $W_k$ 和标签 *label* 由混合标签数据增强模块生成。*loss_batch* 的计算公式为：

$$loss\_batch = \sum_{k=1}^{4} (W[K] * loss\_one(class, label)) \tag{14}$$

## 3 实验方法与结果分析

### 3.1 实验环境与参数设置

本文实验环境为 Ubuntu20.01.1 的操作系统，采用 CUDA11.2，NVIDIA GeForce RTX 3090 GPU，PyTorch1.9.0 的深度学习框架训练并测试网络模型。实验训练的批处理

大小为 4，使用 AdamW 优化器训练网络，权重衰减因子设置为 0.01，采用 SegFormer 的 ImageNet-1k 权重作为预训练权重进行参数初始化。初始学习率设置为 $5\times10^{-5}$，训练 20 000 轮。对图像随机缩放 0.5~2 倍、随机水平翻转、随机裁剪尺寸进行数据增强。

### 3.2　数据集与评价指标

使用遥感数据集 ISAID 验证所提网络对遥感影像的分割效果。ISAID 数据集包含 15 个分类对象，基本涵盖城市遥感影像的重点目标，体现了遥感影像尺寸和特征分布的差异性和尺度变化性。本文仅使用图像级标签的监督信息。

采用领域常用的像素准确率（pixel Accuracy，pAcc）、平均准确率（mean Accuracy，mAcc）、F1 分数（F1）、交并比（Intersection over Union，IoU）和平均交并比（mean Intersection over Union，mIoU）作为评估指标。其中，pAcc 表示预测类别正确的像素集合与总像素集合之比；mAcc 表示各类别准确率的平均值；IoU 表示预测值和真值的交集与并集之比；mIoU 表示对各类别的 IoU 相加求平均；F1 分数为精确率与召回率的调和平均数。计算公式为：

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{15}$$

式中：precision 为精确率，recall 为召回率，均在实验中由混淆矩阵计算得出。

### 3.3　网络性能比较实验

为证明本文所提端到端网络能有效提高训练效率、降低训练复杂度，同时提升分割结果真值像素级标注间的交并比，将其与多阶段网络 RPNet[13]、TransCAM[14]、ReCAM[15]、SEAM[16]、CONTA[17]，以及单阶段网络 ToCo[18]、AFA[19]等 7 种当下主流的基于深度神经网络的弱监督语义分割网络进行比较。

#### 3.3.1　定量分析

表 1 为本文网络与 7 种对照网络生成的伪分割掩码质量情况比较。可以看出，本文网络生成伪分割掩码的 pAcc 指标达 86.411%、F1 指标达 49.768%、mIoU 指标达 35.499%，其中 mIoU 比 RPNet 网络高出 4.961%，比 ReCAM 高出 1.961%，比 SEAM 网络高出 4.877%，比 CONTA 网络高出 3.259%，比 ToCo 网络高出 3.838%，比 AFA 网络高出 4.297%。虽然本文网络的 mIoU 比 TransCAM 网络低 1.016%，但 TransCAM 网络包含分类网络训练、分割网络训练以及后处理 3 个阶段，流程繁琐且耗时，而本文网络生成伪分割掩码时间更短，效率更高。

**Table 1　Comparison of various indicators of pseudo segmentation mask**

**表 1　伪分割掩码各指标比较**

| 类型 | 网络 | pAcc | F1 | mIoU |
|---|---|---|---|---|
| 多阶段 | RPNet | 77.654 | 44.153 | 30.538 |
| | TransCAM | 87.434 | 51.111 | 36.515 |
| | ReCAM | 85.448 | 47.366 | 33.538 |
| | SEAM | 79.243 | 43.597 | 30.622 |
| | CONTA | 86.160 | 46.016 | 32.240 |
| 单阶段 | ToCo | 88.427 | 45.500 | 31.661 |
| | AFA | 66.975 | 44.937 | 31.202 |
| | 本文网络 | 86.411 | 49.768 | 35.499 |

对最终分割结果中所包含的 16 类分割目标进行统计，图 9 为本文网络与多阶段网络生成的 16 类目标分割结果比较，图 10 为本文网络与单阶段网络生成的 16 类目标分割结果比较。可以看出，本文网络最终分割结果的 mIoU 达到 38.836%，对“basketball-court”和“baseball-diamond”类别的最终分割结果分别达到 53.652% 和 58.491%，对“soccer-ball-field”类别的分割结果达到 60.186%，对“large-vehicle”类别的分割效果超过其他所有对照网络。表明本文网络对遥感影像中大目标的分割效果较好。

