## Abstract

主流的深度学习方法极大影响了地表覆盖物分类的准确性。然而采集到的遥感影像被雾气污损时，分类的鲁棒性性能通常会有大幅度下降，这是图像质量较差且存在有大量遮挡等一系列因素所导致。不同于传统的、仅探索模型识别准确率的研究工作，本篇文章的主要内容是基于注意力机制与多模态融合来提升模型分类的鲁棒性。我们采用ConvNext作为骨干网络。在这个基础上，我们基于非局部的注意力机制提出并构建了CREM与CMFM模块。其中的CREM拥有较大的感受野来融合局部与全局的特征信息，从而降低图像中冗余噪声造成的影响。CMFM探求了多模态输入间的交互关系，从而进行信息的矫正。为了验证所提方法的有效性，我们在ISPRS Potsdam与Vaihingen基准数据集上进行了扩展性的消融与对比实验。在污损数据集上的实验结果表明，相比于基准模型，我们的方法在地表覆盖物分类时展现出了卓越的精准度与较强的鲁棒性。

## Introduction

对遥感图像中的地表覆盖物进行精准且鲁棒的分类（a.k.a. 语义分割），是进行一系列遥感任务的前提条件，例如A\B\C。这里的鲁棒性是指，当模型面对复杂的气象条件时（例如，雾气），性能相较于干净的环境时，分类性能损失较少。近年来，一大批学者将深度学习方法应用于遥感领域并取得了较好的效果。基于此的LCC方法大致可以分为四类，分别是FCNs，UNets, HRNets与ViTs。尽管这些方法在干净环境中，对地面覆盖物识别率较好，然而当环境中出现有雾气等污损情况时，模型鲁棒性往往无法得到保证。

图1中反映了在雾气环境中进行LCC时经常会遇到的种种挑战。 图1（a）中的车辆的形状大小有明显的差异。雾气干扰，使得（b）（d）（e）中框内异类物体呈现出高度的同质性。（c）中展示了被树木与雾气遮挡的半个车辆。而由于阴影与浓雾的遮盖，（f）中不同车辆的外观有了明显差异，很难将其与clutter相区别。

为了能较好地应对雾气下进行LCC时遇到的各种困难与挑战，我们构建了鲁棒的LCC模型。Backbone采用的是在图像较差质量环境下表现较好的ConvNeXt，同时我们提出了 CREM与CMFM两大模块以提升鲁棒性。CREM由局部与全局两大分支构成，CREM有着较大的感受野，能够很好地融合全局与本地特征以减少输入中的冗余噪声。基于非局部的结构，我们能提出的CMFM，可以对多模态输入进行融合修正，以此来挖掘多模态间的交互关系。总的来说，我们的贡献可以总结成如下的几个方面：

* 我们设计了端到端的鲁棒LCC模型，本模型在污损的雾天环境下性能表现较好；
* 我们提出了CREM 与CMFM，减少污损图像中的噪声对最终性能的影响；
* 我们生成了雾气下的污损遥感数据集来进行模型鲁棒性的验证。

## backbone and decoder

从图1可以看出，我们的框架首先组装两种模态的数据并将其送入ConvNeXt进行编码，从而得到四种分辨率下的特征图。**其中的低级特征具有较多的细节信息但却包含有大量冗余噪声，高级特征有丰富的语义信息但缺少了对细节边缘的优化。** 我们利用CREM来融合局部与全局特征，利用MRFM来对多模态输入进行有效的融合。得益于CNN内置的归纳偏置的能力与平移不变性，和ViT相比，ConvNeXt·有着更强的鲁棒性。ConvNeXt沿用了ResNext的设计理念，利用分组卷积来拓展模型的宽度。从论文12的综合性比较中可以看出，ResNeXt能够很好地应对污损后的图像。 同时装备的大核卷积使其具有较大的全局感受野。于是我们使用ConvNeXt来作为模型的backbone部分，使得主干网络能够很好地处理低质量遥感图像。解码器部分我们则选用UperNet，他采用了一个自顶向下的结构，通过较小的运算量并通过一个固有的金字塔式结构来将高层的语义表征融合进中底层去。如此一来，它可以统一分析不同层级下的语义表征信息，从而增强模型对复杂场景理解的鲁棒性。

According to the framework overview in Figure 1, both modalities are firstly assembled into ConvNeXt for encoding, which generates feature maps with four various resolutions. Low-level features are more exhaustive but contain redundant noise, whereas high-level features are rich in semantic information but do not depict edges delicately. CREM is proposed to fuse local and global features, while MRFM is conducive to complementary representations extractions from multimodal inputs. In view of inherent inductive bias and translational invariance of CNN, ConvNeXt, which follows the ResNext design philosophy of using group convolution to extend model width, is more robust than ViT. According to the comprehensive comparison in paper 12, ResNeXt is capable of handling corrupted images well. ​Furthermore, equipment of the large kernel convolution provides a large and global receptive field. Therefore, we adopt ConvNeXt as the backbone, which allows the network to effectively cope with low quality RSIs. UperNet is selected as the decoder, which employs a top-down lightweight structure to exploit hierarchical levels of representations through the combinations of parsing visual attributes. Semantic representation information can be unified, improving robustness in understanding complex remote sensing scenarios.

