智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于Transformer的多尺度遥感语义分割网络

邵凯, 王明政, 王光宇

引用本文:
邵凯, 王明政, 王光宇. 基于Transformer的多尺度遥感语义分割网络[J]. 智能系统学报, 2024, 19(4): 920-929.
SHAO Kai, WANG Mingzheng, WANG Guangyu. Transformer-based multiscale remote sensing semantic segmentation network[J]. *CAAI Transactions on Intelligent Systems*, 2024, 19(4): 920-929.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202304026

DOI: 10.11992/tis.202304026
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20240306.1404.004

基于 Transformer 的多尺度遥感语义分割网络

邵凯[1,2,3], 王明政[1], 王光宇[1,2]
（1. 重庆邮电大学 通信与信息工程学院, 重庆 400065; 2. 重庆邮电大学 移动通信技术重庆市重点实验室, 重庆 400065; 3. 重庆邮电大学 移动通信教育部工程研究中心, 重庆 400065）

摘　要: 为了提升遥感图像语义分割效果, 本文针对分割目标类间方差小、类内方差大的特点, 从全局上下文信息和多尺度语义特征 2 个关键点提出一种基于 Transformer 的多尺度遥感语义分割网络（muliti-scale Transformer network, MSTNet）。其由编码器和解码器 2 个部分组成, 编码器包含基于 Transformer 改进的视觉注意网络 (visual attention network, VAN) 主干和基于空洞空间金字塔池化（atrous spatial pyramid pooling, ASPP）结构改进的多尺度语义特征提取模块（multi-scale semantic feature extraction module, MSFEM）。解码器采用轻量级多层感知器（multi-layer perception, MLP）配合编码器设计, 充分分析所提取的包含全局上下文信息和多尺度表示的语义特征。MSTNet 在 2 个高分辨率遥感语义分割数据集 ISPRS Potsdam 和 LoveDA 上进行验证, 平均交并比（mIoU）分别达到 79.50% 和 54.12%, 平均 F_1-score（mF_1）分别达到 87.46% 和 69.34%, 实验结果验证了本文所提方法有效提升了遥感图像语义分割的效果。

关键词: 遥感图像; 语义分割; 卷积神经网络; Transformer; 全局上下文信息; 多尺度感受野; 编码器; 解码器
中图分类号: TP391　　**文献标志码:** A　　**文章编号:** 1673−4785(2024)04−0920−10

中文引用格式: 邵凯, 王明政, 王光宇. 基于 Transformer 的多尺度遥感语义分割网络 [J]. 智能系统学报, 2024, 19(4): 920–929.
英文引用格式: SHAO Kai, WANG Mingzheng, WANG Guangyu. Transformer-based multiscale remote sensing semantic segmentation network[J]. CAAI transactions on intelligent systems, 2024, 19(4): 920–929.

Transformer-based multiscale remote sensing semantic segmentation network

SHAO Kai[1,2,3], WANG Mingzheng[1], WANG Guangyu[1,2]
(1. School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China; 2. Chongqing Key Laboratory of Mobile Communications Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China; 3. Engineering Research Center of Mobile Communications of the Ministry of Education, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract: For improving the semantic segmentation effect of remote sensing images, this paper proposes a Transformer based multi-scale Transformer network(MSTNet) based on the characteristics of small inter-class variance and large intra-class variance of segmentation targets, focusing on two key points: global contextual information and multi-scale semantic features. The MSTNet consists of an encoder and a decoder. The encoder includes an improved visual attention network(VAN) backbone based on Transformer and an improved multi-scale semantic feature extraction module(MSFEM) based on atrous spatial pyramid pooling(ASPP) to extract multi-scale semantic features. The decoder is designed with a lightweight multi-layer perception(MLP) and an encoder, to fully analyze the global contextual information and multi-scale representations features extracted by utilizing the inductive property of transformer. The proposed MSTNet was validated on two high-resolution remote sensing semantic segmentation datasets, ISPRS Potsdam and LoveDA, achieving an average intersection over union(mIoU) of 79.50% and 54.12%, and an average F_1-score(mF_1) of 87.46% and 69.34%, respectively. The experimental results verify that the proposed method has effectively improved the semantic segmentation of remote sensing images.

Keywords: remote sensing image; semantic segmentation; convolutional neural network; Transformer; global contextual information; multiscale receptive field; encoder; decoder

近年来, 随着大量搭载高分辨率影像获取设备的卫星被发射并投入使用, 由此产生了海量高

收稿日期: 2023−04−11.　**网络出版日期:** 2024−03−08.
通信作者: 邵凯. E-mail: shaokai@cqupt.edu.cn.

分辨率遥感图像。遥感语义分割将高分辨率遥感图像按照表达语义含义的不同进行分组分割, 理解分析图像中蕴含的丰富地理信息, 为城市规划[1]、环境保护[2]、土地覆盖测绘[3]等任务带来便利。

高分辨率遥感图像的分割目标呈现出类间方差小、类内方差大的特点, 这给遥感语义分割带来极大的挑战。传统的语义分割方法一般可分为基于阈值、边缘、区域、特征等整体分割法, 不能有效区分遥感图像丰富的光谱信息及纹理信息, 十分依赖图像质量, 容易受到光照等噪声的影响。2015 年, 全卷积神经网络[4](fully convolutional networks, FCN)被首次应用于语义分割, 自此之后卷积神经网络(convolutional neural networks, CNN)凭借其在捕获局部细节信息方面的能力成为主流语义分割架构, 后续的 SegNet[5]、U-Net[6]、PSPNet[7]、DeepLab 系列[8-11]等 CNN 网络获得巨大成功。相较于传统分割方法, 深度学习模型考虑了相邻像素之间的空间关系[12], 因此在处理复杂物体样本时展现出更强大的能力。