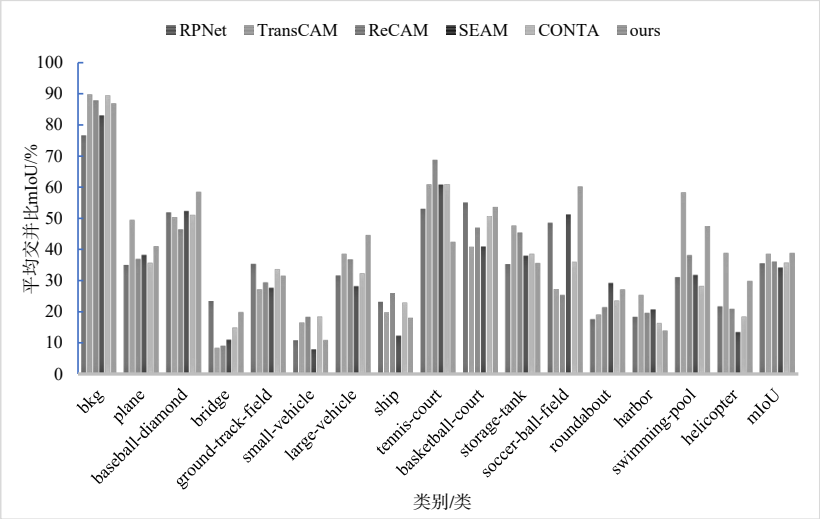



**Fig. 9　Segmentation results of various objects in multi-stage network**

**图 9　多阶段网络各类目标分割结果**

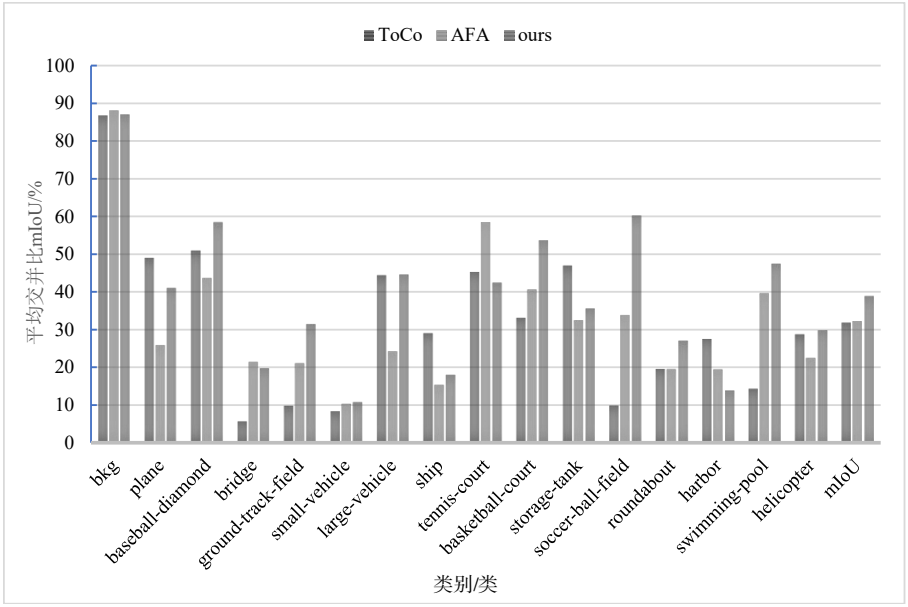


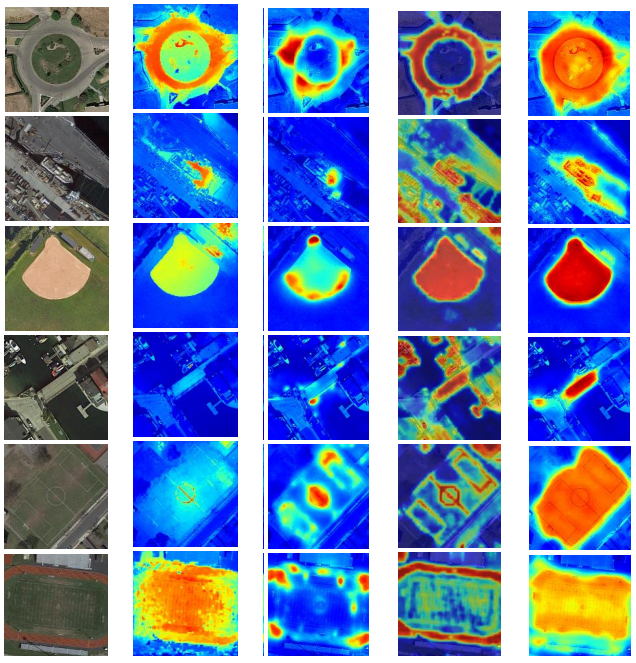

**Fig. 10 Segmentation results of various objects in end-to-end network**

**图 10 单阶段网络各类目标分割结果**

3.3.2 定性分析

图 11 为不同网络在 ISAID 数据集上生成的 CAM。图 12 为不同网络生成的最终分割结果。可以看出，与其他网络相比，本文网络能够完整激活出遥感场景下的大目标区域并且准确激活小目标区域，生成更接近像素级标注的分割结果。

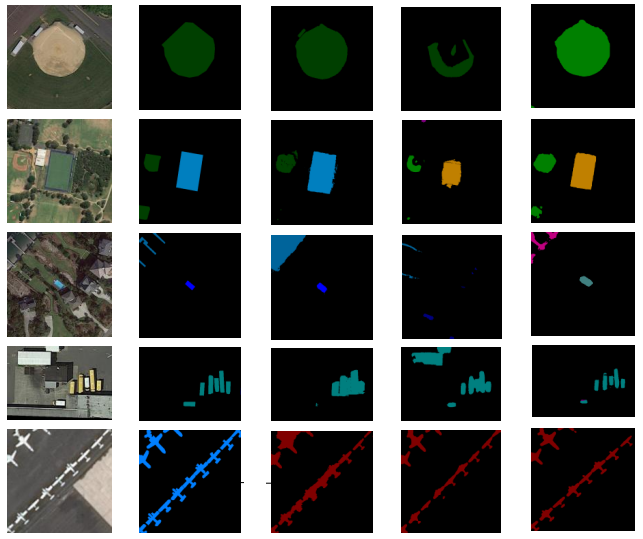
RPNet 通过区域特征比较来识别图像中相似的对象部



(a) Origin images (a) 原图 | (b) SEAM (b) SEAM | (c) CONTA (c) CONTA | (d) ToCo (d) ToCo | (e) The proposed network (e) 本文网络

