Hybrid-Gird：遥感图像细粒度分类可解释方法

朱凯雯1，尤亚楠1，孟钢2

1. 北京邮电大学 人工智能学院, 北京 100876）
2. 北京市遥感信息研究所, 北京 100192）

**摘 要 ：目的** 目前，基于遥感图像的目标细粒度分类网络已经在军事和民用领域得到广泛应用，如何对其内部决策机理进行可解释性研究是当前细粒度算法进一步提高决策依据可信度的关键问题，本文针对目标细粒度分类网络的可解释性分析方法展开了研究。**方法** 本文检验了当前主流的可解释性方法在遥感图像目标细粒度分类网络上的适用性，针对其局限性提出了一种基于网格的细粒度分类本质特征可解释性分析方法，通过像素级特征提取算法、高层特征提取算法和基于网格的特征融合解释算法，提高其对网络决策依据本质特征的精确定位能力。**结果** 通过对Hybrid-Grid及其他可解释性分析方法在ADCC等量化评估指标上进行了比较试验，我们提出的可解释性分析方法取得了最优的效果。**结论** 该解释器能够更准确地揭示了以EFM-Net为例的目标细粒度分类网络的工作原理和视觉注意区域的细粒度细节。

**关键词：**人工智能、可解释性分析方法、目标细粒度分类网络

#### 1 引 言

目标细粒度分类任务是遥感图像智能解译领域的一项重要任务，基于光学遥感图像的目标细粒度分类技术广泛应用于军事和民用领域，例如空域管制、态势评估、重要港口和目标监视等（Zhang等，2023）。不同于大类别的目标分类（Chen等，2016），目标细粒度分类旨在区分从属的对象类别，如不同种类的船舶（Han等，2021）。目标细粒度分类任务的关键在于分类模型能否获得对目标类别具有决定性贡献的本质特征，而非与目标类别相关但贡献度较低的其他特征。近年来，随着深度学习的迅猛发展，许多先进的目标细粒度分类网络被提出，并取得了显著的性能提升。已有的工作（Nie等，2022；Xiong等，2022；Liang等，2020；Fu等，2019）主要基于深度卷积神经网络，即DCNN，通过优化特征提取能力和增强特征来提升模型的性能。我们提出的目标细粒度分类网络EFM-Net（Yi等，2023）[10]利用卷积神经网络的特性和原理，通过局部特征提取、注意力机制、局部区域特征融合以及迁移学习等策略，学习到对细粒度类别特定的本质特征表示，并利用这些特征进行准确的分类，在FGSC-23、FGSC-43、Aircraft-16等具有挑战性的目标细粒度分类数据集上取得了优异的结果。

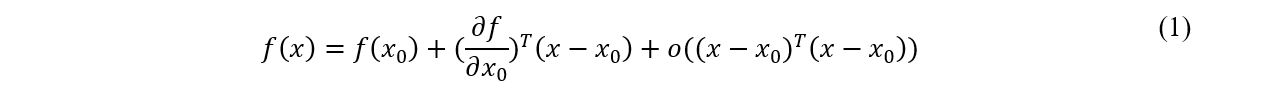
尽管这些网络提升了遥感图像目标细粒度分类任务的精度，但深度学习网络模型的“黑箱”性质使人类很难理解网络是如何做出细粒度分类决策的，更无法确认分类网络是否抓住了对任务最关键的本质特征。这种局限性限制了我们对分类结果的理解和信任，同时也阻碍了目标细粒度网络在军事、无人驾驶等领域的可信应用。围绕遥感图像智能解译任务，当前国内外针对深度神经网络的可解释性分析方法可分为激活值最大化分析法、代理模型分析方法、归因分析法、扰动分析法、类激活图分析法及样例分析法等６类方法（Gong等，2022）。在将现有基于自然影像的可解释性算法应用于遥感影像解译分析任务的研究中，发现IG、GuidedBP、SmoothGrad、occlusion等主流可解释性分析方法存在遥感影像大型场景适用性有限、适用任务范围狭窄等问题（Gong等，2022）。在针对图像分类的深度模型可解释性研究中（Yang等，2022），局部可解释性算法如SmoothGrad，可以通过单一像素的扰动提供细粒度的解释，但解释范围相对较窄，不能关注全局信息。相比之下，CAM系列可解释性方法可以提供更广泛的解释范围，如Grad-CAM（Ramprasaath R. Selvaraju等，2017），它通过梯度反向传播考虑了整个图像的全局信息，从而提供对模型决策的整体解释。因此，在EFM-Net中（Yi等，2023），我们采用Grad-CAM方法（Selvaraju等，2017）对骨干网和多层次注意模块生成内在特征的激活图，并对目标细粒度网络内部进行解释。但我们发现，由于分类网络最后一层卷积层的尺度较小，CAM系列（Perdew等，2016）的可解释性方法将其映射回原图时会导致显著图丢失细节信息，难以生成较为清晰的特征边缘。

总的来说，现有解释器对细粒度特征的解释能力有限，难以理解网络对图像中目标细节信息的关注方式，不能满足解释目标细粒度分类网络的需求。为了更好地刻画目标细粒度分类网络获取的本质特征及其图像域范围，亟需研究更细致的可解释性分析方法来检验分类网络捕捉和利用本质特征做出分类决策的推断逻辑。因此，本文围绕目标细粒度分类网络EFM-Net展开研究，提出了一种针对目标细粒度分类网络的遥感图像细粒度分类可解释方法，该方法沿用CAM系列的思想，通过前向激活图的加权组合得到网络关注的区域性信息（Selvaraju等，2017），使用类激活函数定位重要像素（Baehrens等，2010；Simonyan等，2013），通过添加高斯噪声并对灵敏度图求平均得到网络关注的像素级信息（Smilkov等，2017），最后，使用基于网格的方式对网络关注的信息进行过滤并聚合为特征。该可解释性分析方法通过更直观的可视化结果解释本质特征在空间中的形态，探究网络对细粒度特征的理解，为遥感图像目标细粒度分类领域的可解释性研究和可信应用提供方法参考。

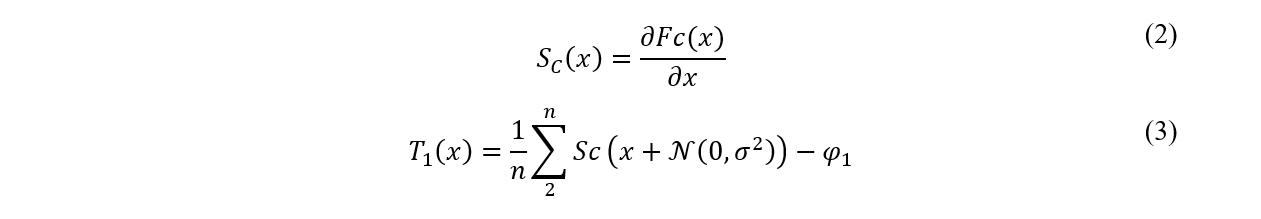
#### 2 研究方法

##### 2.1 像素级特征提取算法

当前与遥感图像中智能解译任务相关的可解释性分析方法中，主要有归因分析法、扰动分析法和类激活图分析法生成显著图来反映影像中不同像素对推断结果的影响，归因分析法主要通过传播梯度信息构成显著性图。基于此，像素级特征可解释性算法以像素为基本单位来解释模型对输入图像的决策行为（Yang等，2022），通过为输入图像中的每个像素分配重要度值或输出贡献值，得到一个与输入图像大小相同的热力图，反映像素级特征对分类网络决策的重要程度。Saliency Map算法（Simonyan等，2013）将深度模型的输出对输入图像的梯度作为解释结果，由于深度模型的复杂度和高度非线性，该方法存在噪声问题、梯度饱和现象、梯度连续性差等缺点。平滑梯度算法（Smilkov等，2017）目前广为应用于改进原始梯度存在的视觉扩散效应，该方法通过引入高斯噪声，在局部邻域内多次采样求平均值来消除噪点，得到平滑锐化的效果。在泰勒展开的框架下，为分类器，为输入图像，为局部邻域内的参考点（Yang等，2022），像素级特征提取算法可以归纳为：

