用变压器进行遥感图像变化检测

陈浩、齐子鹏和史振伟*, *IEEE会员*

***摘要*--**
**现代变化检测（CD）已经通过深度卷积的强大判别能力取得了再一次的成功。然而，由于场景中物体的复杂性，高分辨率遥感CD仍然具有挑战性。具有相同语义概念的物体可能在不同的时间和空间位置显示出不同的光谱特征。最近大多数使用纯卷积的CD管道仍在努力将空间-
时间中的长距离概念联系起来。非局部自我关注的方法通过对像素之间的密集关系进行建模而显示出有希望的性能，但在计算上是低效的。在这里，我们提出了一个比特时空图像变换器（BIT），以便在空间-
时间域内有效地模拟上下文。我们的直觉是，感兴趣的变化的高层次概念可以由几个视觉词，即语义标 记 来 表示。为了实现这一点，我们将位-
时图像表达为几个代币，并使用转化器编码器在紧凑的基于代币的空间-
时间中建立语境模型。然后，学习到的富含上下文的标记被反馈到像素空间，通过转化器解码器完善原始特征。我们将BIT纳入一个基于深度特征差分的CD框架中。在三个CD数据集上进行的广泛实验证明了所提方法的有效性和效率。值得注意的是，我们基于BIT的模型明显优于纯卷积基线，而计算成本和模型参数仅低3倍。基于一个没有复杂结构（如FPN、UNet）的天真骨干（ResNet18），我们的模型超过了几个最先进的CD方法，包括在效率和准确性方面优于最近的四个基于注意力的方法。我们的代码将被公开。**

***索引词*-**
变化检测（CD），高分辨率光学遥感（RS）图像，变压器，注意机制，卷积神经网络（CNN）。

I. 简介

变化检测（CD）是遥感（RS）的主要课题之一。变化检测的目标是通过比较同一地区在不同时间拍摄的共同登记图像，为该地区的每个像素分配二元标签（即变化或无变化）[1]。变化的定义在不同的应用中有所不同，如城市扩张[2]、森林砍伐[3]和损害评估[4]。基于RS图像的信息提取仍然主要依靠人工视觉解释。自动CD技术可以减少丰富的劳动力成本和时间消耗，因此引起了越来越多的关注[2, 5- 13]。

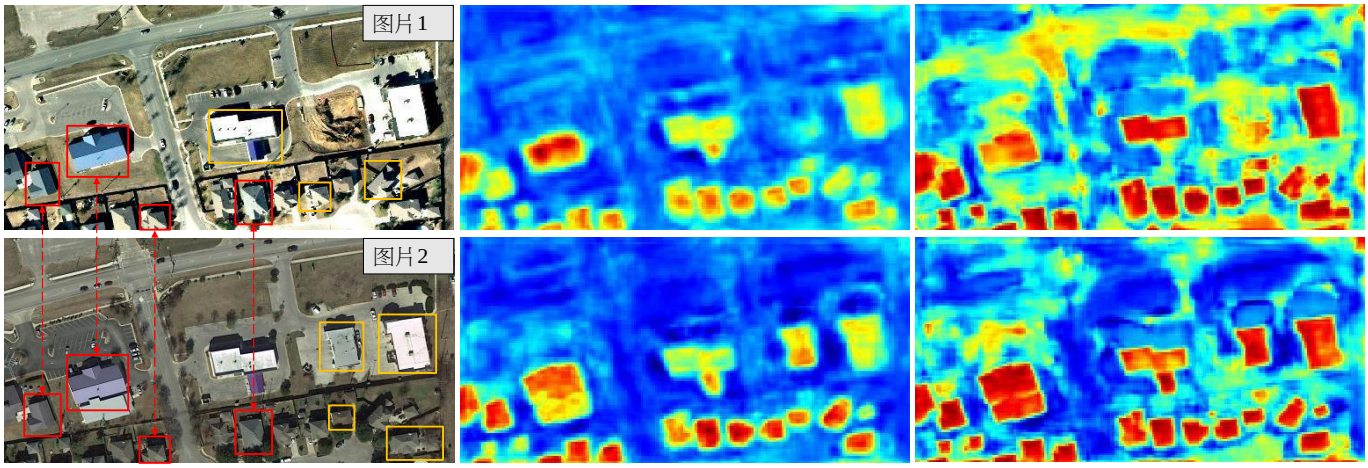
高分辨率（HR）卫星数据和航空数据的出现，为监测土地覆盖和土地利用的精细程度开辟了新的途径。基于高分辨率光学RS图像的CD仍然是一项具有挑战性的任务，这包括两个方面。1）场景中存在的物体的复杂性，2）不同的成像条件。这两点导致具有相同语义概念的物体在不同时间和不同空间位置（时空）表现出不同的光谱特征。例如，如图1（a）所示，场景中的建筑物体具有不同的形状和外观（黄色框内），同一建筑物体在不同时间可能由于光照变化和外观改变而具有不同的颜色（红色框内）。为了识别复杂场景中的兴趣变化，一个强大的CD模型需要。
1) 识别场景中感兴趣的变化的高级语义信息，2）将真正的变化与复杂的无关变化区分开来。

