Machine Translated by Google

遥感图像变化检测
变形金刚

Hao Chen, Zipeng Qi and Zhenwei Shi ，IEEE会员

摘要:现代变化检测（CD)凭借深度卷积强大的判别能力取得了巨大的成功。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在不同时间和空间位置可能表现出不同的光谱特征。

最近使用纯卷积的 CD 管道仍在努力将时空中的远程概念联系起来。非局部自注意力方法通过对像素之间的密集关系进行建模而显示出良好的性能,但计算效率较低。在这里,我们提出了一种双时态图像转换器（BIT)来高效且有效地对时空域内的上下文进行建模。我们的直觉是,兴趣变化的高级概念可以用一些视觉单词（即语义标记)来表示。为了实现这一目标,我们将双时图像表达为几个标记,并使用变换器编码器来对基于紧凑标记的时空中的上下文进行建模。然后,将学习到的上下文丰富的标记反馈到像素空间,以通过转换器解码器细化原始特征。我们将 BIT 纳入基于深度特征差异的 CD 框架中。对三个 CD 数据集的广泛实验证明了所提出方法的有效性和效率。值得注意的是,我们基于 BIT 的模型显着优于纯卷积基线,计算成本和模型参数仅降低了 3 倍。基于没有复杂结构（例如 FPN、UNet)的朴素主干网（ResNet18），我们的模型超越了几种最先进的 CD 方法,包括在效率和准确性方面优于最近四种基于注意力的方法。我们的代码可在 https://github.com/justchenhao/BIT CD 获取。

索引术语 变化检测（CD）、高分辨率光学遥感（RS)图像、变压器、注意力机制、卷积神经网络（CNN）。

一、简介

CHANGE检测（CD)是该领域的主要课题之一遥感（RS）。 CD 的目标是通过比较不同时间拍摄的同一区域的共同配准图像,为区域中的每个像素分配二进制标签（即变化或不变化)[1]。变化的定义因应用而异,例如城市扩张[2]、森林砍伐[3]和损害评估[4]。基于遥感图像的信息提取仍然主要依靠人工目视判读。

该工作得到国家重点研发计划项目2019YFC1510905、国家自然科学基金项目61671037和北京市自然科学基金项目4192034的资助。（通讯作者:施振伟,e-mail: shizhenwei@buaa.edu.cn))

陈浩,齐子鹏,史振伟,北京航空航天大学宇航学院图像处理中心,北京 100191,北京航空航天大学数字媒体北京市重点实验室,北京 100191北京航空航天大学宇航学院虚拟现实技术与系统重点实验室, 北京 100191

自动CD技术可以减少大量的劳动力成本和时间消耗,因而越来越受到人们的关注[2, 5–13]。

高分辨率（HR)卫星数据和航空数据的可用性为精细监测土地覆盖和土地利用开辟了新途径。基于 HR 光学 RS 图像的 CD 在两个方面仍然是一项具有挑战性的任务:1）场景中存在的物体的复杂性,2)不同的成像条件。两者都有助于具有相同语义概念的对象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,场景中的建筑物对象具有不同的形状和外观（黄色框中），并且同一建筑物对象在不同时间可能由于光照而具有不同的颜色（红色框中)变化和外观改变。为了识别复杂场景中的兴趣变化,强大的CD模型需要:1)识别场景中兴趣变化的高级语义信息,2)区分真实变化和复杂的不相关变化。

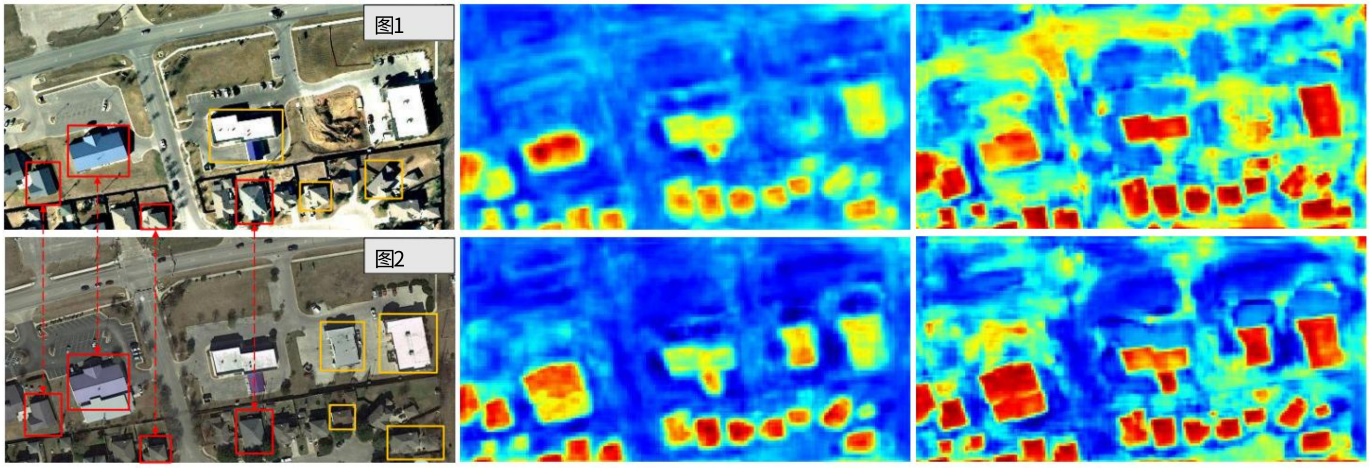
如今,由于其强大的判别能力,深度卷积神经网络（CNN)已成功应用于RS图像分析,并在CD任务中表现出良好的性能[5]。最近的监督 CD 方法 [2, 6-13] 依赖于基于 CNN 的结构从每个时间图像中提取揭示兴趣变化的高级语义特征。