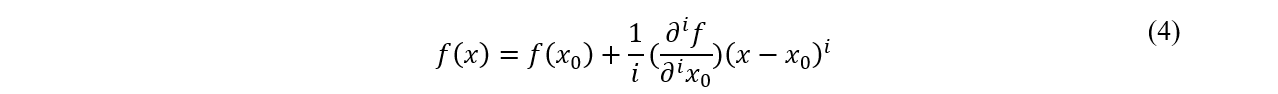


基于此，为了对目标细粒度分类网络进行像素级解释，本文提出的可像素级特征提取算法首先将类激活函数对输入图像求偏导得到灵敏图，对灵敏图添加噪声进行采样，对每个采样图像的结果灵敏度图取平均值，经阈值得到基于像素的目标区域T1，其公式为：

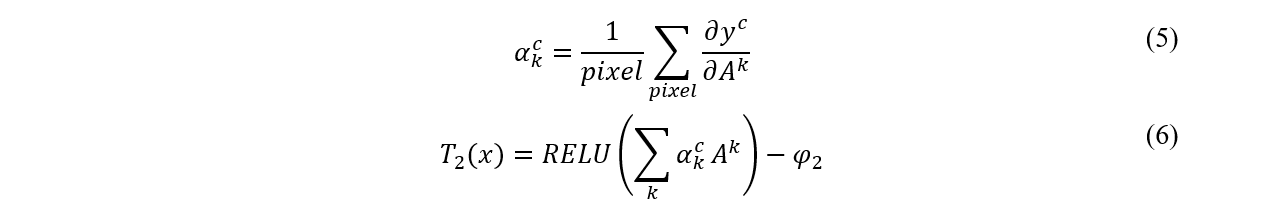


##### 2.2 高层特征提取算法

2017年，Ramprasaath R. Selvaraju等人提出基于梯度的弱监督定位视觉可解释性分析方法，通过计算模型预测类别与特征图之间的梯度信息生成一个热力图来可视化模型对不同特征的注意力分布，该方法能够直观地显示模型对决策的关键特征区域。在泰勒展开的框架下，高层特征提取算法可以归纳为（Deng等，2023）：



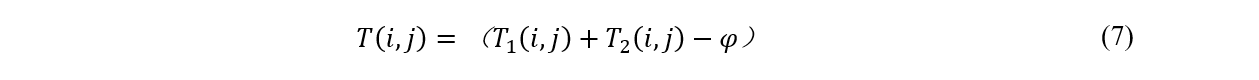
基于此，为了获取网络高层神经元所关注的特征，本文提出了高层特征提取算法，通过对图像分类结果预测类别进行反向传播，利用特征层上的梯度信息计算特征层每个通道的重要程度，根据重要程度对特征层每个通道进行加权求和，经阈值得到基于全局的目标区域T2。其公式为：

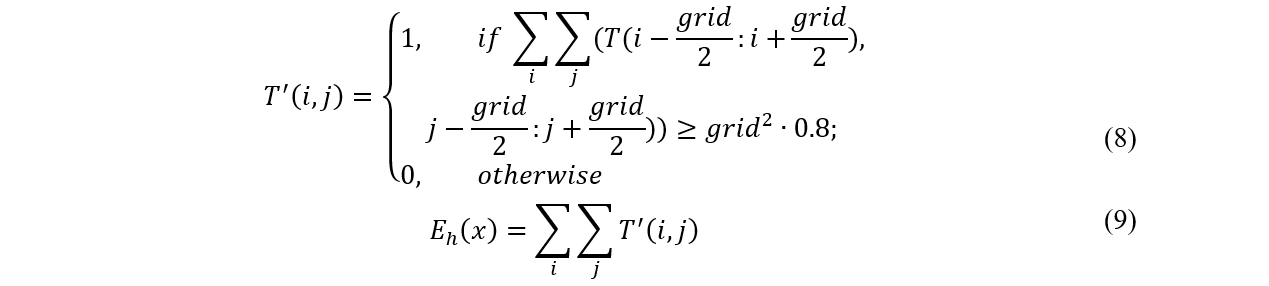


##### 2.3 基于网格的特征融合解释算法

由于特征提取算法都是基于梯度构建，而深度模型的复杂度和高度非线性导致梯度存在梯度噪声、梯度饱和、连续性较差等问题，根据Simonyan等人2014年提出的思想，梯度可视化的视觉效果可以通过对原始梯度进行阈值过滤等操作提升。

因此，本文提出了基于网格的特征融合解释算法，在特征融合过程中经阈值过滤保留对目标细粒度分类最关键的像素级信息，利用网格法将像素点聚合为包含像素级信息和高层信息的本质特征，最终输出特征显著图，该过程如公式（7）-（9）所示。





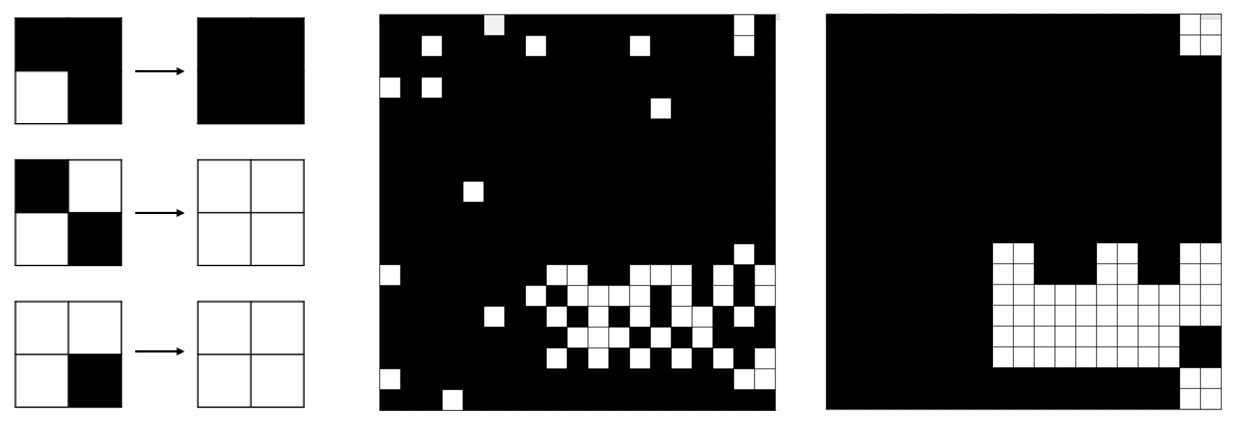


图1 基于网格的特征融合解释算法示意图

以图像尺寸为224x224、遮挡格子尺寸为2x2、阈值为0.5为例，在其遍历过程中，有部分像素点被丢弃，部分像素点被补充，像素点总数基本保持不变。背景中的分散像素点被忽略，突出了前景的目标特征区域，保证了显著图的简洁性和有效性。其原理如图1所示。

##### 2.4 基于网格的本质特征可解释性分析原理

本文提出的基于网格的本质特征可解释性方法主要由三部分组成，分别为2.1节提出的像素级特征提取算法，2.2节提出的高层特征提取算法，以及2.3节提出的基于网格的特征融合解释算法。