如今，由于其强大的判别能力，深度卷积神经网络（CNN）已成功应用于RS图像分析，并在CD任务中表现出良好的性能[5]。最近大多数有监督的CD方法[2,6-13]依靠基于CNN的结构，从每个时间图像中提取高层次的语义特征，揭示感兴趣的变化。

由于空间和时间范围内的情境建模对于识别高分辨率遥感图像中的兴趣变化至关重要，最新的努力集中在增加模型的接收场（RF），通过堆叠更多的卷积层[2，6-8]，使用扩张卷积[7]，以及应用注意机制[2，6，9-13]。与纯粹的基于卷积的方法本质上受限于RF的大小不同，基于注意力的方法（通道注意力[9-12]、空间注意力[9-11]和自我注意力[2，6，13]）能有效地对全局信息进行建模。然而，大多数现有的方法仍然难以将时空中的长距离概念联系起来，因为它们要么将注意力单独应用于每个时空图像以增强其特征[9]，要么简单地使用注意力在通道或空间维度上对融合的位时特征/图像重新加权[10-12, 14]。最近的一些工作[2, 6, 13]通过利用自我注意来模拟时空中任何一对像素之间的语义关系，取得了很好的性能。然而，他们的计算效率很低，需要很高的计算复杂度，随着像素数量的增加而呈二次增长。

为了解决上述挑战，在这项工作中，我们引入了

该工作得到了国家重点研发计划2019YFC1510905项目、国家自然科学基金61671037项目和北京市自然科学基金4192034项目的支持。*(通讯作者:史振伟 (e-mail: shizhenwei@buaa.edu.cn)。*
陈浩、齐子鹏和史振伟在北京航空航天大学宇航学院图像处理中心，北京航空航天大学数字媒体北京市重点实验室，北京，100191；同时也在北京航空航天大学宇航学院虚拟现实技术与系统国家重点实验室，北京，100191。



(a) (b) (c)

图1.说明了上下文建模的必要性和我们的BIT模块的效果。(a)
一个复杂场景的比特时空高分辨率图像的例子。建筑物在不同时间（红框）和不同空间位置（黄框）显示出不同的光谱特征。一个强大的建筑CD模型需要识别建筑对象，并通过利用上下文信息区分真实的变化和不相关的变化。基于高层次的图像特征（b），我们的BIT模块利用时空上的全局背景来增强原始特征。增强后的特征与原始特征之间的差异图像（c）显示了建筑区域特征在不同时空的持续改进。

位时图像变换器（BIT）以一种有效的方式对位时图像中的长范围背景进行建模。我们的直觉是，感兴趣的变化的高层次概念可以由几个视觉词，即语义标记来表示。我们的BIT不是在像素空间中对像素之间的密集关系进行建模，而是将输入的图像表达为几个高层次的语义标记，并在一个紧凑的基于标记的时空中对背景进行建模。此外，我们通过利用每个像素和语义令牌之间的关系来增强原始像素空间的特征表示。图1给出了一个例子，说明我们的BIT对图像特征的影响。考虑到与建筑概念有关的原始图像特征（见图1（b）），我们的BIT通过考虑时空中的全局语境，学会进一步一致地突出建筑区域（见图1（c））。请注意，我们展示了增强后的特征和原始特征之间的差异图像，以更好地展示拟议的BIT的作用。

我们将BIT纳入一个基于深度特征差分的CD框架中。图2说明了我们基于BIT的模型的整体程序。一个CNN骨干网（ResNet）被用来从输入图像对中提取高级语义特征。我们采用空间注意力将每个时间特征图转换为一组紧凑的语义标记。然后，我们使用转换器[15]编码器来模拟这两个标记集内的上下文。由此产生的富含上下文的标记被一个连体转化器解码器重新投射到像素空间，以增强原始像素级特征。最后，我们从两个精炼的特征图中计算出特征差异图（FDI），然后将其送入一个浅层CNN，以产生像素级的变化预测。

我们工作的贡献可以总结为以下几点。

- 我们为遥感图像变化检测提出了一种有效的基于变压器的方法。我们在CD任务中引入变换器，以更好地模拟上下文。在位时图像中，这有利于识别感兴趣的变化并排除不相关的变化。
- 我们的BIT不是在像素空间的任何一对元素之间建立密集的关系，而是将输入的图像表达为几个视觉词，即标记，并在紧凑的基于标记的时空里建立上下文模型。
- 在三个CD数据集上进行的大量实验验证了所提方法的有效性和效率。我们用BIT替换了ResNet18的最后一个卷积阶段，所得到的基于BIT的模型在计算成本和模型参数降低3倍的情况下，明显优于纯粹的卷积对应。基于一个没有复杂结构（如FPN、UNet）的天真CNN主干，我们的模型在效率和准确性方面比最近几个基于注意力的CD方法表现得更好。

本文的其余部分组织如下。第二节介绍了基于深度学习的CD方法的相关工作和最近在RS中的基于变压器的模型。第三节给出了我们提出的方法的细节。第四节报告了一些实验结果。第五节给出了讨论，第六节得出了结论。