由于空间和时间范围内的上下文建模对于识别高分辨率遥感图像的兴趣变化至关重要,因此最新的努力一直集中在通过堆叠更多卷积层来增加模型的接收场（RF)[2，6-8],使用扩张卷积[7],并应用注意力机制[2,6,9-13]。与本质上受限于 RF 大小的纯粹基于卷积的方法不同,基于注意力的方法（通道注意力 [9–12]、空间注意力 [9–11] 和自注意力 [2, 6, 13])在建模全局信息方面是有效的。然而,大多数现有方法仍然在努力将时空中的远程概念联系起来,因为它们要么分别对每个时间图像应用注意力以增强其特征[9],要么简单地使用注意力来重新加权融合的双时特征/通道或空间维度中的图像[10–12, 14]。最近的一些工作 [2,6,13] 通过利用自注意力对时空中任何像素对之间的语义关系进行建模,取得了有希望的性能。然而,它们的计算效率低下,并且需要高计算复杂度,并且计算复杂度随着像素数量呈二次方增长。

Machine Translated by Google

1岁

法师2



(A) (二) (C)

图 1. 说明上下文建模的必要性以及我们的 BIT 模块的效果。 (a) 双时态高分辨率图像中复杂场景的示例。建筑物对象在不同时间（红色框）和不同空间位置（黄色框）表现出不同的光谱特征。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变化和不相关变化。基于高级图像特征 (b),我们的 BIT 模块利用时空全局上下文来增强原始特征。增强后的特征与原始特征之间的差异图像（c)显示了跨时空的建筑区域特征的一致改善。

空间

为了解决上述挑战,在这项工作中,我们引入了双时态图像变换器（BIT)来对双时态高分辨率场景中对象的复杂性进行建模,以高效和图像的方式对双时态图像中的变化区域在上下文进行建模（见图 1,下图 2）。有效的方式。我们的视觉是,兴趣变化的高级概念可以用空间视觉单词来表示的密集关系进行建模,我们的 BIT 不是对像素空间中的是将输入图像表达为一些高级语义标记,并在基于紧凑标记的时空中对上下文进行建模。此外,我们通过利用每个像素和语义标记之间的关系来增强原始像素空间的特征表示。图 1 给出了一个例子来展示我们的 BIT 对图像特征的影响。考虑到与建筑概念相关的原始图像特征（见图 1 (b)）,我们的 BIT 学习通过考虑时空的全局背景来进一步一致地突出显示建筑区域（见图 1 (c)）。请注意,我们显示了增强特征和原始特征之间的差异图像,以更好地展示所提出的 BIT 的作用。

我们将 BIT 纳入基于深度特征差异的 CD 框架中。我们基于 BIT 的模型的整体过程如图 2 所示。CNN 主干网 (ResNet) 用于从输入图像对中提取高级语义特征。我们利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 Transformer [15] 编码器对两个标记集中的上下文进行建模。由此产生的上下文丰富的标记由连体变换器解码器重新投影到像素空间,以增强原始像素级特征。最后,我们根据两个细化的特征图计算特征差异图像（FDI）,然后将它们输入浅层 CNN 以产生像素级变化预测。

我们的工作贡献可概括如下:

·提出了一种基于变压器的高效遥感图像变化检测方法。我们介绍转化为 CD 任务,以更好地对双时图像内的上下文进行建模,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 不是对像素空间中任何元素对之间的密集关系进行建模,而是将输入图像表达为几个视觉单词（即标记）,并在基于紧凑标记的时空中对上下文进行建模。 ·对三个CD 数据集的广泛实验验证了所提出方法的有效性和效率。我们用 BIT 替换 ResNet18 的最后一个卷积阶段,所得到的基于 BIT 的模型的性能优于纯卷积模型,且计算成本和模型参数仅降低了 3 倍。基于没有复杂结构（例如 FPN、UNet)的朴素 CNN 主干,我们的方法在效率和准确性方面表现出比最近几种基于注意力的 CD 方法更好的性能。

本文的其余部分安排如下。第二部分描述了基于深度学习的 CD 方法的相关工作以及 RS 中最近基于 Transformer 的模型。第三节详细介绍了我们提出的方法。第四节报告了一些实验结果。第五节进行讨论,第六节得出结论。

二.相关工作

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监督CD方法图像通常可以分为两个主要流[8]。

一种是两阶段解决方案[16-18],其中训练CNN/FCN对双时图像进行单独分类,然后比较它们的分类结果以做出改变决策。

这种方法仅当变化标签和双时态语义标签都可用时才实用。

Machine Translated by Google

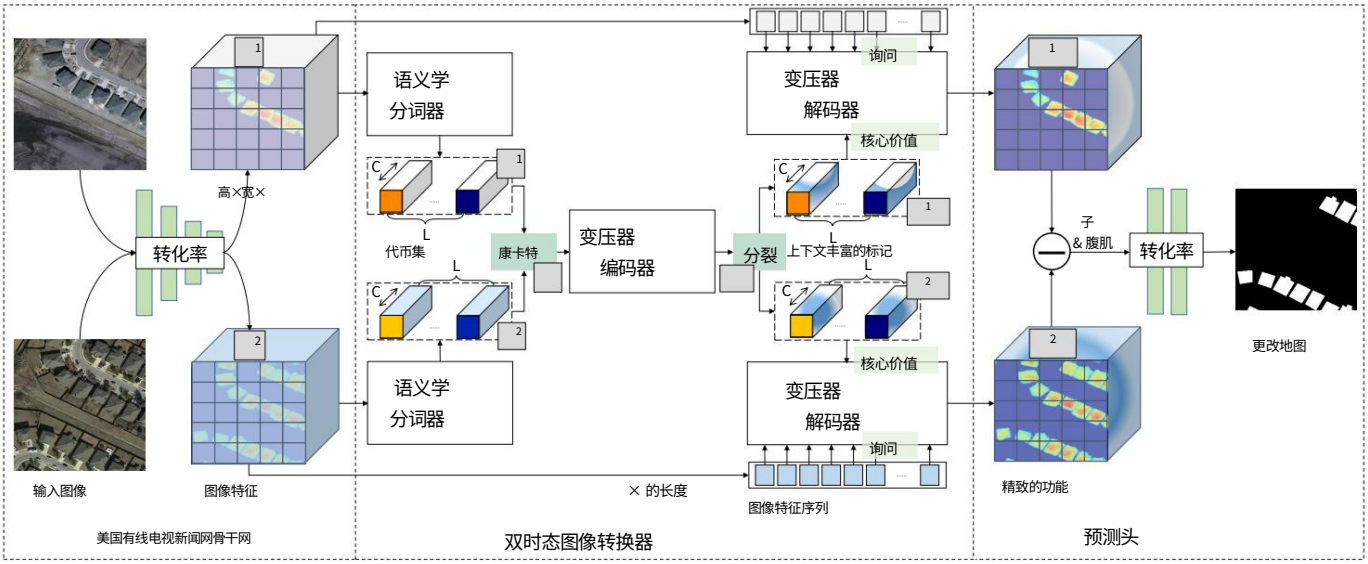


图 2.我们基于 BIT 的模型的图示。我们的语义标记器将 CNN 主干提取的图像特征池化为紧凑的标记词汇集 (L << HW)。然后，我们将连接的双时态标记提供给转换器编码器，以关联基于标记的时空概念。每个时间图像的生成的上下文丰富的标记被投影回像素空间，以通过转换器解码器细化原始特征。最后，我们的预测头通过将计算出的特征差异图像输入浅层 CNN 来生成像素级预测。