在2.1节提出的像素级特征提取算法中，由于神经网络中存在不连续激活函数（如ReLU），直接计算梯度值得归因分析法会引入大量噪声，如何平滑梯度、减少噪声是归因分析法当前最主要的问题（Gong等，2022）。目前，有平均梯度（Smooth-Grad）、平均均多张基于带噪声输入的显著性图（Integrated-Gradients，IG）等方法，基于此，本文于2.3节提出了一种基于网格的特征融合解释算法，根据一定区域内像素级信息密度平均像素的重要度值，达到减少显著性图中的噪声、平滑区域内贡献值快速波动的效果。同时，该算法通过以格子为单位对区域内像素级信息密度进行判定的方式，解决了Smooth-Grad、IG等归因算法显著图像素点过于分散，难以形成“特征的概念”的问题。

本文提出的基于网格的本质特征可解释性方法还结合了类激活图分析法，在2.2节提出的高层特征提取算法中，通过类激活映射方法获得的类似显著性图的网络注意力图引入高层信息，将模型学习到的特征投射回输入像素空间。由于基于梯度的类激活映射技术需要给定卷积层（通常为最后一层）的梯度信息和特征图，而目标细粒度分类网络最后一层的卷积层长宽尺度较小，因此类激活映射分析方法将热力图映射回原图时会导致显著图丢失细节信息，难以生成较为清晰的特征边缘。以EFM-Net为例，其最后一层卷积层输出的特征图尺度为7x7，而输入图像尺度为224x224，则映射回原图时高层信息以224/7=32个像素为单位，以32x32的尺度反应高层信息。由于高层特征提取算法关注的是特征图级别的注意力，它在提供像素级解释方面相对较弱。基于此，2.3节提出的基于网格的特征融合解释算法通过指定区域尺度控制了生成显著图的细度，可以设置特征融合解释算法的网格尺度为输入图像尺度的任意因子。以输入图像尺度为224x224为例，网格尺度可以设置为2x2，4x4，7x7，8x8，14x14，16x16等尺度。当网格尺度为2x2时，本文提出的遥感图像细粒度分类可解释方法可以在显著图中指示细度为输入图像尺度0.9%的特征信息，因此，该方法生成的显著图具有清晰的特征边缘，能够保证对目标细粒度分类任务的高精度解释。

##### 2.5 目标细粒度分类网络

EFM-Net是目标细粒度分类领域目前性能最优的模型之一，由基本特征挖掘模块（Miner）和数据增强模块（Refiner）组成，其结构如图3所示。根据Miner中的特征流，ConvNext主干在内在特征提取阶段从训练样本中提取内在特征。多个层次注意模块（MHAM）通过增强内在特征得到注意特征。将上述得到的特征导入双线性聚合池（BPP）和特征映射网络（FMN），最后生成基本特征，即Miner的核部分进行特征增强。在训练阶段，Refiner可以自适应地增强训练样本，在进一步挖掘图像中的目标特征方面发挥着重要作用。Miner和Refiner可以以无监督的方式相互加强以提取准确的特征。

EFM-Net使用了领域内先进的特征提取与融合方法，包括图像金字塔网络来提取更具代表性的特征（Zhang等，20），结合低级细节、纹理特征和高级语义（Zhang等，20），使用自注意力机制模块增强融合特征（Han等，20；Nie等，20），帮助网络更准确地定位判别区域（Zhang等，20；Liang等，20），抓住对细粒度分类任务最关键的本质特征。由于其结构具有代表性，因此，本文以EFM-Net为研究对象，探究深度学习可解释分析方法对目标细粒度分类网络的理解过程。