II. 相关的工作

A. 基于深度学习的遥感图像变化检测

基于深度学习的光学RS图像的监督CD方法一般可分为两个主流[8]。

一种是两阶段解决方案[16-18]，即训练一个CNN/FCN来分别对位时图像进行分类，然后比较它们的分类结果来决定变化。这种方法只有在变化标签和位时语义标签都可用时才实用。

另一种是单阶段的解决方案，它直接从位时图像中产生变化结果。补丁级方法[19-21]将CD任务建模为一个相似度

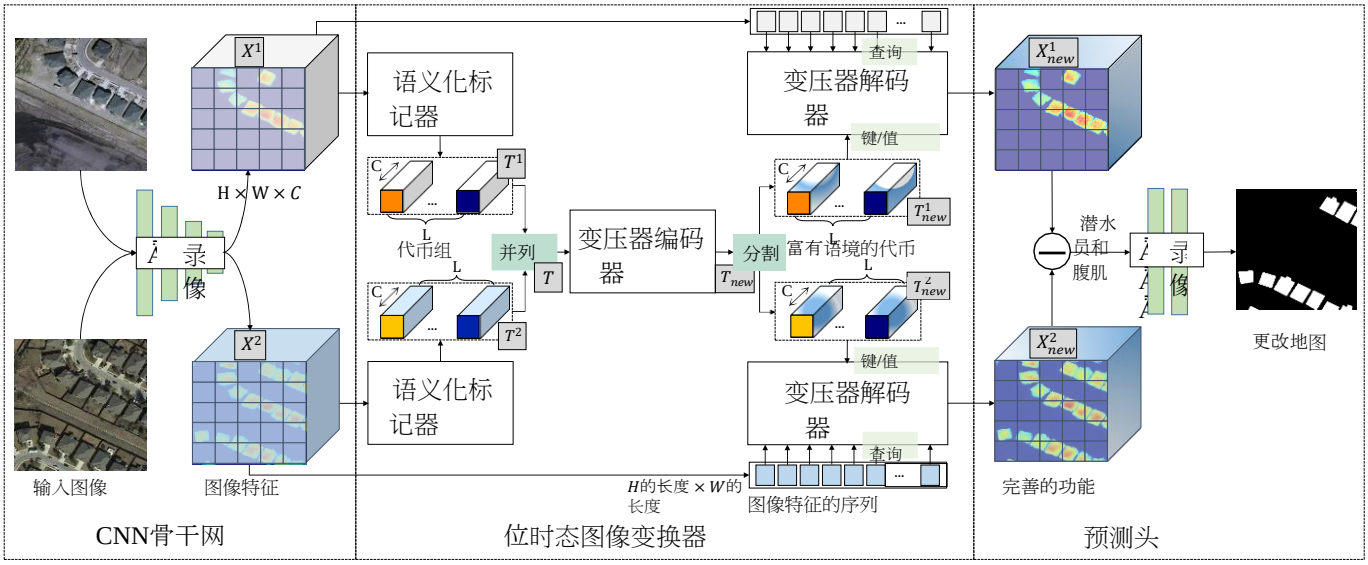


图2.我们基于BIT的模型的说明。我们的语义标记器将CNN骨干网提取的图像特征汇集成一个紧凑的标记词汇集（$L<HW$）。然后，我们将串联的比特时空令牌送入转化器编码器，以便在基于令牌的时空里将概念联系起来。每个时空图像产生的富含上下文的标记被投射回像素空间，以通过转化器解码器完善原始特征。最后，我们的预测头通过将计算出的特征差异图像送入浅层CNN来产生像素级预测。

检测过程中，将位时图像分组为成对的斑块，并在每对斑块上采用CNN来获得其中心预测。像素级的方法[2, 3, 6, 7, 9-13, 22-28]使用FCN直接从两个输入生成高分辨率的变化图，通常比补丁级的方法更有效率和效果。由于CD任务需要处理两个输入，如何融合位时信息是一个重要课题。现有的基于FCN的方法根据位时信息的融合阶段可以大致分为两组。图像层面的方法[3, 22-24, 29]将位时图像连接起来作为语义分割网络的单一输入。特征层面的方法[2, 6, 7, 9-12, 22, 25-28, 30]结合了从神经网络中提取的位时特征，并根据融合的特征做出改变的决定。

最近的许多工作旨在提高神经网络的特征判别能力，通过设计多级特征融合结构[2, 9, 10, 12, 26, 30]，结合基于GAN的优化目标[23, 26, 28, 31]，以及增加模型的接收场（RF），以便在空间和时间范围内更好地进行情境建模[2, 6-13]。