CNN 的成功归功于其具有特定的归纳偏置, 包括局部性和平移不变性, 这使模型能从更少的数据中学习到更通用的解决方案[13], 从而提升了模型的训练效率和泛化能力。然而, CNN 的这种归纳偏置也导致其缺乏对全局上下文信息的提取能力和远距离依赖关系的建模能力, 在语义分割任务中如果只对局部信息建模, 其分割结果往往是模糊的[14]。在语义分割任务中, 由于分割目标尺度不同、类间方差小且类内方差大, 增强网络对全局上下文信息[15]和多尺度语义特征[16-18]的提取是提升分割效果的 2 个关键点。为获取全局上下文信息, 利用空洞卷积来扩大感受野和利用注意力机制[19]来捕获远距离像素之间的依赖关系是 2 个主要的方案。例如文献 [20] 提出的 DANet、文献 [21] 提出的 CCNet、文献 [22] 提出的 ABCNet 和文献 [17] 提出的 MANet。然而, 这 2 个方案还是无法摆脱对 CNN 主干网络的依赖, 以至于对全局上下文信息提取的有效性产生影响[23]。

近年来 Transformer 从自然语言处理(natural language processing, NLP)领域被引入计算机视觉领域, ViT[24](vision transformer)凭借全局上下文信息提取能力和远距离依赖关系建模能力, 极大提升了网络的全局上下文信息提取能力。然而, 目前 ViT 仍存在一些不足需要改进, 首先, Transformer 是基于自注意力机制设计的, 其计算复杂度与图像大小的二次方成正比, 在语义分割这个密集预测任务中, 特别是在高分辨率遥感图像作为输入图像这个背景下, 这是不可接受的[25]; 其次, Transformer 最初是为 NLP 这样以 1D 结构为输入的任务设计的, 它忽略了图像的 2D 结构, 因此不具备 CNN 的归纳偏置, 这导致基于 Transformer 的网络相较于 CNN 网络需要更多数据去训练才能得到较好的结果[25]; 最后, Transformer 只实现了空间适应性而忽略了通道适应性, 在深度神经网络中不同通道往往代表不同对象, 因此通道维度的适应性对视觉任务非常重要[26]。文献 [26] 提出视觉注意网络 (visual attention network, VAN), 将自注意力替换为大核卷积注意力(large kernel attention, LKA), 其将大核卷积替换为空间局部卷积、空间长程卷积和通道卷积。LKA 通过空间局部卷积和空间长程卷积来同时获取全局上下文信息和局部信息, 通道卷积也为 LKA 带来了通道维度的适应性。这样的设计在兼具 CNN 优点的同时相较于自注意力也降低了计算复杂度, 但是 VAN 却忽略了多尺度感受野的重要性, 导致其对多尺度语义特征的提取不足。

多尺度语义特征是遥感语义分割的第 2 个关键点, 可以通过实现多尺度感受野来获取。常用的实现多尺度感受野的方法有: 多尺度金字塔网络、空洞卷积、空间金字塔池化[7](spatial pyramid pooling, SPP)、空洞空间金字塔池[9](atrous spatial pyramid pooling, ASPP)等。其中 SPP 与 ASPP 是更加适合解决 VAN 缺乏多尺度感受野问题的方法。SPP 将输入的特征图划分成不同尺度的网格, 然后在每个网格内进行池化操作, 最终将不同尺度的池化结果拼接起来。但是, SPP 在处理遥感图像这样的具有大尺度变化的图像时仍然会存在信息丢失的问题。ASPP 引入空洞卷积改进了 SPP, 使得池化操作可以在不降低特征图分辨率的情况下增大感受野, 从而获取更多上下文信息。同时, ASPP 还可以通过调整不同层的空洞率, 获取不同尺度的特征, 进一步提高模型的性能。但是空洞卷积会导致 gradding 问题[27], 主要原因之一是其无法建模远距离像素之间的依赖关系。同时 ASPP 结构中的大量$\lfloor n/k \rfloor$卷积也会增大网络的参数量, 导致网络难以训练[28]。故本文引入扩展邻域注意力[29](dilated neighborhood attention, DiNA)来替换空洞卷积, 通过控制不同的膨胀值实现不同大小的感受野来获取多尺度语义特征。

为应对遥感语义分割中分割目标类间方差小、类内方差大的问题, 本文抓住全局上下文信息和多尺度语义特征这 2 个关键点, 提出一种基

于 Transformer 的多尺度遥感语义分割网络 MSTNet(muliti-scale transformer network, MSTNet)。针对关键点提取全局上下文信息，引入基于 ViT 改进的 VAN 作为网络的主干，其在提取全局上下文信息的同时能保持线性复杂度，兼顾了 CNN 和 Transformer 的优点；针对关键点提取多尺度语义特征，使用 ASPP 结构来实现 VAN 主干缺乏的多尺度感受野，并针对 ASPP 中空洞卷积引起的 gradding 问题和大参数量网络难以训练的问题使用 DiNA 替换 ASPP 中的空洞卷积，基于此设计了多尺度语义特征提取模块(multi-scale semantic feature extract-ion module, MSFEM)加强多尺度语义特征提取能力；同时使用多层感知机(multi-layer perception, MLP)作为网络的解码器，在降低参数量的同时可以充分利用编码器所提取的包含全局上下文信息和多尺度表示的语义特征。

1　MSTNet 网络结构

本文提出的基于 Transformer 的多尺度遥感语义分割网络 MSTNet 如图 1 所示，其整体框架分为编码器和解码器 2 个部分。具体来说编码器分为 VAN 主干和 MSFEM 模块 2 个部分，VAN 主干由 4 个 VAN 块组成，每个 VAN 块由不同数量的 VAN 层组成；MSFEM 模块由 4 个不同膨胀率的 DiNA 和 1 个图片池化分支组成。解码器是 MLP 层。输入图片尺寸为 $H\times W\times 3$(H、W分别为图像的高和宽，3 为图像的通道数)，通过 VAN 主干后分别获得 4 个不同尺寸的特征图。首先取 VAN 主干后 3 个阶段的特征图，进行上采样、拼接等操作将其尺寸转换为$(H/8)\times(W/8)\times 512$，这样可得到包含全局上下文信息的语义特征；接下来，使用 MSFEM 模块来加强多尺度语义特征的提取，并将输出特征图的尺寸调整为$(H/4)\times(W/4)\times 512$；然后将第 1 阶段 VAN 提取的浅层特征与之融合，以弥补由于深层特征而忽略的一些局部信息；最后，使用 1 个轻量级的 MLP 解码器，充分利用编码器提取的语义特征，将浅层包含局部信息的特征与深层的全局特征更好地融合，以实现更准确的遥感语义分割。