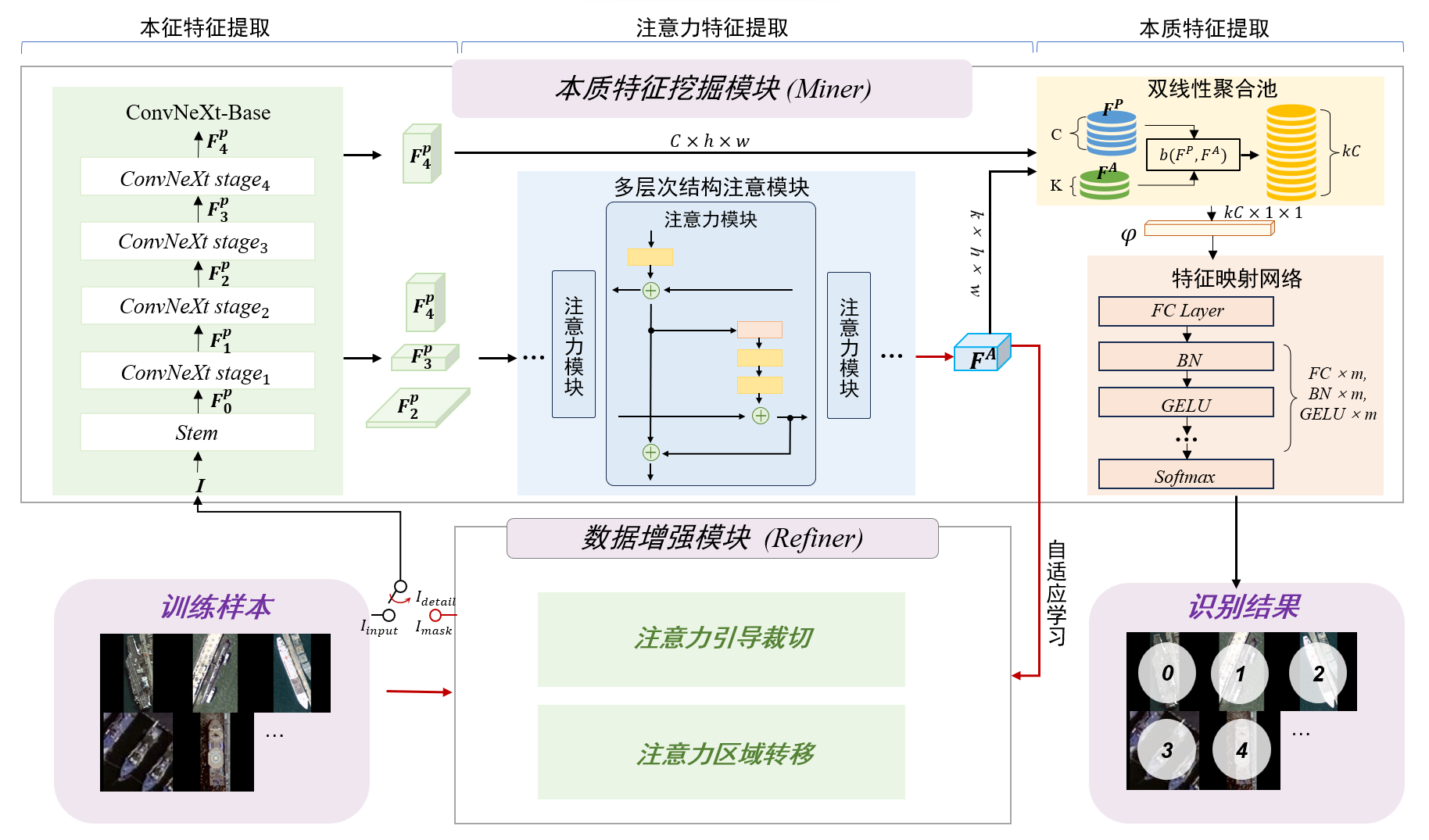


图2 EFM-Net网络结构示意图

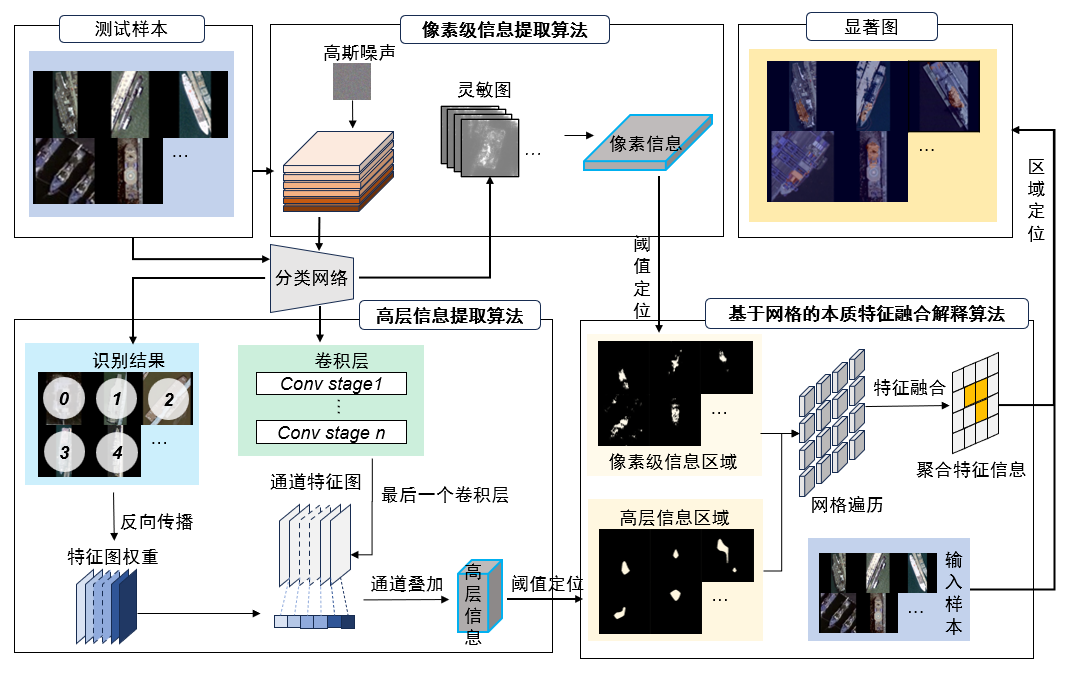


图3 遥感图像细粒度分类可解释方法结构图

#### 3 实 验

##### 3.1 实验数据集

FGSC-23数据集共23类细粒度船舶类的样本。Aircraft数据集共16类细粒度飞机样本，包括云遮挡、小目标和不平衡样本等困难情况。我们在训练集上分别训练了EFM-Net，并在测试集上验证EFM-Net的解释情况，测试集样本分布均衡度如表1所示。

FGSC-23和Aircraft-16数据集的数据样本示例如图4所示。其中，图4（a）、图4（b）分别为FGSC-23和Aircraft-16数据集的图像样本示例。

表1 FGSC-23和Aircraft-16测试集样本分布数量

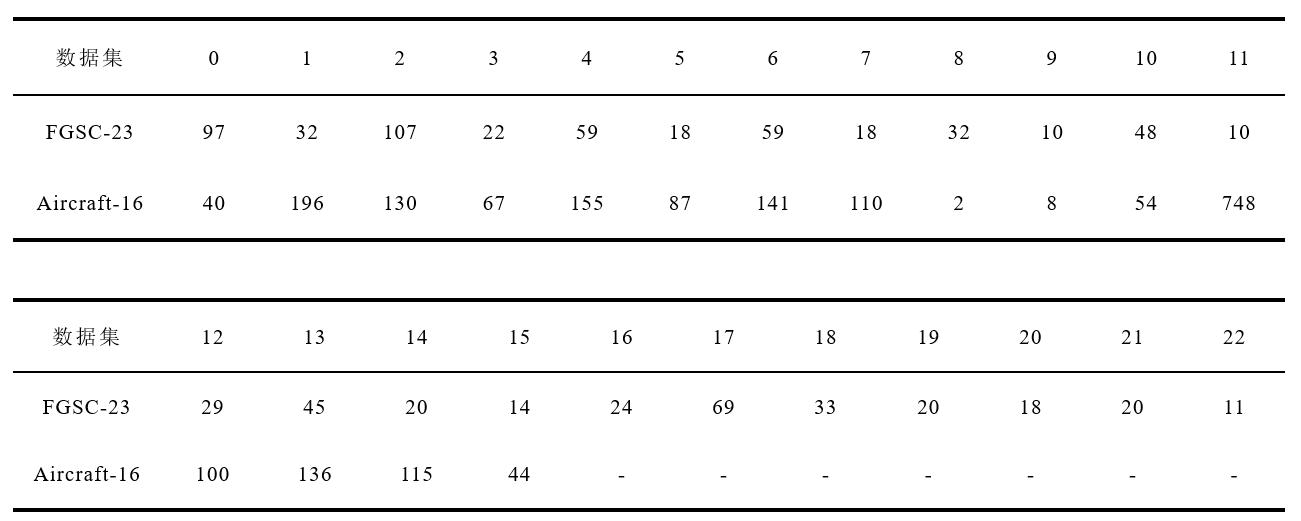




图4 FGSC-23与Aircraft-16数据集样本示例

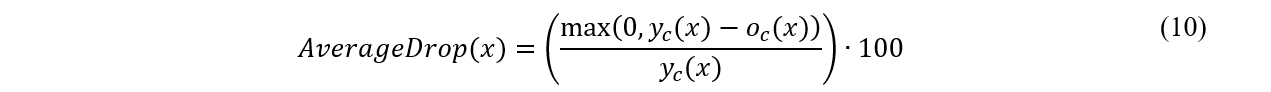
##### 3.2 实验平台

本文实验操作平台为 Ubuntu 18.04 操作系统，使用 CUDA11.3 和 cuDNN8 加速训练，处理器为Intel® Xeon® Silver 4214R CPU@2.40 GHz，GPU 为 NVIDIA GeForce RTX 3090（24 G显存）。使用的编程语言为Python，开发框架为Pytorch。EFM-Net算法的框架见开源链接https://github.com/JACYI/EFM-Net-Pytorch。

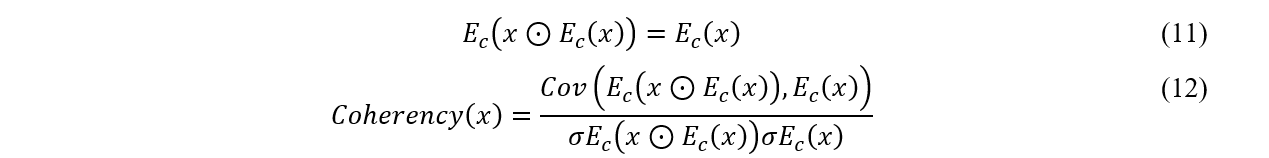
##### 3.3 评价指标

为定量评价解释器的性能，采用平均下降（Average Drop）、一致性（Coherency）、复杂度（Complexity）、ADCC（Average DCC）、删除和精度损失（Deletion and precision loss）作为解释器的评价指标（Samuele Poppi等，2021）：

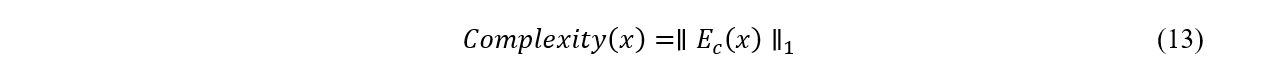
（1）平均下降（Average Drop）：该评价指标衡量了当模型仅看到解释图而不是完整图像时，目标类别c的置信度平均下降的百分比。其公式为：