由于场景中物体的复杂性和图像条件的变化，情境建模对于识别高分辨率遥感图像中的兴趣变化至关重要。为了增加射频大小，现有的方法包括采用更深的CNN模型[2, 6-8]，使用扩张卷积[7]，以及应用注意力机制[2, 6, 9-13]。例如，Zhang等人[7]应用一个深度CNN骨干（ResNet 101[32]）来提取图像特征，并使用扩张卷积来扩大模型的RF大小。考虑到纯粹的卷积网络在本质上受限于每个像素的RF大小，许多最新的努力都集中在引入注意力机制以进一步放大模型的射频，如通道注意[9-12]，空间注意[9-11]，自我注意[2, 6, 13]。然而，他们中的大多数人仍然在努力充分地利用与时间有关的背景，因为他们要么将注意力作为一个特征增强模块分别对待每个时间图像[9]，要么仅仅使用注意力来重新权衡通道或空间维度上的融合位时特征/图像[10-12]。非局部自我注意[2, 6]由于能够利用时空上像素之间的全局关系，显示出有希望的性能。然而，它们的计算效率很低，需要很高的计算复杂度，随着像素数量的增加而呈二次增长。

我们论文的主要目的是学习和利用位时图像中的全局语义信息，以一种高效和有效的方式来提高CD的表现力。与现有的基于注意力的CD方法不同的是，我们从图像中提取一些语义标记，并在基于标记的时空里对上下文进行建模，直接对任何一对元素之间的密集关系进行建模。然后利用所产生的富含语境的标记来增强像素空间中的原始特征。我们的直觉是，场景中的兴趣变化可以用几个视觉词汇（标记）来描述，每个像素的高级特征可以由这些语义标记的组合来表示。因此，我们的方法表现出高效率和高性能。

B. 基于变压器的模型

2017年首次提出的转化器[15]，已被广泛用于自然语言处理（NLP）领域，以解决序列到序列的任务，同时轻松地处理长距离的依赖关系。最近的一个趋势是在计算机视觉（CV）领域采用变换器。由于

基于变换器的强大表示能力，在各种视觉任务中，包括图像分类[33-35]、分割[35-37]，基于变换器的模型显示出与卷积对应模型相当甚至更好的性能。对象检测[36, 38, 39]，图像生成[40, 41]，图像字幕[42]，和超级分辨率[43，44]。

变换器模型在NLP/CV任务上的惊人表现吸引了遥感界研究其在遥感任务中的应用，如图像时间序列分类[45, 46]、高光谱图像分类[47]、场景分类[48]和遥感图像字幕[49, 50]。例如，Li等人[46]提出了一种CNN-变换器方法来进行时间序列图像的作物分类，其中变换器被用来从通过CNN提取的多时相特征序列中学习与土地覆盖语义相关的模式。He等人[47]应用了转化器的一个变体（BERT[51]）。

捕捉高光谱中像素之间的全局依赖性。图像分类。此外，Wang等人[50]采用了转换器来翻译通过以下方式提取的无序词汇CNN从给定的RS图像变成一个格式良好的句子。

在本文中，我们探讨了变压器在二进制CD任务中的潜力。我们提出的基于BIT的方法是有效的并有效地在空间-时间上对全局语义关系进行建模，以利于对感兴趣的变化的特征表示。

III. 高效的基于变压器的变化检测模型

我们的基于BIT的模型的整体程序在图2中得到了说明。我们将BIT纳入一个正常的变化检测管道，因为我们想利用卷积和变换器的优势。我们的模型从几个卷积块开始，以获得每个输入图像的特征图，然后将它们送入BIT以产生增强的位时间特征。最后，产生的特征图被送入预测头以产生像素级预测。我们的关键见解是，BIT学习并联系高层次语义概念的全局背景，并反馈给原始位时特征，使其受益。

我们的BIT有三个主要组成部分。1）一个连体语义标记器，它将像素分组为概念，为每个时空输入生成一个紧凑的语义标记集；2）一个变换器编码器，它在基于标记的时空中模拟语义概念的上下文；3）一个连体变换器解码器，它将相应的语义标记投射回像素空间，以获得每个时空的精细特征图。

我们基于BIT的变化检测模型的推理细节显示在算法1中。

算法1：基于BIT的变化检测模型的推理。

输入。 $\mathbf{I} = (\mathbf{I}^1, \mathbf{I}^2)$ (一对注册的图像)
输出。 $\mathbf{M}$（一个预测变化掩码）

// 第1步：通过CNN骨干网提取高级特征
for i *in* $\{1, 2\}$ **do**
  $\mathbf{X}^i$ = CNN骨干网($\mathbf{I}^i$)
结束
// 第二步：使用BIT来完善位时图像特征
// 计算每个时间特征的标记集
for i *in* $\{1, 2\}$ **do**
  $\mathbf{T}^i$ = Semantic Tokenizer($\mathbf{X}^i$)
结束
$\mathbf{T}$=Concat($\mathbf{T}^1$, $\mathbf{T}^2$)
// 使用编码器来生成富含上下文的标记
$\mathbf{T}_{new}$ = 变频器编码器($\mathbf{T}$)
$\mathbf{T}^1_{新}$, $\mathbf{T}^2_{新}$ =Split($\mathbf{T}_{new}$)
// 使用解码器来完善原始特征
for i *in* {1, 2} **do**
  $\mathbf{X}^i_{新}$ = 变压器解码器($\mathbf{X}^i$, $\mathbf{T}^i_{新}$)
结束
// 第三步：通过预测头获得变化掩码
$\mathbf{M}$ = 预测头($\mathbf{X}^1_{新}$, $\mathbf{X}^2_{新}$)