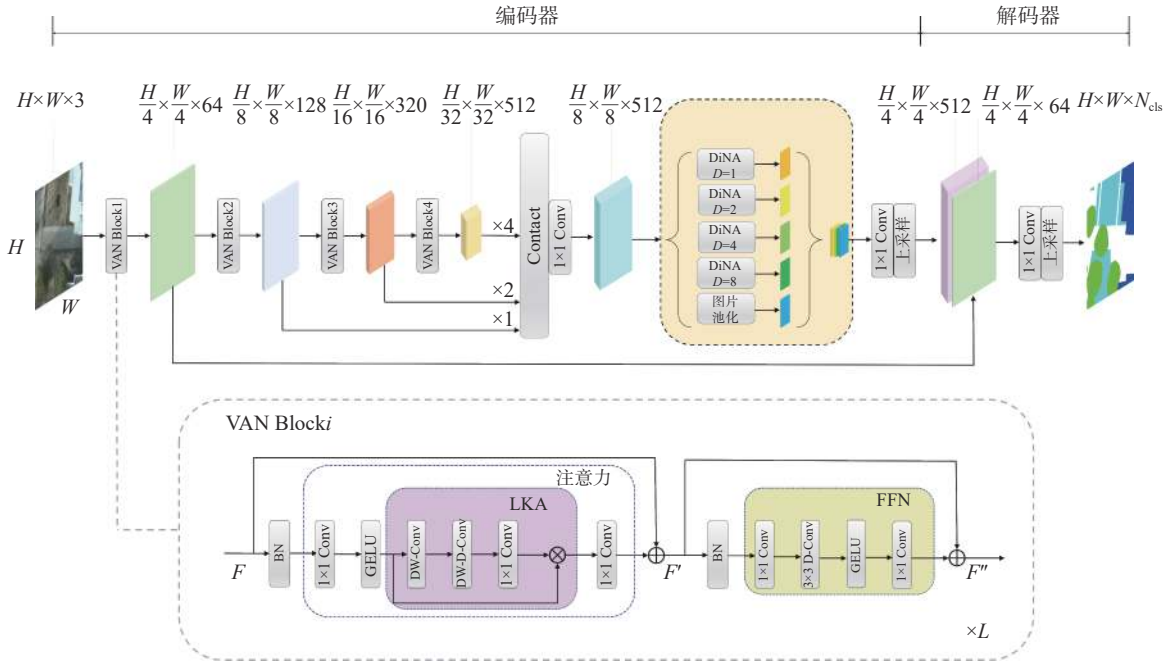


图 1　MSTNet 网络结构

Fig. 1　MATNet network structure

1.1　VAN 主干

针对提取全局上下文信息这个关键点，具有低计算量的 Transformer 网络是最佳解决方案，故本文引人 VAN 作为网络的主干。

1.1.1　LKA

提取全局上下文有 2 种常用方法：1)使用大核卷积，但是这种方式会带来大量的计算开销；2)采用自注意力机制，但是自注意力机制会带来计算量等一系列问题。为克服上述问题，文献 [26] 使用 LKA 注意力机制，该方法通过分解一个大核卷积来捕获远距离依赖关系，同时具备了自注意力机制和大核卷积的优点。

图 2 给出的是大核卷积分解的过程，具体来说是将$K\times K$的大核卷积分解成 1 个$\lceil K/d\rceil\times\lceil K/d\rceil$

的深度扩张卷积，1 个⌊n/k⌋的深度卷积和 1 个⌊n/k⌋通道卷积通过上述分解即可用较小的计算量和参数量获取远距离像素之间的依赖关系。

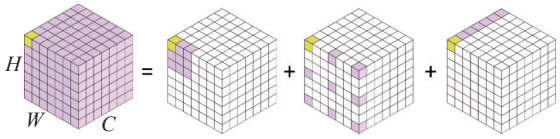


图 2　大核卷积的分解过程

Fig. 2　Large kernel volume integral solution process

1.1.2　VAN

文献 [26] 基于 LKA 设计了一个简单的 Transformer 主干 VAN，如图 1 所示，与 ViT 类似其每个阶段的主要结构还是注意力模块加上前馈层（feed forward networks，FFN）。将一系列的 VAN 块堆叠起来即可得到 VAN 主干。VAN 采用了一种通用的层次结构，其具体包括 4 个空间分辨率递减的阶段。本文选择 VAN-B2 作为网络的主干，其具体网络结构如表 1 所示，其中C代表通道数，L代表不同阶段的 VAN 块的数量，参数量为 26.6×10^6。

表 1　VAN-B2 的网络结构

Table 1　Detailed network structure of VAN-B2

阶段	输出尺寸	VAN-B2
1	$\frac{H}{4}\times\frac{W}{4}\times C$	$C=64$，$L=3$
2	$\frac{H}{8}\times\frac{W}{8}\times C$	$C=128$，$L=3$
3	$\frac{H}{16}\times\frac{W}{16}\times C$	$C=320$，$L=12$
4	$\frac{H}{32}\times\frac{W}{32}\times C$	$C=512$，$L=3$

1.2　MSFEM

多尺度语义特征是语义分割的关键，本文采用 ASPP 结构来实现 VAN 主干缺乏的多尺度感受野从而加强多尺度语义特征的提取。同时针对 ASPP 中空洞卷积引起的 gradding 问题和大参数量网络难以训练的问题，本文利用一种线性复杂度的扩展邻域注意力（dilated neighborhood attention，DiNA）来替换空洞卷积，加强远距离像素之间的依赖关系建模。

文献 [30] 提出了邻域注意力机制（neighborhood attention，NA），其核心思想是允许特征映射中的每个像素只关注其邻近像素，如图 3 所示单个像素的邻域注意可定义为

$$\mathrm{NA}(X_{i,j})=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}_{i,j}\boldsymbol{K}^{\mathrm{T}}_{\rho(i,j)}+B_{i,j}}{\sqrt{d}}\right)\boldsymbol{V}_{\rho(i,j)} \tag{1}$$