输入 图像特征 1×1 编码标记序列

另一种是单阶段解决方案，它直接从双时态图像产生变化结果。补丁级方法 [19-21] 将 CD 任务建模为相似性检测过程，将双时图像分组为补丁对，并在每对上使用 CNN 以获得其中心预测。像素级方法 1×1× c [2,3,6,7,9-13,22-28]使用FCN直接从两个输入生成高分辨率变化图，这通常比补丁级更高效和有效方法。由于CD任务需要处理两个输入，因此如何融合双时态信息是一个重要的课题。现有的基于FCN的方法可以根据双时信息融合的阶段大致分为两类。图像级方法 [3, 22–24, 29] 将双时图像连接起来作为语义分割网络的单个输入。特征级方法 [2,6,7,9-12,22,25-28,30]结合了从神经网络中提取的双时态特征，并根据融合特征做出变更决策。

最近的许多工作旨在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]，并增加模型的接收场 (RF)，以便在空间和时间范围方面更好地进行上下文建模 [2, 6-13]。

由于场景中物体的复杂性和图像条件的变化，上下文建模对于识别高分辨率遥感图像的兴趣变化至关重要。为了增加 RF 大小，现有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，张等人。 [7]应用深度CNN主干（ResNet101[32]）来提取图像特征并使用扩张卷积来扩大模型的RF尺寸。

考虑到纯卷积网络本质上受限于每个像素的 RF 大小，许多最新的工作都集中在引入注意力机制来进一步扩大模型的 RF，例如通道注意力 [9-12]、空间注意力 [9] –11]，自注意力[2,6,13]。然而，他们中的大多数人仍然在努力充分利用与时间相关的上下文，因为他们要么将注意力视为每个时态图像单独的特征增强模块[9]，要么简单地使用注意力来重新加权融合的双时态特征/图像在通道或空间维度上[10-12]。

非局部自注意力 [2, 6] 由于其能够利用时空像素之间的全局关系而表现出有希望的性能。然而，它们的计算效率低下，并且需要高计算复杂度，并且计算复杂度随着像素数量呈二次方增长。

我们论文的主要目的是以高效且有效的方式学习和利用双时图像中的全局语义信息，以提高 CD 性能。与现有的基于注意力的 CD 方法直接对基于像素的空间中任何元素对之间的密集关系进行建模不同，我们从图像中提取一些语义标记并在基于标记的时空中对上下文进行建模。

然后利用生成的上下文丰富的标记来增强像素空间中的原始特征。我们的直觉是，场景内兴趣的变化可以通过一些视觉单词（标记）来描述，并且每个像素的高级特征可以通过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

Transformer 于 2017 年首次推出[15]，已广泛应用于自然语言处理（NLP）领域，以解决序列到序列的任务，同时处理远程任务。

Machine Translated by Google

轻松依赖。最近的趋势是在计算机视觉（CV)领域采用 Transformer。由于 Transformer 强大的表示能力,基于 Transformer 的模型在各种视觉任务中表现出与卷积模型相当甚至更好的性能,包括图像分类 [33–35]、分割 [35–37]、对象检测 [36、 38, 39]、图像生成 [40, 41]、图像字幕 [42] 和超分辨率 [43, 44]。

Transformer 模型在 NLP/CV 任务上的惊人性能引起了遥感界研究其在遥感任务中的应用的兴趣,例如图像时间序列分类 [45, 46]、高光谱图像分类 [47]、场景分类 [48]],以及遥感图像字幕 [49, 50]。例如,李等人。 [46]提出了一种 CNN 变换器方法来执行时间序列图像的作物分类,其中变换器用于从通过 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。他等人。 [47]应用了变压器的变体（BERT [51])来捕获高光谱图像分类中像素之间的全局依赖性。此外,王等人。 [50]利用变压器将CNN从给定的RS图像中提取的无序单词翻译成结构良好的句子。

在本文中,我们探讨了 Transformer 在二进制 CD 任务中的潜力。我们提出的基于 BIT 的方法在建模时空全局语义关系方面是高效且有效的,有利于兴趣变化的特征表示。

三.基于高效变压器的变革检测模型

我们基于 BIT 的模型的整体过程如图 2 所示。我们将 BIT 合并到正常的变化检测管道中,因为我们希望利用卷积和转换器的优势。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,然后将它们输入 BIT 以生成增强的双时特征。最后,将生成的特征图馈送到预测头以产生像素级预测。我们的主要见解是,BIT 学习并关联高级语义概念的全局上下文,以及反馈以有益于原始的双时态特征。

我们的 BIT 具有三个主要组件:1)暹罗语义标记器,它将像素分组为概念,为每个时间输入生成一组紧凑的语义标记;2)变换器编码器,它对基于标记的语义概念的上下文进行建模时空,3)连体变换器解码器,它将相应的语义标记投影回像素空间以获得每个时间的细化特征图。