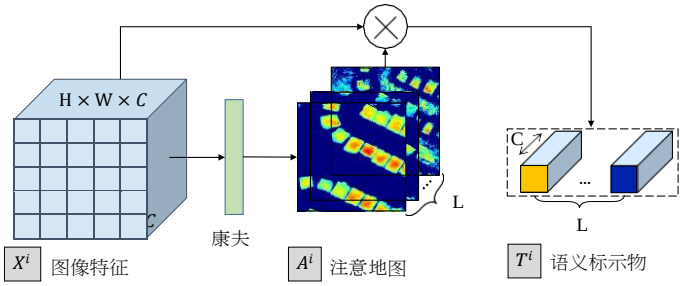


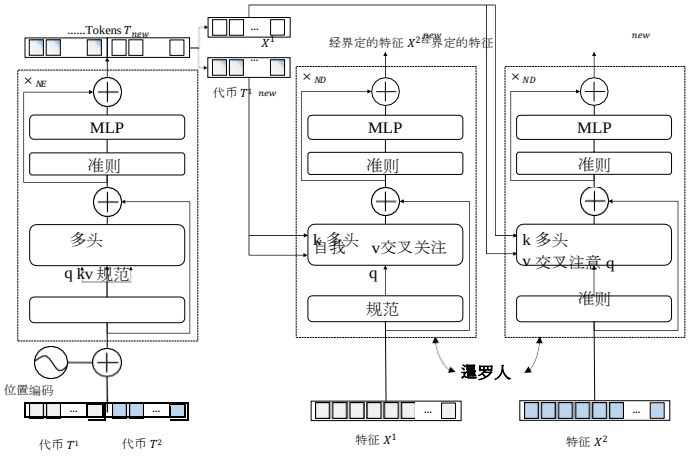
图3.我们的语义标记器的说明。

A. 语义化标记器

我们的直觉是，输入图像中的兴趣变化可以由一些高级概念，即语义标 记 来 描述。而这些语义概念可以由位时图像共享。为此，我们采用了一个连体标记器，从特征中提取紧凑的语义标记每个时空的地图。类似于NLP中的标记器，它将输入的句子分割成几个元素（即单词或短语），并用一个标记向量表示每个元素，我们的语义标记器将整个图像分割成几个视觉单词，每个单词对应一个标记向量。如图3所示，为了获得紧凑的标记，我们的标记器学习了一组空间注意图，将特征图在空间上汇集成一组特征，即标记集。

让$\mathbf{X}^1$, $\mathbf{X}^2 \in R^{HW \times C}$为输入的位时态特征图，其中$H$, W, C为特征图的高度、宽度和通道维度。让$\mathbf{T}^1$, $\mathbf{T}^2 \in R^{L \times C}$为两组标记，其中$L$为标记器的词汇集的大小。

对于i特征图上的每个像素$\mathbf{X}^i_p$ (i = 1, 2)，我们使用点对点的卷积得到L个语义组，每个组表示一个语义概念。然后，我们通过在每个语义组的HW维度上操作的softmax函数计算空间注意力图。最后，我们使用注意力图来计算$\mathbf{X}^i$中像素的加权平均和，以获得一个大小为L的紧凑词汇集，即。



(a) 变压器编码器(b)变压器解码器

图4.我们的变压器编码器和变压器解码器的说明。

语义标记。从形式上看。$\mathbf{T}^i$

$$\mathbf{T}^i = (\mathbf{A}^i)^T \mathbf{X}^i = (\sigma(\varphi(\mathbf{X}^i ; \mathbf{W}))^T \mathbf{X}^i , \quad (1)$$

其中，φ(·)表示与学习型内核$\mathbf{W} \in R^{C\times L的}$点对点卷积，$\sigma$(·)是对每个语义组进行归一化的软键函数，以获得注意力图$\mathbf{A}^i \in R^{HW\times L}$。$\mathbf{T}^i$ 是通过$\mathbf{A}^i$ 和$\mathbf{X}^i$ 的乘法计算出来的。

B. 变压器编码器

在为输入的比特时空图像获得两个语义标记集$\mathbf{T}^1$ ，$\mathbf{T}^2$之后，我们再用一个变换器编码器[15]来模拟这些标记之间的上下文。我们的动机是，基于标记的时空中的全局语义关系可以被转化器充分地利用，从而为每个时空产生丰富的语境标记表示。如图4（a）所示，我们首先将两组标记串联成一个标记集$\mathbf{T} \in R^{2L\times C}$，并将其送入变换器编码器，得到一个新的标记集$\mathbf{T}_{new}$。最后，我们将标记分成两组$\mathbf{T}^i$（$i$=1，2）。

变换器编码器由N_E个，由多头自留地（MSA）和多层感知器（MLP）块组成（图4（a））。与使用后规范化残差单元的原始变换器不同，我们跟随ViT[33]采用前规范化残差单元（PreNorm），即在MSA/MLP之前立即进行层的规范化处理。PreNorm已经被证明比对应的[52]更加稳定和有能力。