式中：$\rho(i,j)$代表在(i,j)处的邻域，$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$为 X 的线性投影（K^T 是 K 的转置），$B_{i,j}$为相对位置偏差，$\sqrt{d}$为缩放参数，d 为嵌入维数。这种邻域注意可扩展到所有像素，这就形成了邻域注意力。如果将每个像素的领域扩展为整个图片，那么其就和自注意力等价了。相较于自注意力的二次复杂度，NA 的复杂度是线性的，并且其复杂度与邻域的大小成线性关系。NA 与卷积类似，它关注每个邻域窗口中的信息，相较于自注意力这种全局注意力机制，NA 可以被叫做是局部注意力。

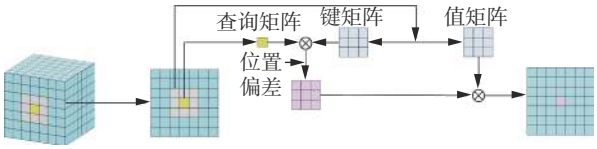


图 3　NA 的具体过程

Fig. 3　The specific process of NA

类比空洞卷积，文献 [29] 基于 NA 提出了扩展邻域注意力 DiNA。DiNA 操作和空洞卷积类似，引入了一个名叫膨胀值的参数。其膨胀值的上限为⌊n/k⌋，其中 n 为输入图分成的令牌的大小，k 代表邻域的大小，当膨胀值为 1 时，DiNA 等同于 NA。通过设置不同的膨胀值，DiNA 可获得一个灵活可变的感受野。相较于空洞卷积，DiNA 可以更好地建模远距离像素之间的依赖关系，尤其是在每个扩张邻域范围内，表现出了更为优异的特性。通过使用多个不同膨胀值的 DiNA 模块，可以轻松获取多尺度特征，从而更加全面地理解输入图像的语义信息。

1.3　解码器

如图 1 所示，MSTNet 的解码器是一个轻量级的 MLP 解码器，其关键思想是利用 Transformer 的诱导特性，即浅层注意力图倾向于保留局部特征，深层注意力图关注全局特征[31]。得益于编码器部分的 VAN 主干和 MSFEM 让网络实现了全局感受野和多尺度感受野，相较于基于 CNN 的网络，MSTNet 不需要去设计复杂的模块去增大网络的感受野和融合多尺度特征。

在解码器部分，MSTNet 首先将 VAN 第一阶段提取的浅层信息与经过 VAN 共 4 个阶段和 MSFEM 的深层信息进行拼接融合，来弥补深层信息中一些局部信息的缺失。然后采用 MLP 层利用融合特征$H\times W\times N_{\mathrm{cls}}$预测分辨率的分割掩膜，其中$N_{\mathrm{cls}}$为预测的类别数。

2　实验与分析

2.1　实验数据集

本文选取国际摄影测量与遥感学会提供的公

共数据集 Potsdam 和武汉大学 RSIDEA 团队提出的 LoveDA 数据集进行实验。

Potsdam 数据集是一个非常具有代表性的历史城市数据集，其中包含了许多大型建筑、狭窄的道路和密集的聚落结构。该数据集共有 38 张图像，每张图像的分辨率为 6000 像素×6000 像素，空间分辨率为 5 cm。此外，该数据集还提供了红外、红色、绿色和蓝色通道以及数字地表模型和标准化数字地表模型等多种数据类型。数据集中都包含最常见的土地覆盖类别，分别是道路、建筑、草地、树木、车辆以及背景。本文使用 ID: 2_10、2_11、2_12、3_10、3_11、3_12、4_10、4_11、4_12、5_10、5_11、5_12、6_7、6_8、6_9、6_10、6_11、6_12、7_7、7_8、7_9、7_10、7_11、7_12 进行训练，使用剩余图片进行测试。在实验中只使用 RGB 三通道正射图像，并且将大图裁剪为 1 024 像素×1 024 像素进行训练和测试。

LoveDA 数据集是武汉大学 RSIDEA 团队提出的地表覆盖分类数据集，其包含来自武汉、南京、常州 3 个城市的 5 987 张 0.3 m 高分辨率影像，每张图像的分辨率为 1 024 像素×1 024 像素。该数据集包括建筑、道路、水、荒地、森林、农田和背景 7 个地表覆盖类别。具体来说，2 522 张图像用于训练，1 669 张用于验证，1 796 张用于测试。

2.2　实现细节

本文的实验平台为 Ubuntu20.04 操作系统，CPU 为 Intel Xeon(R) Platinum 8358P，GPU 为 NVIDIA RTX A5000，显存为 24 GB。开发环境为 Python3.10.8、PyTorch1.13.1、CUDA11.6。训练时采用 Adamw 优化器，初始学习率为 0.0006，衰减率为 0.01，学习策略为 poly，batch size 为 4，损失函数为交叉熵损失函数（cross-entropy loss function, CELoss）。训练计数采用迭代次数（iterations），即每训练一个 batch size 的数据迭代一次，本文将其设置为 80 000 次。2 个数据集训练和测试的输入尺寸都是 1 024 像素×1 024 像素，采用随机水平翻转、随机转置、随机缩放 (缩放比从 0.5 到 2.0) 作为数据增强策略。本文采用平均交并比 (mean intersection over union, mIoU)、平均 F_1-score(mean F_1-score, mF_1)、参数量 Params 作为模型的评价指标。

2.3　对比模型

本文选取了 8 个模型作为对比，其大致可分为基于 CNN 的模型和基于 Transformer 的模型 2 类，其中基于 CNN 的模型为 PSPNet[7]、DeepLabV3+[11]、CCNet[21]、DANet[20]；基于 Transformer 的模型为 ViT-UperNet[24]、Swin-UperNet[25]、SegFormer[30]、VAN-UperNet[26]，除 SegFormer 之外都采用 UperNet[32] 作为网络的解码器。