我们基于 BIT 的变化检测模型的推理细节如算法 1 所示。

A. 语义标记器

我们的直觉是,输入图像的兴趣变化可以通过一些高级概念（即语义标记)来描述。并且语义概念可以共享

算法 1:基于 BIT 的变化检测模型的推断。

输入： I = {(输 1 , 我 2)}（一对注册图像）
出： M（预测变化掩码）

1 // 第 1 步:通过 CNN 主干2 for i in {1, 2} do提取高级特征

3 Xi = CNN Backbone(I 4 end 5 // ）
步骤2:
使用BIT细化双时态图像特征6 // 计算每个时态特征的标记集7 for i in {1, 2} do

8 Ti = 语义分词器(Xi) 9 end

10 T=Concat(T1 , T2 ）
11 // 使用编码器生成上下文丰富的标记
12 Tnew=变压器编码器(T) _
13 T1 新的, T2 new=Split(Tnew)
14 // 使用解码器细化原始特征15 for i in {1, 2} do

16习 新的 = 变压器解码器(Xi , 你是新人）
17结束
Token set (length
L) 18 // 步骤3:通过预测头获取变化掩码
19 M = 预测头(X1 _ 新的, X2 新）

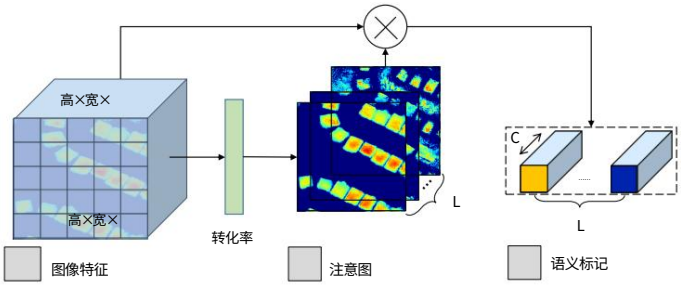


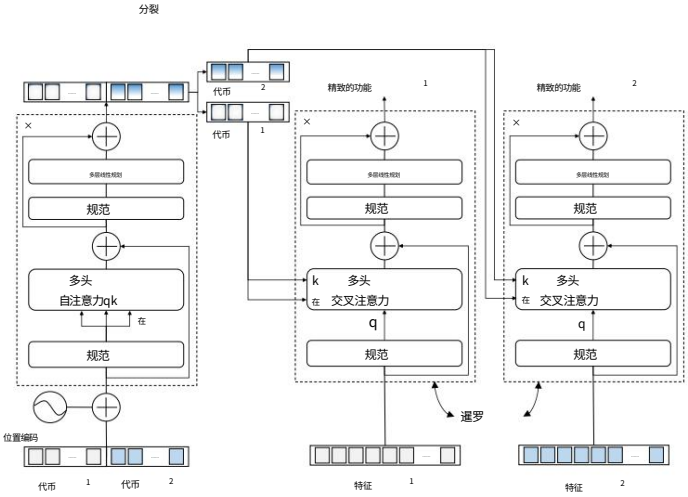
图 3.我们的语义标记器的图示。

通过双时态图像。为此,我们采用 Siamese tokenizer 从每个时间的特征图中提取紧凑的语义标记。类似于 NLP 中的分词器,它将输入句子分割成几个元素（即单词或短语） ,并用一个分词向量表示每个元素,我们的语义分词器将整个图像分成几个视觉单词,每个单词对应一个分词向量。如图3所示,为了获得紧凑的标记,我们的标记器学习一组空间注意力图,以将特征图在空间上池化为一组特征,即标记集。

令X1为输入双时特征图,其中 H、W、C 为特征图的高度、宽度和通道尺寸。令T1 T2 ∈ R 为两组标记,其中 L 为标记词汇集的大小。 高×宽×厚 长×宽 ,

对于特征图Xi (i = 1, 2)，上的每个像素Xi ,我们使用逐点卷积获得 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 维度上运行的 softmax 函数来计算空间注意力图。最后,我们使用注意力图来计算像素的加权平均和

Machine Translated by Google



(a) 变压器编码器 (b) 变压器解码器

图 4. 我们的 Transformer 编码器和 Transformer 解码器的图示。

Xi中获得大小为 L 的紧凑词汇集,即语义标记Ti 。正式地,

$$T_i = (\sigma(\phi(X_i; W)))^{T} X_i \quad (1)$$

其中 ϕ(·) 表示与可学习内核 W ∈ R C×L的逐点卷积， σ(·) 是 softmax 函数,用于标准化每个语义组以获得注意力图Ai ∈ R HW×L。 Ti由A和Xi相乘计算得出

。

B. 变压器编码器

获得两个语义标记集合T1后 ，T2对于输入双时图像,然后我们使用变压器编码器对这些标记之间的上下文进行建模[15]。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换器充分利用,从而为每个时间生成上下文丰富的标记表示。如图4（a)所示,我们首先将两组令牌连接成一个令牌集T ∈ R,并将其输入到变压器编码器中以获得新的令牌集Tnew 。最后,我们将标记分成两组Ti new(i = 1, 2)。

2L×C ,

Transformer 编码器由NE层多头自注意力（MSA)和多层感知器（MLP)块组成（图 4（a））。与使用后范数残差单元的原始变压器不同,我们遵循ViT[33]采用前范数残差单元（PreNorm),即层归一化发生在MSA/MLP之前。 PreNorm 已被证明比对应的版本更稳定、更有能力 [52]。

在每一层 l 中,自注意力的输入是根据输入T(l－1) ε 计算得到的三元组（查询 Q、键 K、值 V）右2L×C 作为:

$$Q = T(l-1)W_q \quad ,$$
$$K = T(l-1)W_k \quad , \quad (2)$$
$$V = T(l-1)W_v \quad ,$$

其中Wl－1 ,Wl－1 ,Wl－1是三个线性投影 ∈ R C×d 是可学习的参数
q k
层的三者,d 是三元组的通道维度。一个注意力头被表述为:

$$\text{那}(Q, K, V) = \sigma\left(\frac{QKT}{\sqrt{d}}\right) \text{在}, \quad (3)$$

其中σ(·)表示在通道维度上运行的softmax函数。

Transformer编码器的核心思想是多头自注意力。 MSA 并行执行多个独立的注意力头,并将输出连接起来,然后投影以产生最终值。 MSA的优点是它可以共同关注来自不同位置的不同表示子空间的信息。正式地,

$$MSA(T(l-1)) = Concat(head1, .., head)WO, T(l-1)Wv),$$
$$\text{其中 } head_j = Att(T(l-1)Wq_j, T(l-1)Wk_j, j \quad (4)$$