在每一层，自我注意的输入是一个三联体（查询$\mathbf{Q}$，键$\mathbf{K}$，值$\mathbf{V}$）从输入$\mathbf{T}^{(l-1)} \in R^{2L\times C}$为。

$$\mathbf{Q}=\mathbf{T}^{(l-1)}W_q, \quad \mathbf{K}=\mathbf{T}^{(l-1)}W_k, \quad \mathbf{V}=\mathbf{T}^{(l-1)}W_v 。 \quad (2)$$

其中，$\mathbf{W}_q^{l-1}, \mathbf{W}_k^{l-1}, \mathbf{W}_v^{l-1} \in R^{C\times d}$是可学习参数。$d$是通道的长度三个线性投影层的等值线，尺寸的三倍。一个注意头被制定为：。

$$\mathrm{Att}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \sigma\left(\frac{QKT}{\sqrt{d}}\right)\mathbf{V}, \quad (3)$$

其中σ( )表示在信道维度上操作的softmax函数。

变换器编码器的核心思想是多头自我关注。MSA并行地执行多个独立的注意力头，输出结果被串联起来，然后投射出最终的数值。MSA的优点是，它可以在不同的位置上共同关注来自不同表征子空间的信息。从形式上看。

$$\mathrm{MSA}(\mathbf{T}^{(l-1)}) = \mathrm{Concat}(\mathrm{head}_1, ..., \mathrm{head}_h)\mathbf{W}^O , \text{其中} \mathrm{head}_j = \mathrm{Att}(\mathbf{T}^{(l-1)}\mathbf{W}_j^q , \mathbf{T}^{(l-1)}\mathbf{W}_j^k , \mathbf{T}^{(l-1)}\mathbf{W}_j^v) 。 \quad (4)$$

其中，$\mathbf{W}_j^q , \mathbf{W}_j^k , \mathbf{W}_j^v \in R^{C\times d}, \mathbf{W}^O \in R^{hd\times C}$为线性投影矩阵，$h$是注意头的数量。

MLP模块由两个线性转换层组成，中间有一个GELU[53]激活。输入和输出的维度是C，内层的维度是2C。从形式上看。

$$\mathrm{mlp}(\mathbf{t}^{(l-1)}) = \mathrm{gelu}(\mathbf{t}^{(l-1)}\mathbf{w}_1)\mathbf{w}_2 \quad (5)$$

其中，$\mathbf{W}_1 \in R^{C\times 2C}$，$\mathbf{W}_2 \in R^{2C\times C}$是线性投影矩阵。

请注意，在将令牌序列$\mathbf{T}$送入转换层之前，我们将可学习的位置嵌入（PE）$\mathbf{W}_{PE} \in R^{2L\times C}$加入到令牌序列中。我们的经验证据（第四章D节）表明，有必要对令牌补充PE。PE编码了元素在基于标记的时空中的相对或绝对位置的信息。这种位置信息可能有利于语境建模。例如，时间上的位置信息可以指导转化器利用与时间有关的语境。

C. 变压器解码器

到目前为止，我们已经获得了两组富含上下文的标示物$\mathbf{T}^i$ $(i = 1, 2)$，用于每个时空图像。这些背景-新丰富的标记包含紧凑的高级语义信息这很好地揭示了兴趣的变化。现在，我们需要将概念的表述投射回像素空间，以获得像素级的特征。为了实现这一点，我们使用一个改进的连体变换器解码器[15]来完善每个时空的图像特征。如图4(b)所示，给定一个特征序列$\mathbf{X}^i$ ，转化器解码器利用每个像素和标记集$\mathbf{T}^i$之间的关系，获得精炼的特点 $\mathbf{X}^i_{新}$。我们把$\mathbf{X}$中的像素^{i}作为查询和标记作为键。我们的直觉是，每个像素可以通过紧凑的语义标记的组合来表示。

我们的变换器解码器由N_D层的多头交叉注意（MA）和MLP块组成。与[15]中的原始实现不同，我们删除了MSA块，以避免大量计算像素间的密集关系

在$\mathbf{X}^i$
。我们采用PerNorm和相同的MLP配置作为变换器编码器。在MSA中，查询、密钥和值都来自同一个输入序列，而在MA中，查询来自图像特征$\mathbf{X}^i$
，而密钥和值来自标记$\mathbf{T}^i_{新}$。形式上，在每层l，MA是定义为：

$$\mathrm{MA}(\mathbf{X}^{i,(l-1)}, \mathbf{T}^i_{新}) = \mathrm{Concat}(\mathrm{head}_1, ..., \mathrm{head}_h)\mathbf{W}^O$$

$$其中头部_j = \mathrm{Att}(\mathbf{X}^{i,(l-1)}\mathbf{W}^q_j, \mathbf{T}^i_{新}\mathbf{W}^k_j, \mathbf{T}^i_{新}\mathbf{W}^v_j) \quad (6)$$

其中，$\mathbf{W}^q_j \in \mathbb{R}^{C\times d}$，$\mathbf{W}^k_j \in \mathbb{R}^{C\times d}$，$\mathbf{W}^v_j \in \mathbb{R}^{C\times d}$，$\mathbf{W}^O \in \mathbb{R}^{hd\times C}$是线性投影矩阵，$h$是注意头的数量。