2.4　实验结果

Potsdam 数据集上的实验结果如表 2 所示，LoveDA 数据集上的实验结果如表 3 所示，其中各项指标中最好的数据都加粗单独标出。根据表 2 的数据，本文提出的 MSTNet 在大多数评估指标上优于其他模型，仅在参数量指标方面比 SegFormer 稍逊。与 VAN 模型相比，MSTNet 在 mIoU 和 mF_1 指标上分别提高了 0.68% 和 0.63%；与传统的 CNN 模型相比，MSTNet 在 mIoU 和 mF_1 指标上分别提高了约 2%。根据表 3 的实验结果，本文提出的 MSTNet 在 LoveDA 数据集上的部分指标优于其他模型，并且在综合表现上是最均衡的。与基于 CNN 模型相比，基于 Transformer 的模型在 mIoU 和 mF_1 指标上提升为 2%~3%，并且本文提出的 MSTNet 更是在这 2 个指标上取得了将近 4% 的提升。

表 2　Potsdam 数据集实验结果

Table 2　Experimental results of the Potsdam dataset

模型	主干	道路	建筑	草地	树木	汽车	mIoU/%	mF_1/%	Params/10^6
PSPNet	ResNet50	86.53	93.50	76.36	78.00	90.99	77.44	85.88	48.97
DeepLabV3+	ResNet50	86.68	93.31	76.43	78.22	91.04	77.63	86.07	43.58
CCNet	ResNet50	86.49	93.43	76.45	78.13	90.53	77.08	85.55	49.82
DANet	ResNet50	86.24	92.93	76.42	77.98	90.93	77.22	85.73	49.82
ViT-UperNet	ViT-b	85.66	92.68	75.18	73.88	88.18	75.81	84.91	144.06
Swin-UperNet	Swin-s	87.18	93.35	77.45	79.32	90.73	78.39	86.68	81.15
SegFormer	Mit-b2	87.12	93.78	77.92	79.82	90.7	78.56	86.76	**24.72**
VAN-UperNet	VAN-B2	87.23	93.88	77.57	80.04	90.43	78.82	87.03	56.44
MSTNet(本文模型)	VAN-B2	**88.02**	**94.50**	**78.58**	**80.06**	**91.47**	**79.50**	**87.66**	32.85

表 3　LoveDA 数据集实验结果

Table 3　Experimental results of LoveDA dataset

模型	主干	背景	建筑	道路	水	荒地	森林	农田	mIoU/%	mF_1/%	Params/10^6
PSPNet	ResNet50	53.02	63.77	53.16	57.02	25.97	39.39	47.10	48.49	64.43	48.97
DeepLabV3+	ResNet50	51.90	63.25	54.05	58.69	30.08	41.76	46.27	49.43	65.48	43.58
CCNet	ResNet50	53.56	64.66	53.19	61.47	29.07	39.15	47.66	49.82	65.68	49.82
DANet	ResNet50	53.42	64.78	53.42	61.23	29.43	39.39	46.84	50.17	65.91	49.82
ViT-UperNet	ViT-b	52.58	63.00	56.13	**70.27**	29.18	40.04	55.07	52.32	67.72	144.06
Swin-UperNet	Swin-s	54.52	62.24	54.48	66.03	**38.11**	**43.16**	48.45	52.86	68.62	81.15
SegFormer	MiT-b2	53.34	**65.77**	55.44	69.97	34.23	37.97	52.14	52.67	68.14	**24.72**
VAN-UperNet	VAN-B2	52.98	64.65	57.20	70.02	29.24	38.99	**56.42**	52.81	68.08	56.44
MSTNet(本文模型)	VAN-B2	**54.62**	64.92	**57.84**	69.81	33.59	41.77	54.86	**53.92**	**69.14**	32.85

为了验证模型训练参数的合理性和收敛性，本文对经典的 CNN 模型 DeepLabV3+、改进的 Transformer 模型 VAN 以及在 VAN 基础上改进的 MSTNet 模型在 Potsdam 和 LoveDA 数据集上进行了 80 000 个迭代次数的训练，并绘制了它们在准确率曲线和 Loss 曲线上的表现，如图 4 所示。结果显示，在 Potsdam 数据集上，3 个模型的收敛速度相似，但相较于 DeepLabV3+，VAN 和 MSTNet 表现更加稳定。在 LoveDA 数据集上，基于 Transformer 的 VAN 和 MSTNet 相对于 DeepLabV3+表现出较大的优势，不仅在收敛速度上表现更好，而且在准确率上也取得了明显的领先。同时相较于 VAN，MSTNet 在收敛速度和准确率上都有一定的提升。