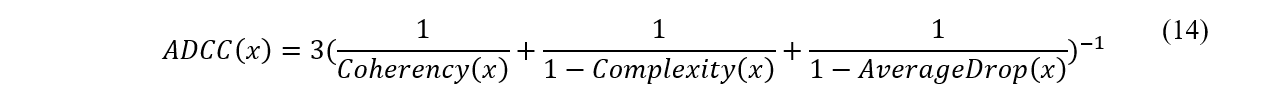
（2）一致性（Coherency）：该评价指标用于衡量解释器的显著图在解释预测时是否包含所有相关特征并以连贯的方式去除无用特征。因此，对于给定的输入图像x和感兴趣的类别c，x的显著图在以显著图本身为条件时不应发生变化。其公式为：



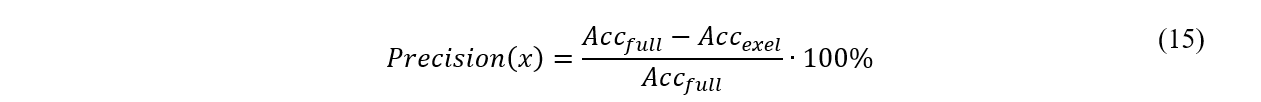
（3）复杂度（Complexity）：该评价指标要求显著图尽可能简单，即它必须包含最少的像素来解释预测结果。当归因方法仅突出显示少量像素时，复杂度最小化。这个评价指标有助于评估解释器的简洁性，即使用最少的像素来解释预测结果。其公式为：



（4）ADCC（Average DCC）：该评价指标提出于论文《重新评估用于解释性的类激活映射：一种新的评估指标和实验分析》（Samuele Poppi等，2021年），用于对解释器进行解释性评估。ADCC综合考虑了平均减少、一致性和复杂度等多个方面，从而提供了对解释器可解释性的全面和公正的评估。其公式为：



（5）删除和精度损失（Deletion and precision loss）：该评价指标衡量在从图像中移除重要特征（由解释器提供）后，分类器精度的下降程度。该指标的数值越大，表示删除的特征对分类器的预测精度影响越大，反映了模型对于关键特征的依赖程度。其公式为：



#### 4 讨论与分析

##### 4.1 目标细粒度分类网络的可解释性分析

EFM-Net是目标细粒度分类领域目前性能最优的模型，使用了领域内先进的特征提取与融合方法，其结构具有代表性。因此，在开源算法库的基础上，以我们提出的EFM-Net为例，探究当前深度学习可解释性分析方法对目标细粒度分类网络的解释效果和适用性。

EFM-Net包含FPN（Feature Pyramid Network）、MHAM（Multi-Head Attention Module）和BPP（Bi-directional Prediction Pyramid）等模块，为了逐步解释EFM-Net模型的决策过程，我们使用Grad-CAM（Ramprasaath R. Selvaraju等，2017）从骨干网和多层次注意模块注意的特征生成类激活图。如图5所示，从左到右分别为骨干网、多层次注意模块和双向预测金字塔模块的类激活图，其中红色和蓝色区域分别代表我们模型的注意力和抑制。从激活图中，我们可以观察到，EFM-Net骨干网所关注的区域离散松散，伴随着较多的噪声，对小目标的特征不敏感，而多层次注意模块注意的特征则集中在我们感兴趣的区域，使用通道间注意力有助于关注重要区域、去除干扰。因此，在Miner中使用多层次注意模块可以很好地解决噪声干扰和目标损耗小的问题。此外，这两种特征融合后产生的本质特征包含了两者的优点。根据EFM-Net在不同模块和阶段下的注意力分布变化，可以证实EFM-Net决策的可信性和可靠性。

本文还使用现有的主流归因分析法对EFM-Net的全局决策进行解释，包括Smooth-Grad、Deep-Lift和Integrated-Gradients，其解释效果如图6（b）-（d）所示。我们观察到EFM-Net在决策过程中对目标的关注区域具有一致性，具体而言，对于船舶细粒度分类任务，以FGSC-23数据集为例，EFM-Net主要依赖于船舶的停机坪、舰岛、相控阵雷达等区域进行判断。类似地，在飞机目标细粒度分类任务中，以Aircraft-16数据集为例，EFM-Net主要依赖于飞机的机翼、发动机等区域进行判断。这与人类的判断依据相吻合，因此，我们可以确定EFM-Net在分类过程中关注的特定区域包含了目标的本质特征，从而深入理解网络的决策过程。

通过使用Smooth-Grad和Grad-CAM等解释方法，应用于EFM-Net，在FGSC-23和Aircraft-16数据集上进行解释并生成显著图，遮挡显著图指示的重要区域并生成遮挡样本以进行遮挡实验，其遮挡样本如图4（b）-（d）、（f）-（e）所示。

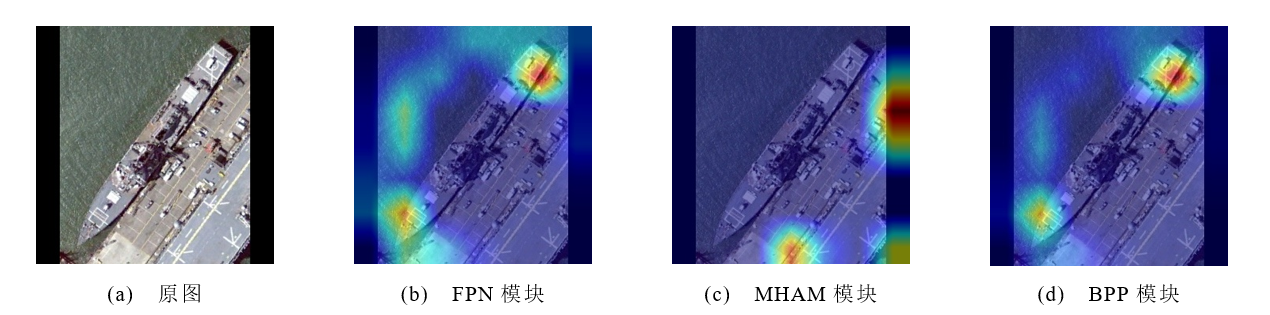


图5 Grad-CAM解释EFM-Net模块

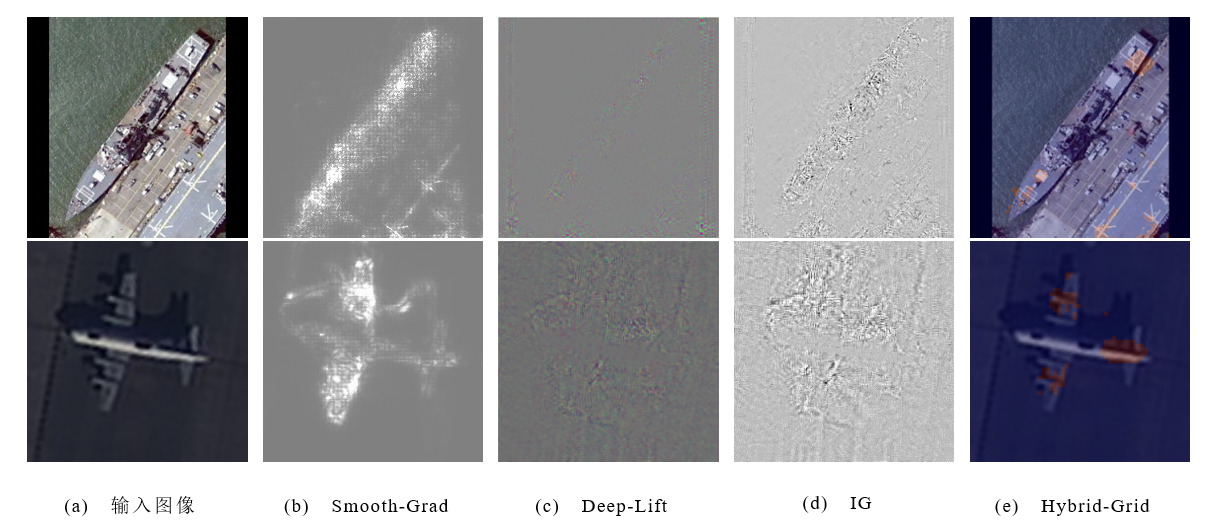


图6 不同解释方法对EFM-Net的解释效果

##### 4.2 基于网格的本质特征可解释性分析效果评估

在第2节中，我们已经论述了本文所提出的遥感图像细粒度分类可解释方法具有原理上的优越性，它通过像素级特征提取算法、高层特征提取算法以及基于网格的特征融合解释算法，可以得到包含了全局信息和像素级信息的本质特征。为了对本文提出的遥感图像细粒度分类可解释方法进行效果评估，本节中，我们以EFM-Net作为遥感图像目标细粒度分类实现方法，通过与Smooth-Grad、Deep-Lift和Integrated-Gradients等其他可解释性分析方法进行横向比较，证明遥感图像细粒度分类可解释方法的优越性。