请注意，我们没有在输入的查询中加入PE，因为我们的经验证据（第四章D节）显示，加入PE后没有相当的收益。

D. 网络细节

CNN骨干。我们使用修改过的ResNet18[32]来提取位时图像特征图。我们将最后两个阶段的步长改为1，并在ResNet后面增加了一个点对点的对话（输出通道C=32）以减少特征维度，然后是一个双线性插值层，从而得到了下采样系数为4的输出特征图以减少空间细节的损失。我们将这个骨干层命名为ResNet18 S5。为了验证所提方法的有效性，我们还使用了两个较轻的骨干网，即ResNet18 S4/ResNet18 S3，它只使用了ResNet18的前四/三层。

位时态图像变换器。根据第四章E节的参数实验，我们设定令牌长度L=4。我们将变换器编码器的层数设为1，变换器解码器的层数设为8。MSA和MA中的头数h被设定为8，每个头的信道尺寸d
被设置为8。

预测头。受益于CNN主干和BIT提取的高层次语义特征，一个非常低的FCN被用于变化识别。给出两个来自BIT输出的上采样特征图$\mathbf{X}^{1*}, \mathbf{X}^{2*} \in \mathbb{R}^{H_0\times W_0\times C}$（$H_0$，$W_0$分别是原始图像的高度和宽度），预测头要生成预测的变化概率图$P \in \mathbb{R}^{H_0\times W_0\times 2}$，其公式为

$$P = \sigma(g(D)) = \sigma(g(|X^{1*} - X^{2*}|)), \quad (7)$$

其中，特征差异图（FDI）$D \in \mathbb{R}^{H_0\times W_0\times C}$是两个特征图减去的元素绝对值，$g: \mathbb{R}^{H_0\times W_0\times C} \to \mathbb{R}^{H_0\times W_0\times 2}$是变化分类器，$\sigma(\cdot)$表示一个对分类器输出的通道维度进行像素化操作的softmax函数。我们的变化分类器的配置是两个带有BatchNorm的3-3卷积层。每个卷积的输出通道是 "32, 2"。

在推理阶段，预测掩码$M \in \mathbb{R}^{H_0\times W_0}$是通过对$P$的通道维度进行像素化的Argmax操作来计算的。

损失函数。在训练阶段，我们最小化交叉熵损失以优化网络参数。从形式上看，损失函数定义为：。

$$L = \frac{1}{H_0 \times W_0} \sum_{h=1,w=1}^{H,W} l(P_{hw}, Y_{hw}), \quad (8)$$

其中$l(P_{hw}, y) = -log(P_{hwy})$是交叉熵损失。

Y_{hw}是位置（h，w）的像素的标签。

IV. 实验结果和分析

A. 实验设置

我们对三个变化检测数据集进行了实验。

LEVIR-CD[2]是一个公共的大规模建筑CD数据集。它包含637对高分辨率（0.5米）的RS图像，尺寸为1024 1024。我们遵循其默认的数据集分割（训练/验证/测试）。由于GPU内存容量的限制，我们将图像切割成大小为256 256的小斑块，没有重叠。因此，我们得到了7120/1024/2048对斑块，分别用于训练/验证/测试。

WHU-CD[54]是一个公共建筑CD数据集。它包含一对尺寸为32507 × 15354的高分辨率（0.075m）航空图像。由于[54]中没有提供数据分割方案，我们将图像裁剪成大小为256 256的小斑块，没有重叠，并随机分割成三部分：6096/762/762，分别用于训练/验证/测试。

DSIFN-CD[10]是一个公共的二元CD数据集。它包括六对大型高分辨率（2米）卫星图像，分别来自中国的六个主要城市。该数据集包含多种土地覆盖物的变化，如道路、建筑、耕地和水体。我们沿用了作者提供的默认的512 512大小的裁剪样本。我们有3600/340/48个样本，分别用于训练/验证/测试。

为了验证我们基于BIT的模型的有效性，我们设置了以下模型进行比较。

- **基础**：我们的基线模型，由CNN主干（ResNet18 S5）和预测头组成。
- **BIT**：我们基于BIT的模型，有一个轻型骨架（ResNet18 S4）。

为了进一步评估所提方法的效率，我们另外设置了以下模型。

- **基础S4**：一个轻型CNN骨架（ResNet18_S4）+预测头。
- **基础S3**：一个轻得多的CNN骨干（ResNet18_S3）+预测头。
- **BIT_S3**：我们基于BIT的模型，有一个更轻的背骨（ResNet18 S3）。

实施细节。我们的模型是在PyTorch上实现的，并使用单个NVIDIA Tesla V100

GPU进行训练。我们对输入的图像斑块进行正常的数据增强，包括翻转、重新缩放、裁剪和高斯模糊。我们使用带有动量的随机梯度下降法（SGD）来优化模型。我们设定动量为0.99，权重衰减为0.0005。学习率最初被设定为0.01，并线性衰减到0，直到训练了200个epochs。验证是这样进行的