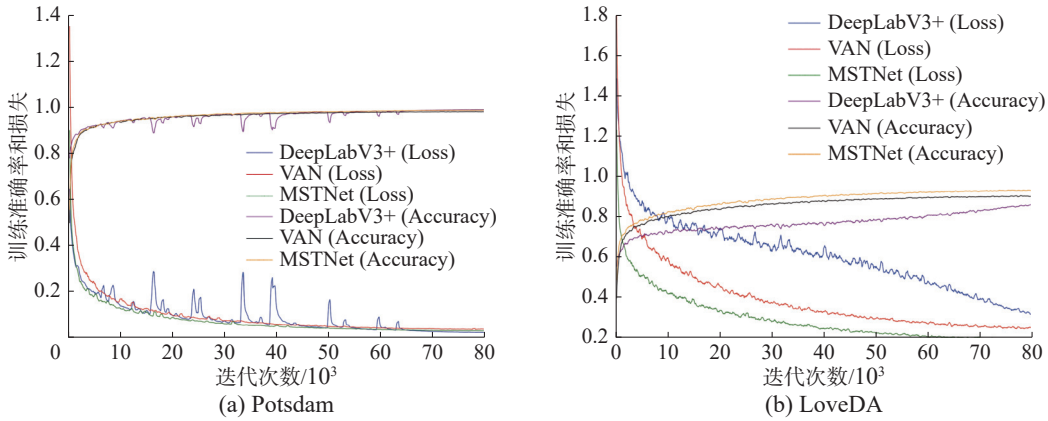


图 4　Postdam 和 LoveDA 数据集上的准确率曲线和 Loss 曲线

Fig. 4　Accuracy curves and Loss curves on Potsdam and LoveDA datasets

2.5　消融实验

为验证各个模块的有效性，分别在 Potsdam 和 LoveDA 数据集上进行消融实验，具体结果如表 4 所示。表 4 中，不含 VAN 代表主干用的是 ResNet50，不含 MSFEM 代表消去这个模块，不含 MLP 代表使用 UperNet 作为解码器。实验 1 和实验 2 分别采用 CNN 主干 ResNet 和 Transformer 主干 VAN 进行实验，结果表明 Transformer 主干表现要优于 CNN 主干，在 Potsdam 数据集上的 mIoU 和 mF_1 指标分别提升 3.28% 和 2.72%，在 LoveDA 数据集上这 2 个指标分别提升 4.12% 和 3.78%。实验 1 和实验 3 的对比与实验 2 和实验 4 的对比都说明 MSFEM 模块给网络带来了性能上的提升。实验 5 采用 UperNet 作为解码器，相比之下实验 4 的 MLP 解码器降低了 30.41×10^6 的参数量，并且提升了模型性能。

为验证 MSFEM 的优越性，选择 ResNet50 作为网络的主干，在此基础上分别添加 SPP、ASPP、MSFEM 进行实验，其结果如表 5 所示。结果显示 MSFEM 的参数量相比于 SPP 和 ASPP 分别下降了 18.66×10^6 和 13.27×10^6。在 Potsdam 数据集上相比于 SPP 和 ASPP，MSFEM 模块在 mIoU 和 mF_1 这 2 个指标上分别提升 0.40%、0.44% 和 0.21%、0.25%。在 LoveDA 数据集上 MSFEM 模块相比于 SPP 和 ASPP 在这 2 个指标上分别提升 1.23%、1.38% 和 0.29%、0.33%。

表 4　VAN 与 MSFEM 消融实验数据

Table 4　VAN and MSFEM ablation experimental data

序号	模块			Params/10^6	Potsdam/%		LoveDA/%	
	VAN	MSFEM	MLP		mIoU	mF_1	mIoU	mF_1
实验1	×	×	√	23.99	75.85	84.52	48.55	64.47
实验2	√	×	√	26.55	79.13	87.24	52.67	68.25
实验3	×	√	√	30.30	77.84	86.32	49.72	65.81
实验4	√	√	√	32.85	79.50	87.66	53.92	69.14
实验5	√	√	×	63.26	78.93	87.25	53.24	68.43

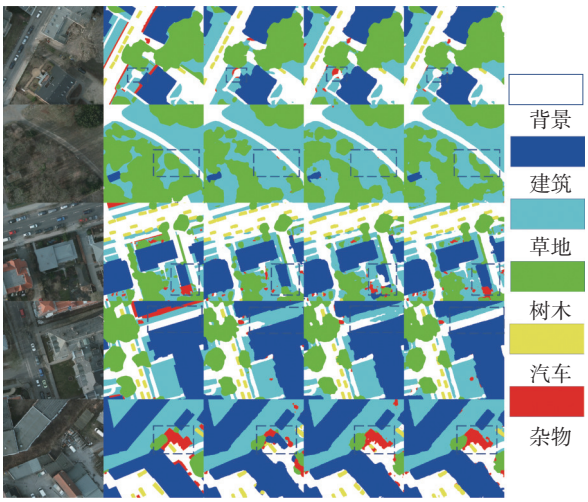
表 5　MSFEM 优越性实验数据

Table 5　Experimental data on the superiority of MSFEM

模型	Params/10^6	Potsdam/%		LoveDA/%	
		mIoU	mF_1	mIoU	mF_1
ResNet50+SPP	48.97	77.44	85.88	48.49	64.43
ResNet50+ASPP	43.58	77.63	86.07	49.43	65.48
ResNet50+MSFEM	30.31	77.84	86.32	49.72	65.81

2.6　可视化分析

为了更好地展示 MSTNet 的性能，本文进行了分割结果可视化对比。具体地，本文选取了 3 种模型进行对比，包括基于 CNN 并且含有多尺度提取模块的 DeepLabV3+模型、基于 Transformer 不含多尺度提取模块的 VAN 以及基于 Transformer 并且含有多尺度模块的 MSTNet。分割结果可视化对比的结果如图 5 和图 6 所示，其中明显改善的地方已经用黑色虚线框标记出来。

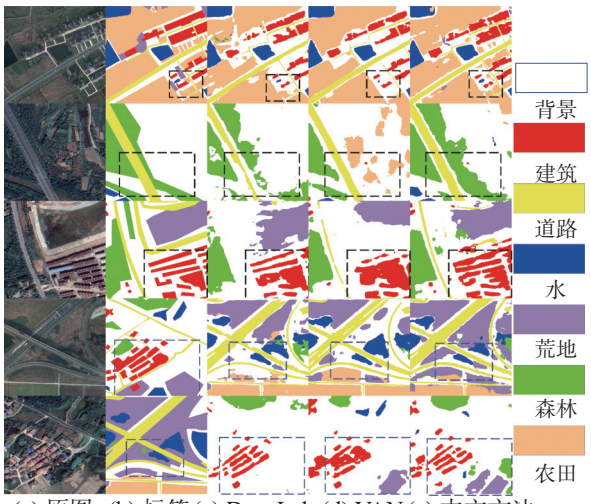


(a) 原图　(b) 标签 (c) DeepLab V3+ (d) VAN (e) 本文方法

图 5　Potsdam 数据集上分割结果对比

Fig. 5　Comparison chart of segmentation results on the Potsdam dataset

图 5 给出了在 Potsdam 数据集上 3 个模型的分割结果对比。由于 Transformer 具有全局上下文提取能力，因此 VAN 和 MSTNet 的总体分割结果要好于 DeepLabV3+。此外，由于 MSFEM 模块的加入，MSTNet 在细节方面要优于 VAN。例如，在第 1 和第 5 行中，VAN 对杂物类的误判；第 3 行虚线框中 VAN 分割结果建筑的部分缺失；第 4 行虚线框中 VAN 对草地的误判等问题都在 MSTNet 的结果中得到改善。