## CREM

如果直接将雾气下采集到的、污损后的多模态数据送入多模态融合模块，会给网络带来较多的冗余和噪声。为了提升多模态间的互补性与兼容性，降低噪声对鲁棒性的干扰，我们设计了由局部与全局分支融合而成的语义表征增强模块来增强模型提取有用信息的能力，其具体结构如图2所示。总体上看，习得的全局表征能很好地适应图像中光照与可视角度的变化，而局部表征对边缘类的细节信息、几何形态与材质更为敏感。于是，将两种分支的特征进行融合可以很好地获得互补性信息，从而对噪声进行滤除与修正。其中的局部分支由深度可分离的ASPP(DW-ASPP)、空间与通道大核注意力与残差模块构成。由于DS-ASPP中各个并行的空洞卷积有着不同的空洞率，它能够以很小的计算量来高效应对遥感图像中各类物体的尺度多样性问题。各层的DS卷积是由depthwise与pointwise卷积构成，深度卷积对每个输入通道分别进行空间卷积，而点卷积将深度卷积的输出结合起来。DW-ASPP输出结果concat后通过1×1的卷积进行embedding操作。原始的自注意力机制忽视了各通道的自适应性，而SCLA由三种卷积串联而成，同时与embedding后的结果做元素的乘积，可以对每个通道、每个位置的特征图重要性进行建模，从而很好地捕获RSIs中的长距离依赖关系。输出结果经由1×1的卷积进行通道对齐操作，与原始输入特征相加后经过残差模块来降低数据中信息的冗余度。

## MRFM

传统的多模态融合算法往往会忽视模态间的独特特性与联系，从而无法有效地融合多模态数据。我们设计的MRFM模块，基于自注意力机制可以很好地针对不同位置的输入来触发相应的表征联系这一特征。但不同于单纯的自制力机制，我们很好地结合了RGB输入中丰富的上下文信息与DSM输入中稀疏的高质量高度信息，从而有效地挖掘相邻上下文间的表征联系，使得输出的特征图中汇聚了较强的双模态表征能力。MRFM具体结构如图所示。

## EXP

为了比较所构建架构中各模块的有效性，我们对各部分进行了消融实验。表格1是关于CREM的，我们单独加入各个元素然后融合进backbone中，我们选定A作为reference。 结合表格1中的A、B可以得出，Convnext在进行下游语义分割任务时，要比作为SOTA的Segformer，在分类精度与鲁棒性方面分别有3.42%与2.78%提升。横向对比表格中的A，C，D，E各个组成部分后可以发现，CREM中的SCLA对增强鲁棒性与分类精度方面影响最大（CD减少了13.86%）。而最终融合了所有组件的模型，有着相对最好的性能表现（相较于reference，分类精度提升了3.19%，而CD降低了17.49%）。表格2是关于CMFM的。SNL（simple nonlocal）是GCNet中的简化非局部模块。FM为图2中蓝色虚线框内关于多模态交互融合的部分。Concat将RGB与DSM简单地联结在一起。QE在图2中的query的部分多加入一个类似于Key与Value Embedding部分的组件。表格2中的A、B反映出，SNL的结构可以有效提升模型在clean与fog情况下的分类精度同时增强鲁棒性。而从BC、BD的数据中可以得出，FM相较于单纯的联结多模态输入，其在污损环境下分类精度提升了4.07%，鲁棒性提升了8.17%。E与F差别在于E中多出的一个Query embedding结构，虽然clean下有0.2%的提升，但是多出了0.83GLOPs且CD并无降低。综上所述，我们的CMFM很好的平衡了模型大小、精度与鲁棒性各个方面。

图5展示了模型应对RSI中各种污损时的性能表现。其中，\*代表的是训练集中同时存在有污损与干净数据，训练阶段使模型同时接触两种场景能有效提升其鲁棒性。同时，模型应对高斯模糊时的能力较强，而在雪天场景下的分类能力较弱。如图6，我们利用Seg-GradCAM生成了针对于Car这一类别的特征图。从图中可以看出，融合有两部分的模型明显增强了捕捉远距离依赖关系的能力与泛化性，从而减少了污损环境下的性能损失。

Comparison experiment

表格3中展示的是我们进行的对比实验。对于数据集中标注的5个类别物体，模型对于Car的性能提升最大，相比于综合表现第二名的Swin-Unet，模型对于污损图像的分类精度提升了5.64%，而CD值降低了7.93%。对比表格中的CNN与Transformer为基础的方法，我们发现Transformer类的方法往往在clean下表现较好，而在Fog下性能损失较为严重，这是因为CNN方法有较好的归纳偏置，而ViTs需要有较多的数据作为预训练以提升泛化性。而我们的模型相比于其他的SOTA，则在精度与鲁棒性方面都有着较好的表现。