矩阵,h 是注意力头的数量,WO ∈ R 其中Wq ,Wv jj投影 高清×厚度 是线性的
j

MLP 块由两个线性变换层组成,中间有一个 GELU [53] 激活。输入和输出的维度为C,内层的维度为2C。正式地,

$$MLP(T(l-1)) = ICE(T(l-1)W1)W2 \quad (5)$$

其中W1 ∈ R 矩阵。C×2C ,W2 ∈ R 2C×C 是线性投影

请注意,我们在输入之前将可学习的位置嵌入（PE)添加到令牌序列 T 中
WP E ∈ R it 到变压2L×C
器层。我们的经验证据（第 2 节）
IV-D)表示需要对代币补充PE。 PE 对基于令牌的时空中元素的相对或绝对位置的信息进行编码。这样的位置信息可以有利于上下文建模。例如,时间位置信息可以指导变压器利用时间相关的上下文。

C. Transformer 解码器

到目前为止,我们已经为每个时间图像获得了两组上下文丰富的标记Ti new(i = 1, 2)。这些上下文丰富的标记包含紧凑的高级语义信息,可以很好地揭示兴趣的变化。现在,我们需要将概念的表示投影回像素空间以获得像素级特征。为了实现这一点,我们使用改进的 Siamese Transformer 解码器 [15] 来细化每个时间的图像特征。如图 4 (b) 所示,给定特征序列Xi ,Transformer 解码器利用每个像素与标记集Ti之间的关系来获得细化特征Xi

,

新的 新的。 我们将Xi中的像素视为查询,将标记视为键。我们的直觉是每个像素都可以由紧凑语义标记的组合来表示。

我们的 Transformer 解码器由多头交叉注意 (MA) 和 MLP 块的ND层组成。与 [15]中的原始实现不同,我们删除了MSA块以避免像素之间密集关系的大量计算

Machine Translated by Google

在奚。我们采用 PerNorm 和与 MLP 相同的配置作为 Transformer 编码器。在MSA中,查询、键和值来自相同的输入序列,而在MA中,查询来自图像特征Xi ,键和值来自标记Ti 。正式地,在每一层l,定义MA作为:

,

新的。

MA(Xi,(l−1) ,其中 时间new) = Concat(head1, ..., head)WO,

headj = Att(Xi,(l−1)Wq j , T 我新一周 j , T 新Wv), (6)

其中Wq ,Wk 量。× d ,W注意力头的数 R jj投影矩阵,h 是 高清×厚度 是线性的

请注意,我们不会将 PE 添加到输入查询中,因为我们的经验证据（第 IV-D 节)显示添加 PE 时没有显着的收益。

D. 网络详细信息

CNN骨干。我们使用修改后的 ResNet18 [32] 来提取双时图像特征图。原始ResNet18有5个阶段,每个阶段下采样2。我们将最后两个阶段的步长替换为1,并在ResNet后面添加逐点卷积（输出通道C = 32)以减少特征维度,然后是双线性插值层,从而获得下采样因子为4的输出特征图,以减少空间细节的损失。我们将此主干命名为 ResNet18 S5。为了验证所提出方法的有效性,我们还使用了两个更轻的主干,即ResNet18 S4/ResNet18 S3,其仅使用ResNet18的前四个/三个阶段。

双时图像转换器。根据第 2 节中的参数实验。 IV-E,我们设置 token 长度 L = 4。我们将 Transformer 编码器的层数设置为 1,Transformer 解码器的层数设置为 8。MSA 和 MA 中的头数 h 设置为 8,通道维度设置为 d每个头设置为 8。

预测头。受益于 CNN 主干和 BIT 提取的高级语义特征,采用非常浅的 FCN 进行变化判别。给定 BIT 输出的两个上采样特征图X1* X2* ∈ R H0×W0×C （H0、W0分别是原始图像的高度、宽度）,预测头将生成预测变化概率图 P ε R H0×W0×2由下式给出 ,

,

$$P = \sigma(g(D)) = \sigma(g(|X1^* - X2^*|)), \quad (7)$$

其中特征差异图像（FDI)D ∈ R H0×W0×C
是两者相减的元素级绝对值

特征图,g : R H0×W0×C → R H0×W0×2是变化分类器,σ(·) 表示对分类器输出的通道维度进行逐像素操作的 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为“32, 2”。

在推理阶段,预测掩码 M ∈ R H0×W0通过对 P 的通道维度进行像素级 Argmax 运算来计算。

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

$$L = \frac{1}{H0 \times W0} \sum_{h=1,w=1}^{高、宽} l(Phw, Yhw), \quad (8)$$

其中l(Phw, y) = −log(Phwy)是交叉熵损失, Yhw是位置 (h, w) 处像素的标签。

四.实验结果与分析

A.实验设置我们在三个变化检

测数据集上进行实验。

LEVIR-CD [2] 是一个公共的大型建筑 CD 数据集。它包含 637 对尺寸为 1024 × 1024 的高分辨率 (0.5m) RS 图像。我们遵循其默认数据集分割（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的小块,没有重叠。因此,我们分别获得了 7120/1024/2048 对用于训练/验证/测试的补丁。

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)航拍图像。由于[54]中没有提供数据分割解决方案,我们将图像裁剪成尺寸为 256 × 256 的小块,没有重叠并随机分割它分为三部分:6096/762/762分别用于训练/验证/测试。

DSIFN-CD [10] 是一个公共二进制 CD 数据集。它包括分别来自中国六个主要城市的六对大型高分辨率（2m)卫星图像。该数据集包含道路、建筑物、农田、水体等多种土地覆盖对象的变化。我们遵循作者提供的大小为 512 × 512 的默认裁剪样本。

我们分别有 3600/340/48 个样本用于训练/验证/测试。

为了验证基于 BIT 的模型的有效性,我们
设置以下模型进行比较:

·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 的轻量主干模型 _(ResNet18 S4）。

为了进一步评估所提出方法的效率,我们另外设置了以下模型： · Base S4:轻CNN主干（ResNet18 S4）+预测头。 · Base S3:一

个轻量级的 CNN 主干 (ResNet18 S3) +

预测头。

· BIT S3:我们基于BIT 的模型,具有更轻的主干结构(ResNet18 S3)。

实施细节。我们的模型在 PyTorch 上实现,并使用单个 NVIDIA Tesla V100 GPU 进行训练。我们将正常的数据增强应用于输入图像块,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（SGD)来优化模型。我们将动量设置为 0.99,权重衰减设置为 0.0005。学习率最初设置为 0.01,并线性衰减到 0,直到训练 200 个 epoch。已执行验证