(a) 原图　(b) 标签 (c) DeepLab V3+ (d) VAN (e) 本文方法

图 6　LoveDA 数据集上分割结果对比

Fig. 6　Comparison chart of segmentation results on the LoveDA dataset

图 6 给出了在 LoveDA 数据集上的分割结果对比。在这样较大的数据集上，基于 Transformer 的 VAN 和 MSTNet 的优势更加明显。例如，在第 3 行中，DeepLabV3+的分割结果中道路部分缺失，

而 VAN 和 MSTNet 都能够准确预测。在第 4 行和第 5 行中, DeepLabV3+缺失了荒地类别, 而 VAN 和 MSTNet 都能够预测出部分荒地。此外, 得益于 MSFEM 模块的加入, MSTNet 相对于 VAN 在细节方面具有明显的优势。例如, 在第 3 行和第 5 行中, 对于建筑部分的预测, MSTNet 的分割结果边缘更加清晰, 而 VAN 的分割结果已经连在一起。

2.7　实验结论

1)增强网络全局上下文提取能力的 Transformer 主干有效地提升了模型的分割效果。表 2~4 的实验结果都表明这一点, 图 5 和图 6 的可视化对比也明显展示出 Transformer 模型的优势。

2)多尺度语义特征是语义分割的关键。表 2~4 的实验结果都说明多尺度语义特征提升了分割效果。图 5、图 6 的可视化实验可直观地看出多尺度语义特征显著提升了分割细节。

3)本文改进 ASPP 的 MSFEM 模块, 在参数量、性能、提升模型收敛速度方面有很大提升。表 5 的对比实验可看出 MSFEM 模块有效降低了参数量提升了模型性能。图 4 表明 MSFEM 相比 ASPP 提升了稳定性, 可有效提升模型的收敛速度。

4)MLP 解码器在降低参数量的同时有效提升模型性能。表 2~4 的实验结果表明 MLP 解码器在降低参数量和提升模型性能方面十分有效。

5)本文提出的 MSTNet 在不同数据集上有不错的泛化能力。表 2 和表 3 的数据显示, 在 Potsdam 这样较小的数据集上基于自注意力的 ViT、Swin 等表现不佳, 在 LoveDA 这样的大数据集上表现很好, 这是由于大量数据训练弥补了其缺乏的归纳偏置。本文提出的 MSTNet 在 2 个不同的数据集上都有不错的表现, 有一定的泛化能力。

3　结束语

本文提出了一种基于 Transformer 的多尺度遥感图像语义分割网络 MSTNet, 该网络旨在应对遥感图像语义分割中目标类间方差小、类内方差大的问题, 从而提升分割效果。为实现这一目标, MSTNet 的设计关注提取全局上下文信息和多尺度语义特征这 2 个关键点, 分别引入 VAN 主干和 MSFEM 模块, 在提升分割精度的同时减少了参数量。总体来说, 本文提出的 MSTNet 一定程度上提升了遥感语义分割的精度, 但是由于现在高分辨率遥感数据集较少并且已有数据集的数据量较少, 基于 Transformer 的模型在小数据集上的表现不佳。因此, 下一步的研究方向考虑基于弱监督学习设计遥感语义分割网络来缓解此问题。

参考文献:

[1] 皮新宇, 曾永年, 贺城墙. 融合多源遥感数据的高分辨率城市植被覆盖度估算 [J]. 遥感学报, 2021, 25(6): 1216–1226.
PI Xinyu, ZENG Yongnian, HE Chengqiang. High-resolution urban vegetation coverage estimation based on multi-source remote sensing data fusion[J]. National remote sensing bulletin, 2021, 25(6): 1216–1226.

[2] 高吉喜, 万华伟, 王永财, 等. 大尺度生态质量遥感评价方法构建及应用 [J]. 遥感学报, 2023, 27(12): 2860–2872.
GAO Jixi, WAN Huawei, WANG Yongcai, et al. New framework for large-scale ecological quality evaluation and application research using remote sensing data[J]. National remote sensing bulletin, 2023, 27(12): 2860–2872.

[3] 刘红超, 张磊. 面向类型特征的自适应阈值遥感影像变化检测 [J]. 遥感学报, 2020, 24(6): 728–738.
LIU Hongchao, ZHANG Lei. Adaptive threshold change detection based on type feature for remote sensing image[J]. Journal of remote sensing, 2020, 24(6): 728–738.

[4] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 3431−3440.

[5] 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.

[6] RONNEBERGER O, FISCHER P, BROX T. U-NET: convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference. Munich: Springer International Publishing, 2015: 234−241.

[7] ZHAO Hengshaung, SHI Jiangping, QI Xiaojuan, et al. Pyramid scene parsing network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 2881−2890.

[8] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Semantic image segmentation with deep convolutional nets

and fully connected crfs[EB/OL]. (2014−12−22) [2022−06−07]. https://arxiv.dosf.top/abs/1412.7062.

[9] 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.

[10] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation[EB/OL]. (2017−06−17)[2021−11−12].https://arxiv.org/abs/1706.05587.

[11] CHEN L C, ZHU Yukun, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the European Conference on Computer Vision. Cham: Springer, 2018: 801−818.

[12] 邵凯, 闫力力, 王光宇. 压缩感知重构算法的两步深度展开策略研究 [J]. 智能系统学报, 2023, 18(5): 1117–1126.
SHAO Kai, YAN Lili, WANG Guangyu. Two-step deep unfolding strategy for compressed sensing reconstruction algorithms[J]. CAAI transactions on intelligent systems, 2023, 18(5): 1117–1126.

[13] D'ASCOLI S, TOUVRON H, LEAVITT M L, et al. Convit: improving vision transformers with soft convolutional inductive biases[C]//International Conference on Machine Learning. Vienna: PMLR, 2021: 2286−2296.

[14] ZHANG Cheng, JIANG Wanshou, ZHANG Yuan, et al. Transformer and CNN hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery[J]. IEEE transactions on geoscience and remote sensing, 2022, 60: 1–20.

[15] GAO Liang, LIU Hui, YANG Minhang, et al. Stransfuse: fusing swin transformer and convolutional neural network for remote sensing image semantic segmentation[J]. IEEE journal of selected topics in applied earth observations and remote sensing, 2021, 14: 10990–11003.