**Fig. 11 Class activation mapping generated by each network**

**图 11 各网络生成的 CAM**



(a) Origin images (a) 原图 | (b) Ground truth with pixel-level annotations (b) 真值像素级标注 | (c) TransCAM (c) TransCAM | (d) AFA (d) AFA | (e) The proposed network (e) 本文网络

**Fig. 12 Segmentation results generated by each network**

**图 12 各网络分割结果**

分；ReCAM 具有卷积网络的固有特点，忽略了低级语义相关性；SEAM 利用自注意机制关注像素亲和度并用于细化类别激活映射图；CONTA 的最后特征图采用 8 倍下采样，对小目标极不友好；TransCAM 将 Transformer 分支与卷积分支相结合，增加了网络的计算量级；TransCAM、RPNet 和 ReCAM 均采用 Deeplab 系列的分割网络，流程复杂；ToCo 网络需要配置预训练模型才能达到预期性能；AFA 以 Transformer 为基础，由于约束不足存在过度激活的现象。

因此，总体效果均不甚理想。本文网络在遥感影像中得到了质量更高的伪分割掩码，而高精度的伪分割掩码作为监督信息，可减少不同前景类别间错误的信息传播。此外，本文网络对图像中的大目标十分友好，可以较完整地分割出遥感影像中的“baseball-diamond”、“soccer-ball-field”和“plane”；对“swimming-pool”和“large-vehicle”等包含 2~3 类复杂场景图片的分割性能良好。