Machine Translated by Google

在每个训练周期之后,验证集上的最佳模型用于测试集上的评估。

评估指标。我们使用变化类别的F1分数作为主要评价指标。 F1-score由测试的精确率和召回率计算得出,如下:

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

此外,还报告了更改类别的精确度、召回率、并交交集 (IoU) 以及总体准确度 (OA)。上述指标定义如下:

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

其中TP、FP、FN分别表示真阳性、假阳性、假阴性的数量。

B. 与最先进的方法的比较我们与几种最先

进的方法进行比较,包括三种纯基于卷积的方法(FC-EF [22]、FC-Siam-Di [22]]、FC-Siam-Conc [22])和四种基于注意力的方法(DTCDSCN [9]、STANet [2]、IFNet [10] 和 SNUNet [14])。 · FC-EF [22]:图像级融合方法,其中双孔图像作为单个输入连接到完全卷积网络。

· FC-Siam-Di [22]:特征级融合方法,采用 Siamese FCN 提取多级特征,并利用特征差异来融合双时信息。

· FC-Siam-Conc [22]:特征级融合方法,采用 Siamese FCN 提取多级特征,并使用特征级联来融合双时态信息。

· DTCDSCN [9]:多尺度特征级联方法,将通道注意力和空间注意力添加到深层Siamese FCN中,从而获得更具判别性的特征。请注意,他们还在每个时间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Metric-based Siamese FCN的方法,它集成了时空注意力机制以获得更多的判别性特征。

· IFNet [10]:多尺度特征串联方法,将通道注意力和空间注意力应用于解码器每一级的串联双时特征。深度监督(即计算解码器每个级别的监督损失)用于更好地训练中间层。 · SNUNet [14]:多尺度特征级联方法,结合Siamese网络和NestedUNet[55]来提取高分辨率的高级特征。通道注意力应用于特征

在解码器的每个级别。还采用深度监督来增强中间特征的辨别能力。

我们使用具有默认超参数的公共代码来实现上述 CD 网络。

标签。我报告了 LEVIR-CD、WHU-CD 和 DSIFN-CD 测试集的总体比较结果。定量结果表明,我们基于 BIT 的模型在这些数据集中始终优于其他方法,并且具有显着优势。例如,我们的 BIT 的 F1 分数在三个数据集上分别超过最近的 STANet 2/1.6/4.7 点。请注意,我们的 CNN 主干网只是纯 ResNet,我们没有应用 [2] 中的 FPN 或 [9,10,14,22] 中的 UNet 等复杂的结构,这些结构通过融合对像素级预测任务非常强大具有高空间精度的低级特征和高级语义特征。我们可以得出结论,即使使用简单的主干,我们基于 BIT 的模型也可以实现卓越的性能。这可能归因于我们的 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中的特征表示。

三个数据集上的方法的可视化比较如图5所示。为了更好地观察,使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 的模型比其他模型取得了更好的结果。首先,我们基于BIT的模型可以更好地避免由于对象的外观与兴趣变化的外观相似而导致的误报(例如,图5(a),(e),(g),(i))。例如,如图 5(a)所示,大多数比较方法错误地将游泳池区域分类为建筑物变化(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 可以减少此类错误检测。图5(c)中,由于道路颜色相似,传统方法将道路误认为是建筑物变化由于接收场有限,建筑物的行为和这些方法无法排除这些伪变化。其次,我们的BIT还可以很好地处理由季节差异或土地覆盖要素外观变化引起的无关变化(例如图5(b)、(f)和(l))。图 5(f)中建筑物非语义变化的示例说明了我们的 BIT 的有效性,它学习时空域内的有效上下文,以更好地表达真实的语义变化并排除不相关的变化。最后,我们的 BIT 可以针对大范围的变化生成相对完整的预测结果(例如,图 5(c)、(h)和(j))。例如,在图5(j)中,由于接收场有限,某些比较方法无法完全检测到图像2中的大型建筑区域(显示为绿色),而我们基于BIT的模型呈现出更完整的结果。

C. 模型效率和有效性

为了公平地比较模型效率,我们在配备 Intel Xeon Silver 4214 CPU 和 NVIDIA Tesla V100 GPU 的计算服务器上测试了所有方法。标签。

II 报告不同方法在 LEVIR-CD、WHU-CD 和 DSIFN-CD 测试集上的参数数量 (Params.)、每秒浮点运算次数 (FLOPs) 和 F1/IoU 分数。