[16] 龙丽红, 朱宇霆, 闫敬文, 等. 新型语义分割 D-UNet 的建筑物提取 [J]. 遥感学报, 2023, 27(11): 2593–2602.
Long Lihong, Zhu Yuting, Yan Jingwen, et al. New building extraction method based on semantic segmentation[J]. National remote sensing bulletin, 2023, 27(11): 2593–2602.

[17] LI Rui, ZHENG Sunyi, ZHANG Ce, et al. Multiattention network for semantic segmentation of fine-resolution remote sensing images[J]. IEEE transactions on geoscience and remote sensing, 2021, 60: 1–13.

[18] 王潇棠, 闫河, 刘建骐, 等. 一种边缘梯度插值的双分支 deeplabv3+语义分割模型 [J]. 智能系统学报, 2023, 18(3): 604–612.
WANG Xiaotang, YAN He, LIU Jianqi, et al. A new deeplabv3+ semantic segmentation model of edge gradient interpolation with double branch structure[J]. CAAI transactions on intelligent systems, 2023, 18(3): 604–612.

[19] 李涛, 高志刚, 管晟媛, 等. 结合全局注意力机制的实时语义分割网络 [J]. 智能系统学报, 2023, 18(2): 282–292.
LI Tao, GAO Zhigang, GUAN Shengyuan, et al. Global attention mechanism with real-time semantic segmentation network[J]. CAAI transactions on intelligent systems, 2023, 18(2): 282–292.

[20] FU Jun, LIU Jing, TIAN Haijie, et al. Dual attention network for scene segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 3146−3154.

[21] HUANG Zilong, WANG Xinggang, HUANG Lichao, et al. CCNET: criss-cross attention for semantic segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Long Beach: IEEE, 2019: 603−612.

[22] LI Rui, ZHENG Shunyi, ZHANG Ce, et al. ABCNet: attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery[J]. ISPRS journal of photogrammetry and remote sensing, 2021, 181: 84–98.

[23] WANG Libo, LI Rui, WANG Dongzhi, et al. Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images[J]. Remote sensing, 2021, 13(16): 3065.

[24] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: transformers for image recognition at scale[EB/OL]. (2020−10−22)[2021−01−03]. https://arxiv.dosf.top/abs/2010.11929.

[25] LIU Ze, LIN Yutong, CAO Yue, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Nashville: IEEE, 2021: 10012−10022.

[26] GUO Menghao, LU Chengze, LIU Zhengning, et al. Visual attention network[J]. Computational visual media, 2023, 9(4): 733–752.

[27] WANG Panqu, CHEN Pengfei, YUAN Ye, et al. Understanding convolution for semantic segmentation[C]//2018 IEEE Winter Conference on Applications of Computer Vision. Lake Tahoe: IEEE, 2018: 1451−1460.

[28] BAI Haiwei, CHENG Jia, HUANG Xia, et al. HCANet: a hierarchical context aggregation network for semantic segmentation of high-resolution remote sensing images[J]. IEEE geoscience and remote sensing letters, 2021, 19: 1–5.

[29] HASSANI A, SHI H. Dilated neighborhood attention transformer[EB/OL]. (2022−09−29)[2023−01−16]. https://arxiv.dosf.top/abs/2209.15001.

[30] HASSANI A, WALTON S, LI J, et al. Neighborhood attention transformer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 6185−6194.

[31] XIE Enze, WANG Wenhai, YU Zhiding, et al. SegFormer: simple and efficient design for semantic segmentation with transformers[J]. Advances in neural information processing systems, 2021, 34: 12077-12090.

[32] XIAO Tete, LIU Yingcheng, ZHOU Bolei, et al. Unified perceptual parsing for scene understanding[C]//Proceedings of the European Conference on Computer Vision. Cham: Springer, 2018: 418−434.

作者简介：

邵凯，副教授，主要研究方向为智能感知与信息系统、信号与信息智能处理。发表学术论文 40 余篇。E-mail：shaokai@cqupt.edu.cn。

王明政，硕士研究生，主要研究方向为深度学习、遥感图像语义分割。E-mail：2889220059@qq.com。

王光宇，教授，主要研究方向为数字信号处理、滤波器组理论。出版学术专著 2 部，发表学术论文 30 余篇。E-mail：wangguangy@cqupt.edu.cn。

2024 年度智慧医疗专题学术会议

为贯彻落实《中共中央关于制定国民经济和社会发展第十四个五年规划和二〇三五年远景目标的建议》和《“健康中国 2030”规划纲要》等文件精神，由中国人工智能学会、中国体视学会主办，齐鲁工业大学承办的 2024 年度智慧医疗专题学术会议将于 2024 年 8 月 24 日在山东济宁曲阜尼山圣境召开。

2024 年度会议将邀请国内外智慧医疗领域的著名专家学者参加。为进一步推进智慧医疗技术在医疗领域的深入应用，构建智慧医疗体系和平台助力医院向数字化、信息化转型发展，促进产学研用之间的深度合作，会议计划征集智慧医疗领域的学术论文，被会议录用论文将收录至智慧医疗会议论文集中，还将推荐其中的优秀中文论文到《工程科学学报》《北京邮电大学学报》《智能系统学报》和《中国体视学与图像分析》等期刊，优秀英文论文推荐至 Frontiers in Microbiology（中科院二区 TOP，IF=5.1）AI in Pathogenic Microorganism（病原微生物中的 AI）专题。会议还将评选优秀论文并在会议期间颁发奖励证书。欢迎智慧医疗领域的专家学者、专业人士和研究生踊跃投稿，欢迎相关企业投稿并做报告。

重要日期：

全文投稿截止：2024 年 6 月 23 日

全文录用通知：2024 年 7 月 15 日

正式会议日期：2024 年 8 月 24 日

征文范围包括（但不局限于）：

面向医学应用的人工智能；

主动健康与老龄化科技应对；

深度学习及智能辅助诊断技术；

病原微生物中的 AI；

医学大数据和人工智能实践与案例分析等相关方向。

投稿邮箱：

caai_impc_forum@163.com