综上所述，定量和定性分析结果表明，本文网络生成的伪分割掩码精确率、召回率以及最终分割结果精度均达到较高水平。

### 3.4　消融实验

本文消融实验设计验证多类别标记编码模块、亲和伪标签生成模块和混合标签数据增强模块 3 个模块，以及混合损失函数 loss_batch 的有效性。表 2 为本文网络在 ISAID 数据集上的消融实验结果，以 AFA 网络为基线网络。可以看出，多类标记分类模块的多头注意力扩大了网络感受野，利用语义关系引导自注意；亲和伪标签生成模块通过图像降噪帮助网络生成可靠的亲和伪标签作为分割模块的像素级监督；混合标签数据增强模块增加了训练集的多样性，帮助网络学习到良好的表征。在反向传播中对损失函数的优化引导网络及时调整模型权重，以提升网络性能。在引入多类别标记编码模块、亲和伪标签生成模块、混合标签数据增强模块和混合损失函数 loss_batch 后，本文网络生成分割结果的 mIoU 比基线网络提高了 6.595%，证实了各模块的有效性。

**Table 2　The results of ablation experiments on the ISAID dataset using the proposed network**

**表 2　本文网络在 ISAID 数据集上的消融实验结果**

| 网络 | 多类别标记编码模块 | 亲和伪标签生成模块 | 混合标签数据增强模块 | loss_batch | MIoU |
|---|---|---|---|---|---|
| 基线网络 AFA | | | | | 32.241 |
| 本文网络 | √ | | | | 32.611 |
| | √ | √ | | | 33.302 |
| | √ | √ | √ | | 37.260 |
| | √ | √ | √ | √ | 38.836 |

### 3.5　损失权重比较实验

为验证不同权重损失函数取值对本文网络分割性能的影响，并确定最适合的损失函数权重取值，设计损失权重比较实验。其中，$\lambda_1$ 负责调节分类分支在整个网络损失函数中的比重，$\lambda_2$ 和 $\lambda_3$ 分别调节亲和伪标签分支和分割分支所占比重。图 13 为不同取值损失权重下分割结果的 mIoU，实验将 3 个超参数取值分为 10 组，例如[0.5，0.1，0.1]代表 $\lambda_1$ 取值为 0.5、$\lambda_2$ 取值为 0.1，$\lambda_3$ 取值为 0.1。实验结果表明，当 $\lambda_1$ 为 1、$\lambda_2$ 为 0.1、$\lambda_3$ 为 0.1 时分割结果的 mIoU 最高。

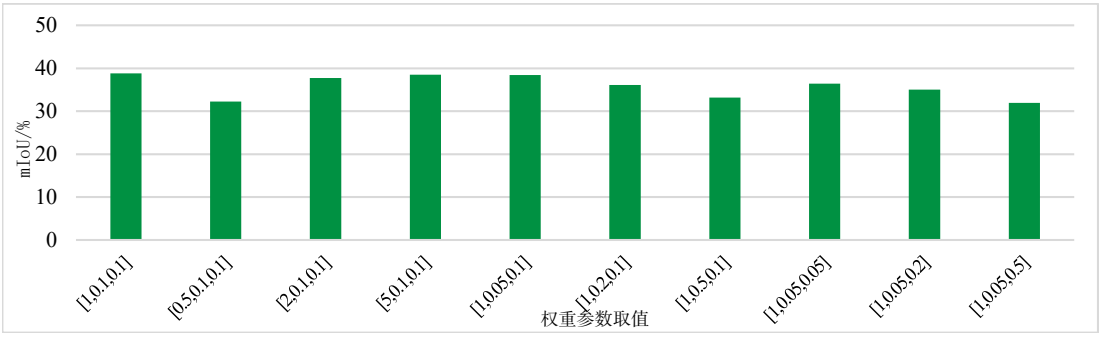



**Fig. 13　mIoU of segmentation results generated by different loss weights**

**图 13　不同损失权重分割结果的 mIoU**

## 4　结语

针对传统卷积和 ViT 的局限性，本文引入多类别标记编码模块和亲和伪标签生成模块，产生更加完整且准确的类别激活映射图作为初始伪标签，对初始伪标签优化为细化伪标签，同时学习到的亲和关系可以对细化伪标签进一步优化，得到可作为分割模块像素级监督的高精度亲和伪标签。混合标签数据增强模块建立了跨图像的语义联系并且实现了多尺度特征融合数据增强，使网络具备更强的遥感数据处理能力。在 ISAID 数据集上经过多次实验，证明本文所提网络提高了初始伪标签的质量和最终分割结果的准确性。本文网络针对遥感影像小目标的分割细粒度仍有提升空间，且只提供图像级标签作为监督信息，虽然点、线和矩形框形式的弱标注时间更长，但是可以得到更多的监督信息。因此，未来研究考虑结合多元弱标注数据展开工作，在减轻标注负担的同时进一步细粒度分割结果以提升网络性能。

**参考文献：**

[1] LIU Y L, ZHANG J, WANG S S, et al. Review of the annual report on global ecological environment remote sensing monitoring: 2012-2021 [J]. Journal of Remote Sensing, 2022, 26(10): 2106-2120.
刘一良，张景，王丝丝，等. “全球生态环境遥感监测年度报告”回顾：2012—2021[J]. 遥感学报，2022，26(10)：2106-2120.