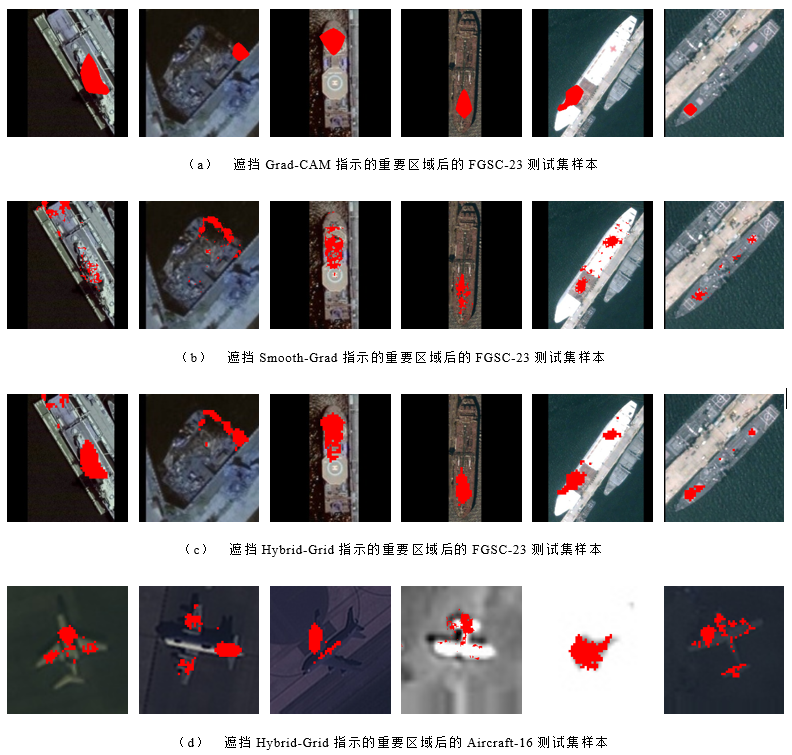


图7 依据解释器指示区域进行遮挡的样本示例

如图7所示，我们可以发现，Deep-Lift和Integrated-Gradients生成的显著图存在噪声和解释效果不佳的问题。由于Integrated-Gradients通过引入与输入相似的多个样本来计算梯度，深度模型内部对这些尺度缩放图像的处理可能存在很大的差异，因此会引入虚假相关性（Yang等，2023），导致生成的显著图错误地强调或忽视一些特征，不能够抓住EFM-Net判断依据的本质特征。由于Deep-Lift方法涉及对参考激活的选择，参考激活的不准确或不合适会导致无法捕捉到神经元间复杂的相互作用和依赖关系，此外，梯度传播受限也会导致信息传递不完整等问题，因此解释效果欠佳。与之相比，Smooth-Grad和Grad-CAM则给出了较为符合目视解译判断依据的显著图。具体而言，Smooth-Grad可以生成更清晰、细节较多的显著性图，但会错误地指示飞机的影子等并不属于目标细粒度分类本质特征的干扰信息。Grad-CAM可以关注到对全局决策更重要的块状区域，然而由于全局池化操作的使用，Grad-CAM会丢失一些细节区域的信息，同时，受限于目标细粒度分类网络最后一层卷积层的尺度问题，将热力图映射回原图时导致显著图难以生成较为清晰的特征边缘。 综上所示，目前现有的可解释性分析方法都不能对目标细粒度分类网络的决策过程进行可信赖、稳定的解释，并给出符合人类感知的本质特征表示。

我们提出的遥感图像细粒度分类可解释方法——Hybrid-Grid对于遥感图像目标细粒度分类网络具有更好的解释效果。如图6（e）所示，红色响应为EFM-Net在目标细粒度分类任务中抓取到的关键特征。从视觉连贯性和判别性的角度来看（Simonyan等，2013；Zeiler和Fergus，2014；Springenberg等，2014；Selvaraju等，2016；Sundararajan等，2017），Hybrid-Grid在解释区域的定位上表现更准确，其指示区域在目标细粒度分类网络感兴趣的细粒度特征上，同时不易受到影子等干扰因素的影响。从视觉感知角度来看，Hybrid-Grid给出的显著图具有较小的噪声，在解释具体目标的分类时保持了归因分析法所具有的像素级信息和细节性，相对于类激活图映射系列方法具有更清晰的特征边缘。另外，Hybrid-Grid的显著图具有较好的像素连贯性，能够实现以区域响应的形式定位目标物体，显示分类器决策依赖的本质特征的区域。综上所述，本文提出的遥感图像细粒度分类可解释方法同时具有先进的原理和较好的视觉解释效果，能给出对目标细粒度分类网络更准确和直观的解释结果。

##### 4.3 基于网格的特征融合解释算法参数

Hybrid-Grid的先进性之一体现于它对像素级特征和高层特征的良好平衡，通过调节基于网格的特征融合解释算法的参数，一方面可以达到减少噪声、平滑波动、聚合特征的效果，另一方面可以控制生成显著图的细度，使其在保持对目标细粒度分类网络高精度解释效果的情况下适应不同尺度的遥感图像。以224x224的输入图像尺度为例，可以设置特征融合解释算法的网格尺度为输入图像尺度的任意因子。如图8所示，从左到右特征融合解释算法的网格尺度逐渐减小，随着该参数的变化，Hybrid-Grid生成的显著图细节性越来越好，对目标细粒度分类网络决策依赖的本质特征的边界描述得更加清晰，而特征像素间的聚合性越来越差，对目标决策重要区域响应的定位变得分散。

通过本文提出的Hybrid-Grid所指示的本质特征区域，根据3.3节提出的可解释性分析方法量化评价指标，计算Hybrid-Grid对EFM-Net在FGSC-23数据集上的删除和精度损失，考察Hybrid-Grid解释器能否抓住目标细粒度分类决策依据的本质特征，探究特征融合解释算法的网格参数对其产生的影响，其结果如表2所示。我们发现，遮挡Hybrid-Grid指示的本质特征所在区域可以使EFM-Net的Top-1准确率、Top-5准确率和F1指标产生大幅下降，同时，随着掩膜尺寸参数的减小，删除和精度损失指标呈现出一种特殊的变化模式。具体而言，我们通过调整掩膜尺寸从32x32到2x2来观察精度损失的变化，其结果如表2所示。具体而言，删除和精度损失指标，也即，Hybrid-Grid对EFM-Net的解释性能，在掩膜尺寸减小的初期阶段有所下降。这是因为较大的网格尺寸使Hybrid-Grid的特征融合解释算法的定位区域较广，过大的掩膜遮挡面积使得目标细粒度分类网络的准确率大幅下降。但与图8所展示的可视化效果结合来看，较大的网格尺寸可能会导致特征融合解释算法不能良好地融合像素级特征信息和高层特征信息，使Hybrid-Grid无法准确地定位细粒度的本质特征信息，导致可解释性分析方法的性能下降。因此，应进一步减小特征融合解释算法的网格参数。