Machine Translated by Google

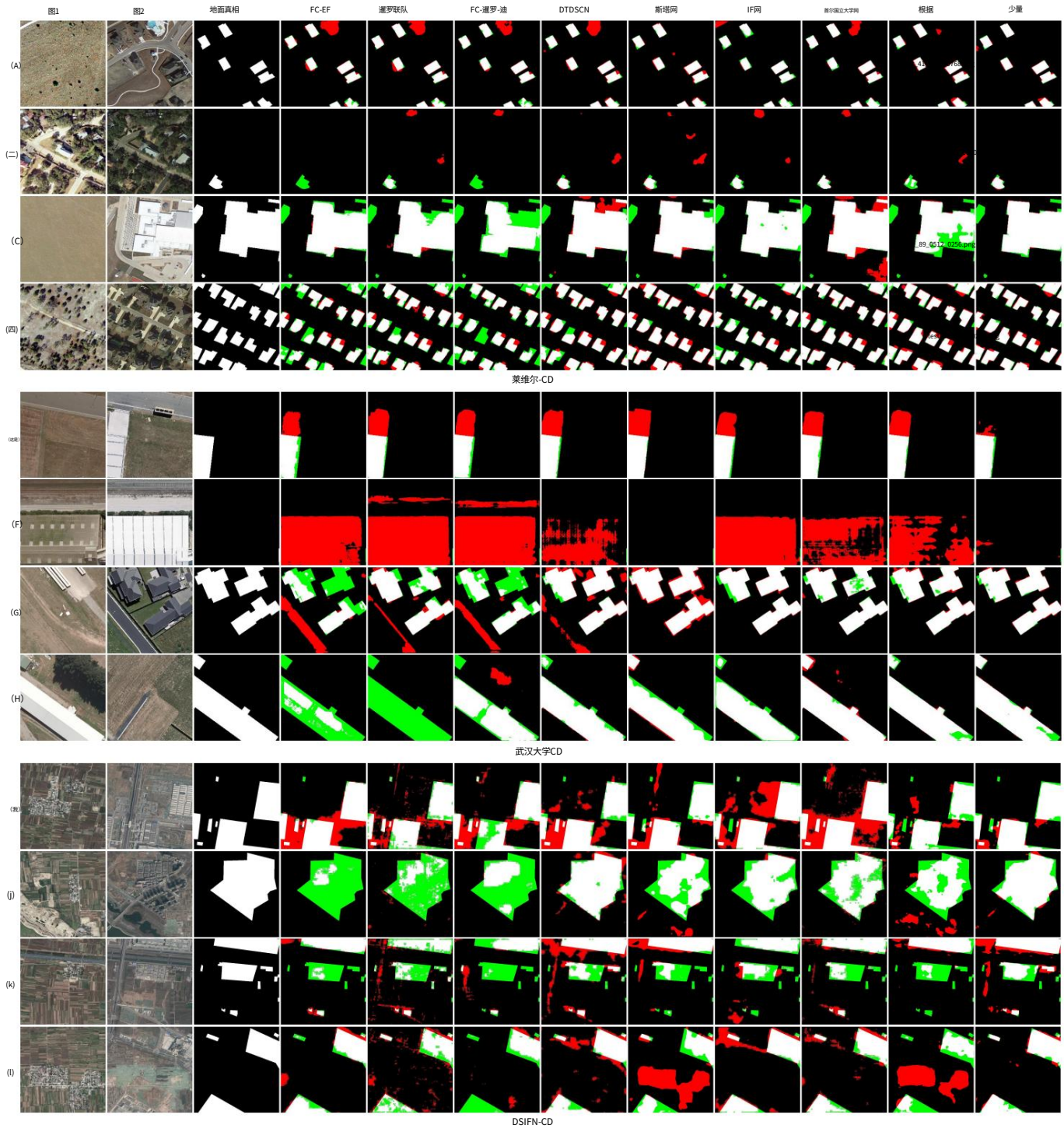


图 5. 不同方法在 LEVIR-CD、WHU-CD 和 DSIFN-CD 测试集上的可视化结果。使用不同的颜色以获得更好的视图，即白色为真阳性，黑色为真阴性，红色为假阳性，绿色为假阴性。

Machine Translated by Google

表一

三个CD测试集的比较结果。最高分以粗体标记。所有分数均以百分比表示

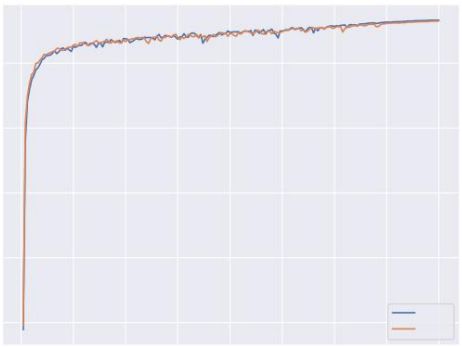
(%)。

	LEVIR-CD Pre。 / Rec. / F1 / IoU / OA	WHU-CD预。 / 推荐。 / F1 / IoU / OA	DSIFN-CD 预。 / 推荐。 / F1 / IoU / OA
FC-EF [22]	86.91 / 80.17 / 83.40 / 71.53 / 98.39	71.63 / 67.25 / 69.37 / 53.11 / 97.61	72.61 / 52.73 / 61.09 / 43.98 / 88.59
FC-暹罗-迪 [22]	89.53 / 83.31 / 86.31 / 75.92 / 98.67	47.33 / 77.66 / 58.81 / 41.66 / 95.63	59.67 / 65.71 / 62.54 / 45.50 / 86.63
暹罗联合FC [22]	91.99 /76.77/83.69/71.96/98.49	60.88/73.58/66.63/49.95/97.04	66.45/54.21/59.71/42.56/87.57
DTDSCN [9]	88.53 / 86.83 / 87.67 / 78.05 / 98.77	63.92 / 82.30 / 71.95 / 56.19 / 97.42	53.87 / 77.99 / 63.72 / 46.76 / 84.91
STA网络 [2]	83.81 / 91.00 / 87.26 / 77.40 / 98.66	79.37 / 85.50 / 82.32 / 69.95 / 98.52	67.71 / 61.68 / 64.56 / 47.66 / 88.49
IF网 [10]	94.02 / 82.93 / 88.13 / 78.77 / 98.87	96.91 / 73.19 / 83.40 / 71.52 / 98.83	67.86 / 53.94 / 60.10 / 42.96 / 87.83
新加坡国立大学网 [14]	89.18 / 87.17 / 88.16 / 78.83 / 98.82	85.60 / 81.49 / 83.50 / 71.67 / 98.71	60.60 / 72.89 / 66.18 / 49.45 / 87.34
根据	88.24 / 86.91 / 87.57 / 77.89 / 98.76	81.80 / 81.42 / 81.61 / 68.93 / 98.53	73.30 / 48.65 / 58.48 / 41.32 / 88.26
少量	89.24 / 89.37 / 89.31 / 80.68 / 98.92	86.64 / 81.48 / 83.98 / 72.39 / 98.75	68.36 / 70.18 / 69.26 / 52.97 / 89.41