[2] CHEN X Y, YANG K, WANG J S. Extraction of impervious surface in mountain cities by combining sentinel images and feature optimization [J]. Software Guide, 2022, 21(4): 214-219.
陈鑫亚，杨昆，王加胜. 结合 Sentinel 影像与特征优选的山地城市不透水面提取[J]. 软件导刊，2022，21(4)：214-219.

[3] SUN X Y, LI J J, ZHANG R J, et al. Research on the relationship between reconstruction of ancient water system network and urban planning in Xiong'an New Area based on remote sensing [J]. Natural Resources Remote Sensing, 2023,35(1): 132-139.
孙禧勇,李晶晶,张瑞江,等. 基于遥感的雄安新区古水系网重构与城镇规划关系研究[J]. 自然资源遥感,2023,35(1):132-139.

[4] BADRINARAYANAN V, HANDA A, CIPOLLA R. Segnet: a deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017,39(12):2481-2495.

[5] XU Z H, LIU Y, QUAN J C, et al. Semantic segmentation of building objects in remote sensing images based on VGG16 precoding [J] Science, Technology and Engineering, 2019, 19(17):250-255.
徐昭洪,刘宇,全吉成,等. 基于VGG16预编码的遥感影像建筑物语义分割[J]. 科学技术与工程, 2019,19(17):250-255.

[6] NIU R. HmaNet: hybrid multiple attention network for semantic segmentation in aerial images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022,60:1-18.

[7] HUNG W C, TSAI Y H, LIOU Y T, et al. Adversarial learning for semi-supervised semantic segmentation [DB/OL]. https://arxiv. org/abs/1802. 07934.

[8] SUN X, SHI A, HUANG H, et al. BAS$^4$Net: boundary-aware semi-supervised semantic segmentation network for very high resolution remote sensing images[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 13: 5398-5413.

[9] HE Y, WANG J, LIAO C, et al. ClassHyPer: ClassMix-based hybrid perturbations for deep semi-supervised semantic segmentation of remote sensing imagery[J]. Remote Sensing, 2022, 14(4): 879.

[10] XU L, OUYANG W, BENNAMOUN M, et al. Multi-class token transformer for weakly supervised semantic segmentation[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition,2022: 4310-4319.

[11] XU R, WANG C, SUN J, et al. Self correspondence distillation for end-to-end weakly-supervised semantic segmentation[DB/OL]. https://arxiv.org/abs/2302. 13765.

[12] TAKAHASHI R, MATSUBARA T, UEHARA K. Data augmentation using random image cropping and patching for deep CNNs[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 30(9): 2917-2931.

[13] LIU W, KONG X, HUNG T Y, et al. Cross-image region mining with region prototypical network for weakly supervised segmentation [J]. IEEE Transactions on Multimedia,2023,25:1148-1160.

[14] LI R, MAI Z, ZHANG Z, et al. Transcam: Transformer attention-based cam refinement for weakly supervised semantic segmentation[J]. Journal of Visual Communication and Image Representation, 2023, 92: 103800.

[15] CHEN Z, WANG T, WU X, et al. Class re-activation maps for weakly-supervised semantic segmentation[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition,2022: 969-978.

[16] WANG Y, ZHANG J, KAN M, et al. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 12275-12284.

[17] ZHANG D, ZHANG H, TANG J, et al. Causal intervention for weakly-supervised semantic segmentation[J]. Neural Information Processing Systems, 2020, 33: 655-666.

[18] RU L, ZHENG H, ZHAN Y, et al. Token contrast for weakly-supervised semantic segmentation[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition,2023: 3093-3102.

[19] RU L, ZHAN Y, YU B, et al. Learning affinity from attention: end-to-end weakly-supervised semantic segmentation with transformers [C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 16846-16855.

(责任编辑:尹晨茹)