在每个训练纪元之后，验证集上的最佳模型被用于测试集的评估。

评价指标。我们使用与变化类别有关的F1分数作为主要的评价指标。F1-
分数由测试的精度和召回率计算，具体如下。

$$F1 = \frac{2}{召回率^{-1} + 精度^{-1}} \tag{9}$$

此外，还报告了精确性、召回率、变化类别的交叉联合（IoU）和总体准确性（OA）。上述指标定义如下。

$$\begin{aligned} 精度 &= TP / (TP + FP) \\ 召回率 &= TP / (TP+FN) \\ IoU &= TP / (TP+FN+FP) \\ OA &= (TP+TN) / (TP+TN+FN+FP) \end{aligned} \tag{10}$$

其中TP、FP、FN分别代表真阳性、假阳性、假阴性的数量。

B. 与最先进的技术比较

我们与几个最先进的方法进行了比较，包括三个纯粹基于卷积的方法（FC- EF[22]，FC-Siam-Di[22]，FC-Siam-Conc[22]）和四个基于注意的方法（DTCDSCN[9]，STA Net[2]，IFNet
[10]和SNUNet[14]）。

- FC-
  EF[22]。图像级别的融合方法，其中比特图像被串联起来，作为一个完全卷积网络的单一输入。
- FC-Siam-Di [22]:特征级融合方法，采用Siamese FCN提取多级特征，并使用特征差异来融合位时信息。
- FC-Siam-
  Conc[22]。特征级融合方法，采用连体FCN提取多级特征，并使用特征串联来融合位时信息。
- DTCDSCN[9]。多尺度特征串联法，在深度连体FCN中加入通道注意和空间注意，从而获得更多的判别特征。请注意，他们还在每个时态的标签图的监督下训练了两个额外的语义分割解码器。为了公平比较，我们省略了语义分割解码器。
- STANet[2]。基于公制的连体FCN方法，它整合了空间-时间注意力机制，以获得更多的辨别特征。
- IFNet[10]。多尺度特征串联方法，在解码器的每一级对串联的位时特征应用通道关注和空间关注。深度监督（即在解码器的每一层计算监督损失）被用来更好地训练中间层。
- SNUNet[14]。多尺度特征串联方法，结合连体网络和NestedUNet[55]，提取高分辨率的高层次特征。通道关注被应用于特征
  在解码器的每一级。深度监督也被用来增强中间特征的辨别能力。

我们使用他们的公共代码和默认的超参数来实现上述CD网络。

表一报告了LEVIR-CD、WHU-CD和DSIFN-
CD测试集的总体比较结果。I报告了LEVIR-CD、WHU-CD和DSIFN-
CD测试集的总体比较结果。定量结果显示，我们基于BIT的模型在这些数据集上始终以明显的优势优于其他方法。例如，我们的BIT的F1分数在这三个数据集上分别超过了最近的STANet的2/1.6/4.7分。请注意，我们的CNN骨干只是纯粹的ResNet，我们没有应用复杂的结构，如FPN在
[2]或[9, 10, 14, 22]中的UNet，它们对于像素
通过融合具有高空间精度的低层次特征和高层次语义特征来完成明智的预测任务。我们可以
结论是，即使使用一个简单的骨干网，我们基于BIT的模型可以实现卓越的性能。这可能归功于我们的BIT能够在全局高度抽象的时空范围内对上下文进行建模，并利用上下文来增强像素空间的特征表示。

图5显示了这些方法在三个数据集上的可视化比较。为了更好地观察，用不同的颜色来表示TP（白色）、TN（黑色）、FP（红色）、FN（绿色）。我们可以观察到，基于BIT的模型取得了比其他模型更好的结果。首先，我们的基于BIT的模型可以更好地避免假阳性（例如，图5（a），（e），（g），（i）），由于物体的外观与兴趣的变化相似。例如，如图5(a)所示，大多数比较方法会错误地将游泳池区域归类为建筑物的变化(视图为红色)，而基于通过全局上下文建模增强的判别特征，STANet和我们的BIT可以减少这种错误检测。在图5（c）中，道路被传统的方法误认为是建筑物的变化，因为道路具有与建筑物相似的颜色行为，而这些方法由于其有限的接收域而无法排除这些伪变化。其次，我们的BIT也能很好地处理由季节性差异或土地覆盖元素的外观改变引起的不相关变化（例如，图5（b）、（f）和（l））。图5（f）中建筑的非语义变化的例子说明了我们的BIT的有效性，它在时空域内学习有效的上下文，以更好地表达真正的语义变化并排除不相关的变化。最后，我们的BIT可以对大面积的变化产生相对完整的预测结果（例如，图5（c），（h）和（j））。例如，在图5（j）中，由于一些比较方法的接收范围有限，图像2中的大面积建筑区域不能被完全检测出来（查看为绿色），而我们基于BIT的模型却能渲染出更完整的结果。

C. 模式的效率和效果

为了公平地比较模型的效率，我们在配备了英特尔至强银4214 CPU和英伟达Tesla V100 GPU的计算服务器上测试了所有的方法。Tab.II报告了不同方法在LEVIR-CD、WHU-CD和DSIFN-
CD测试集上的参数数量（Params.）、每秒浮点运算（FLOPs）和F1/IoU得分。