随着掩膜尺寸进一步减小，我们发现删除和精度损失指标持续上升，这是因为较小的网格尺寸能够使特征融合解释算法更好地捕捉到目标细粒度分类网络决策关注的细微特征，从而提高Hybrid-Grid的解释性能。当掩膜尺寸达到2x2时，Hybrid-Grid的删除和精度损失指标达到了最大值，这意味着在输入图像为224x224的情况下，当特征融合解释算法的网格尺寸为2x2时，Hybrid-Grid能够以最佳的方式解释EFM-Net进行目标细粒度分类的决策依据，并提供最准确、详细的本质特征定位。

表2不同掩膜尺寸对目标细粒度分类网络造成的精度损失

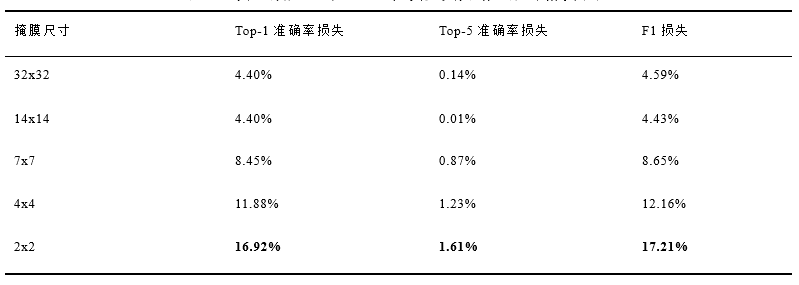




图8 不同参数的基于网格的特征融合解释算法对EFM-Net的解释效果

##### 4.4 量化指标评价可解释性分析方法

根据3.3节所述，本节使用平均下降、连贯性、复杂度和ADCC指标对Hybrid-Grid及其他现有可解释性分析方法在EFM-Net上进行量化指标评估，以定量衡量不同可解释性分析方法在目标细粒度分类网络上的性能。对于普通分类网络来说，包括ResNet18、VGG16在内， Samuele Poppi等人于2021年发现CAM系列的Score-CAM在ADCC指标上取得了最好的性能指标，这表明Score-CAM方法能够有效地解释传统分类网络的决策过程，因此在本节中，使用Score-CAM作为分类网络的基准可解释性分析方法。

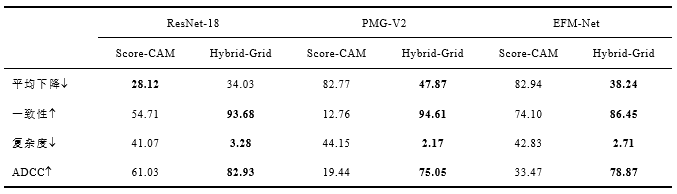
ResNet-18可以有效地解决传统分类任务（He，2015），Samuele Poppi等人将ResNet18应用于传统分类任务，并用Score-CAM对该过程进行解释。类似地，我们将ResNet18应用于目标细粒度分类任务，并用Score-CAM对该过程进行解释，以比较使用Score-CAM解释ResNet-18应用于传统分类任务和目标细粒度方法的性能。在表3中，我们报告了Score-CAM在解释ResNet-18应用于FGSC-23测试集的过程中使用现有指标进行评价的值，与在传统分类任务中相比，这四个指标都产生了显著的下降。其中，ADCC指标与Samuele Poppi提出的最优数值77.30相比下降了16.27。这是因为ResNet-18在目标细粒度分类任务上的适应性不足，无法有效地捕捉到重要的目标区域进行分类。

为了评估Hybrid-Grid在解释目标细粒度分类网络方面的优势，我们选择EFM-Net作为目标模型，并使用上述量化指标来比较我们提出的可解释性分析方法Hybrid-Grid和基准方法Score-CAM对EFM-Net应用于FGSC-23测试集的解释效果。结果如表3所示，Hybrid-Grid在所有指标上都取得了最好的结果，该指标证明它能够对提供一致且准确的解释；去除冗余特征、突出本质特征；其解释具有更高的可靠性和稳定性。与之相比，Score-CAM虽然在解释传统分类任务时表现良好，但在解释目标细粒度分类网络时存在一定的局限性，这也证明了现有可解释性分析方法对目标细粒度分类网络关注的本质特征的解释能力有限。

为了验证Hybrid-Grid解释方法在其他目标细粒度分类网络上的通用性和有效性，我们选择了2021年提出的PMG-V2网络（Ruoyi Du等，2021）作为测试网络，使用FGSC-23作为测试集，将Hybrid-Grid对PMG-V2网络的分类过程进行可解释性分析，并与基准方法Score-CAM进行比较，实验结果如表3所示。Hybrid-Grid对PMG-V2的解释效果在所有指标（包括平均下降、一致性、复杂度、ADCC）上均优于Score-CAM，并展现出优秀的数值。这表明对于目标细粒度分类网络PMG-V2来说，Hybrid-Grid具有更加准确和稳定的解释能力。该实验证明了Hybrid-Grid不仅适用于之前提出的EFM-Net网络，还可以应用于其他目标细粒度分类网络，且在这些网络上也具备良好的解释性能。

综上所述，量化指标结果证明了EFM-Net在目标细粒度分类任务上的先进性，以及Hybrid-Grid解释器在解释目标细粒度分类模型的通用性和优越性。

表3量化指标评估Hybrid-Gird解释器

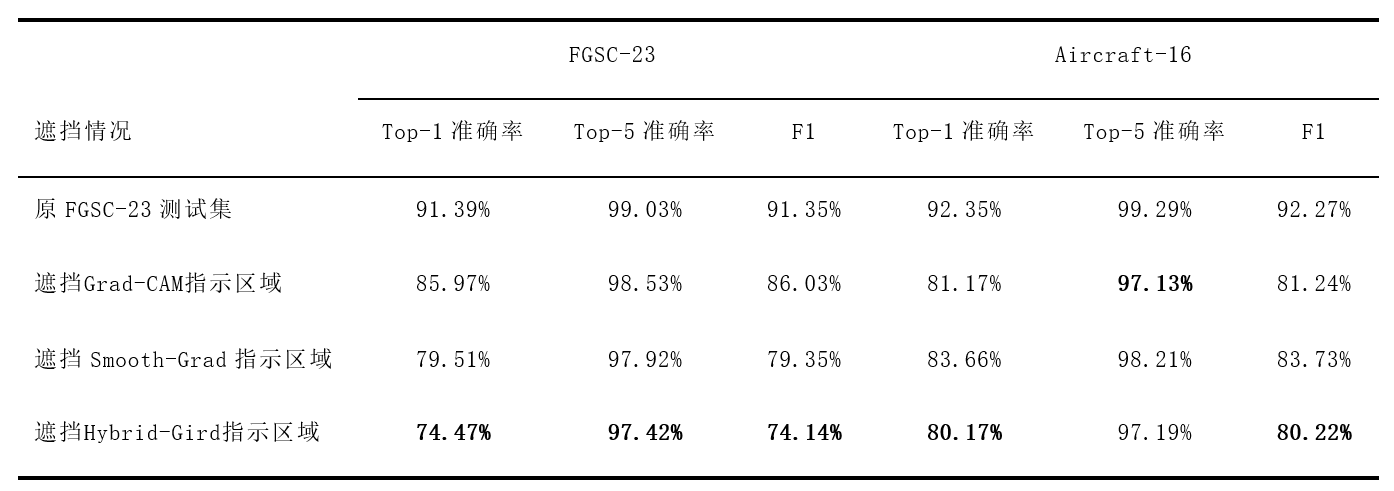


##### 4.5 删除实验评价可解释性分析方法

根据3.3节所述，本节使用删除及精度损失指标对Hybrid-Grid及其他现有可解释性分析方法进行评估，我们使用在FGSC-23数据集上训练后的EFM-Net，通过在原FGSC-23测试集以及根据不同可解释性分析方法使用掩膜遮挡了测试图像特定区域的测试集进行测试，考察EFM-Net模型的精度损失情况。考察的精度指标包括Top-1准确率、Top-5准确率和F1。Top-1准确率是指模型在对单个样本进行分类时，预测的最高置信度类别与实际类别相符的比例；Top-5准确率是指模型在对单个样本进行分类时，在前五个最高置信度的预测类别中，至少有一个与实际类别相符的比例；F1综合考虑了模型的精确率（预测为正例的样本中，真正为正例的比例）和召回率（实际为正例的样本中，被正确预测为正例的比例）的指标，可以评估模型的分类性能。

如表4所示，遮挡Hybrid-Gird指示的重要区域后，EFM-Net在FGSC-23测试集上的Top-1准确率下降了16.92%，Top-5准确率下降了1.61%，F1下降了17.21%。通过与根据其他可解释性分析方法的遮挡结果进行比较，Hybrid-Grid指示的重要区域对Top-1、Top-5准确率和F1指标产生的影响高于其他两个解释器。如表4所示，遮挡Hybrid-Gird解释器所指示的重要区域后，EFM-Net在Aircraft-16测试集上的Top-1准确率下降了12.18%，Top-5准确率下降了2.1%，F1下降了11.95%。通过与其他解释器结果进行比较，Hybrid-Grid解释器指示的重要区域对Top-1、F1指标产生的影响高于其他两个解释器，对Top-5准确率的影响略低于Grad-Cam解释器。综合来看，Hybrid-Grid抓取的特征确实是对EFM-Net分类决策过程最关键的本质特征，证明Hybrid-Grid对目标细粒度分类网络的解释性能优于现有可解释性分析方法。

表4遮挡实验对EFM-Net性能指标的影响



##### 4.6 可解释性分析方法的适应性和可行性

在本研究中，我们对Hybrid-Gird解释器在不同数据集、不同分类网络上的适应性和可行性进行了实验。我们选择了一个常用的分类网络——传统的ResNet18，以及两个具有代表性的目标细粒度分类网络——PMG-V2和EFM-Net，并将它们应用于两个具有代表性的FGSC23和Aircarf16数据集，获得它们在这些数据集上的分类结果。接下来，我们使用本文提出的可解释性分析方法Hybrid-Gird对这些分类网络的决策过程进行解释，通过该解释器，我们能够可视化输入图像中对网络模型决策具有关键作用的本质特征，这些解释结果如图7所示。同时，我们还使用了多种评价指标来量化评价Hybrid-Gird解释器的性能。这些评价指标包括平均下降（Average Drop）、一致性（Coherency）、复杂度（Complexity）、ADCC（Average DCC）、删除和精度损失（Deletion and precision loss）等，用于评估Hybrid-Gird解释器在不同数据集和不同目标细粒度分类网络上的表现，以判断解释器的适应性和可行性，其结果如表3、表4所示。

上述实验结果证明，在不同的数据集和分类网络中，Hybrid-Gird可以准确捕捉到输入图像中与目标分类密切相关的关键特征，并通过可视化结果展示出来。这表明Hybrid-Gird解释器能够应用于不同分类网络、不同数据集进行解释。

#### 5 结 论

目前，基于遥感图像的目标细粒度分类网络已经在军事和民用领域得到广泛应用，如何对其内部决策机理进行可解释性研究是当前细粒度算法进一步提高决策依据可信度的关键问题。面向解耦目标细粒度分类网络复杂决策过程、提高现有细粒度网络在安全敏感任务中可信应用的现实需求，本文针对目标细粒度分类网络的可解释性分析方法展开了研究。

本文检验了当前主流的可解释性方法在遥感图像目标细粒度分类网络上的适用性，针对其局限性提出了一种基于网格的细粒度分类本质特征可解释性分析方法，通过像素级特征提取算法、高层特征提取算法和基于网格的特征融合解释算法，其对网络决策依据本质特征的精确定位能力进一步提升了领域内对目标细粒度网络的解释能力。

围绕EFM-Net、ResNet-18、PMG-V2细粒度分类网络，采用ADCC等量化评估指标对若干典型可解释性分析方法进行了对比试验，评价结果证明本文提出的Hybrid-Grid可解释性分析方法取得了较好效果，证实了该解释器能够更准确地揭示目标细粒度分类网络的工作原理和视觉注意区域的细粒度细节。该可解释性分析方法的泛化性在FGSC-23和Aircraft16数据集上得到了充分验证。

参考文献：

1. Xiong W, Xiong Z and Cui Y. An explainable attention network for fine-grained ship classification using remote-sensing images. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-14

2. Nie Y, Bian C and Li L. Adap-EMD: Adaptive EMD for aircraft fine-grained classification in remote sensing. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 1-5

3. Simonyan K, Vedaldi A and Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. [DOI:10.48550/arXiv.1312.6034]

4. Liang W , Li J , Diao W ,et al. FGATR-Net: Automatic Network Architecture Design for Fine-Grained Aircraft Type Recognition in Remote Sensing Images. Remote Sensing, 2020, 12(24):4187.[DOI:10.3390/rs12244187]

5. Han Y, Yang X, Pu T, et al. Fine-grained recognition for oriented ship against complex scenes in optical remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1-18

6. Baehrens D, Schroeter T, Harmeling S, et al. How to explain individual classification decisions. The Journal of Machine Learning Research. [DOI:10.1007/s10846-009-9348-4]

7. Yang P B, Sang J T, Zhang B, et al. 2023. Interpretability of depth models for image classification. Journal of Software， 34 (1) : 25（杨朋波,桑基韬,张彪,等. 2023. 面向图像分类的深度模型可解释性研究综述.软件学报。34(1):25）

8. Geirhos R , Rubisch P , Michaelis C ,et al. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. 2018. [DOI:10.48550/arXiv.1811.12231]

9. Zhang W , Jiao L , Li Y ,et al. Laplacian Feature Pyramid Network for Object Detection in VHR Optical Remote Sensing Images. IEEE Transactions on Geoscience and Remote Sensing, 2022(60-):60.

10. Fu K , Dai W , Zhang Y ,et al. MultiCAM: Multiple Class Activation Mapping for Aircraft Recognition in Remote Sensing Images.Remote Sensing, 2019, 11(5). [DOI:10.3390/rs11050544]

11. Ding Y, Chong Y, Pan S, et al. Spatial-spectral unified adaptive probability graph convolutional networks for hyperspectral image classification. IEEE Transactions on Neural Networks and Learning Systems, 2021.

12. Huang S, Zhang H, Pižurica A. Subspace clustering for hyperspectral images via dictionary learning with adaptive regularization. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1-17.

13. Everingham M, Van Gool L, Williams C K I, et al. The pascal visual object classes (voc) challenge. International journal of computer vision, 2010, 88: 303-338.

14. Lapuschkin S ,Wäldchen, Stephan, Binder A ,et al. Unmasking Clever Hans predictors and assessing what machines really learn.Nature Communications, 2019, 10(1). [DOI:10.1038/s41467-019-08987-4]

15. Nguyen A , Yosinski J , Clune J .Understanding Neural Networks via Feature Visualization: A Survey[C]//2019. [DOI:10.1007/978-3-030-28954-6\_4]

16. Ribeiro M T , Singh S and Guestrin C ."Why Should I Trust You?": Explaining the Predictions of Any Classifier.ACM, 2016. [DOI:10.1145/2939672.2939778]