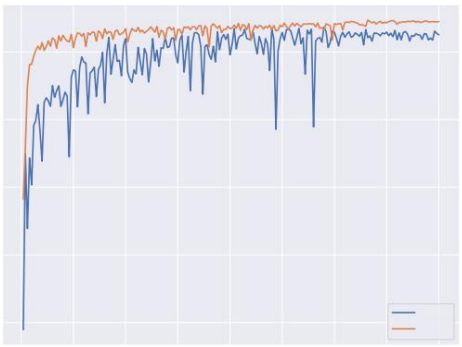
首先，我们通过以下方式验证我们提出的双边投资条约的效率：
比较卷积对应物。标签。二、显示
基于 Base S3/Base S4 构建，该模型添加了 BIT
（BIT S3/BIT S4）比那更有效和高效
（Base S4/Base S5）具有更多卷积层。例如，BIT S4 优于 Base S5 1.7/2.4/10.8 点
三个测试集上的 F1 分数，同时使用 3 次
模型参数数量更少，降低 3 倍
计算成本。此外，我们可以观察到，相比
到Base S4，添加更多的卷积层只会引入
微小的改进（即 F1- 的 0.16/0.75/0.18 点）
三个测试集的得分），而 BIT 的改进
比 CNN 多得多（即 4 60 倍）。

其次，我们对四种基于注意力的方法进行比较
方法（DTCDSCN、STANet、IFNet 和 SNUNet）。作为
如表所示。 II，我们的 BIT S4 在 F1/IoU 分数上优于四个对手，并且有很大的优势
计算复杂度和模型参数小。有趣的是，即使骨干重量轻得多（大约 10 倍）
较小），我们基于 BIT 的模型（BIT S3）仍然优于
大多数数据集上的四种比较方法。比较
结果进一步证明了我们的有效性和效率
基于BIT的模型。

培训可视化。图 6 说明了平均 F1 分数
每个训练时期的训练/验证集。我们
可以观察到，虽然 Base 和 BIT 模型有
训练准确度方面表现相似，BIT 优于
稳定性方面验证准确性的基础
和有效性。这表明BIT的训练是
更稳定、更高效，我们基于BIT的模型有更多
泛化能力。这可能是由于其学习紧凑的能力
上下文丰富的概念，有效地代表了变化
出于兴趣。



(a) LEVIR-CD 数据集上的训练准确性。



(b) LEVIR-CD 数据集的验证准确性。

图 6. 每个训练时期模型的准确性。平均 F1 分数为
报道称。

D. 消融研究

情境建模。我们对 Transformer 进行消融
编码器（TE）验证其在上下文建模中的有效性，
其中多头自注意力是TE中的核心组件
用于建模上下文。从选项卡。三、我们可以观察到一致
从 BIT 中删除 TE 后，LEVIR-CD、WHU-CD 和 DSIFN-CD 数据集上的 F1 分数显着
下降。它
表明TE中的self-attention对于建模至关重要
基于令牌的时空关系。此外，我们替换
我们的 BIT 具有非局部 [56] 自注意力层，即
能够对基于像素的时空内的关系进行建模。这
比较结果见表。 III 显示我们的 BIT 在三个测试集上的表现明显优于 Non-local。它
可能
因为我们的 BIT 在基于令牌的空间中学习上下文，
结构更紧凑，信息密度更高

Machine Translated by Google

表二

模型效率的消融研究。我们报告参数数量(PARAMS.)、每秒浮点运算
（失败），以及三个CD测试集上的F1和IOU分数。模型的输入图像的大小为256 × 256 × 3
计算失败次数。

模型	参数 (M)	失败次数 (G)	莱维尔-CD F1	莱维尔-CD IOU	武汉大学CD F1	武汉大学CD IOU	DSIFN-CD F1	DSIFN-CD IOU
DTDSCN [9]	41.07	7.21	87.67	78.05	71.95	56.19	63.72	46.76
STA网络 [2]	16.93	6.58	87.26	77.40	82.32	69.95	64.56	47.66
IF网 [10]	50.71	41.18	88.13	78.77	83.40	71.52	60.10	42.96
新加坡国立大学网 [14]	12.03	27.44	88.16	78.83	83.50	71.67	66.18	49.45
基地S3	1.28	1.78	82.23	76.24	79.52	66.00	56.00	38.88
+ CNN（基础S4）	3.38	4.09	87.41	77.64	80.86	67.87	58.30	41.15
+ 位（位 S3）	1.45	2.05	88.51	79.39	81.38	68.60	69.00	52.67
底座S4	3.38	4.09	87.41	77.64	80.86	67.87	58.30	41.15
+CNN（基础S5）	11.85	12.99	87.57	77.89	81.61	68.93	58.48	41.32
+位（位S4）	3.55	4.35	89.31	80.68	83.98	72.39	69.26	52.97

表三

我们的位在三个CD数据集上的消融研究。消融是
在Tokenizer (T)、 Transformer Encoder (TE)和
变压器解码器(TD)。我们还将非本地添加到
比较基线。报告 F1 分数。注意
TE和TD的深度设置为1。

模型	T	TE	TD	LEVIR	WHU	DSIFN
底座S4	×	×	×	87.41	80.86	58.30
+非本地	×	×	×	87.56	80.93	59.94
少量				88.93	82.34	67.38
少量	×			87.58	81.68	61.76
少量		×		87.35	81.05	62.93
少量			×	88.07	79.16	64.47
少量		×	×	87.38	80.82	59.54

表四

三CD位置嵌入(PE)的消融研究
数据集。我们在变压器编码器中对PE进行烧蚀
(TE)和变压器解码器(TD)。报告 F1 分数。
请注意， TE和TD的深度设置为1。

TE 中的型号	PE	TD	LEVIR	中的	PE	WHU	DSIFN
少量	×	×	87.77			82.06	60.81
少量		×	88.93			82.34	67.38
少量	×		87.87			81.40	60.23
少量			89.07			82.01	65.68

表五

令牌长度的影响。该位的F1/IOU分数为
在LEVIR-CD、WHU-CD和DSIFN-CD测试集上进行评估。
请注意， TE和TD的深度设置为1。

长度	莱维尔-CD F1	莱维尔-CD IOU	武汉大学CD F1	武汉大学CD IOU	DSIFN-CD F1	DSIFN-CD IOU
32	87.76	78.18	81.53	68.82	62.40	45.35
16	88.45	79.74	81.79	69.19	63.07	46.06
8	88.19	78.88	81.83	69.27	64.28	47.36
4	88.93	80.07	82.34	70.00	67.38	50.80
2	88.90	80.02	82.02	69.53	65.13	48.29

表六

变压器深度的影响。我们进行分析
关于编码器深度（ED）和解码器深度（DD）
位，并报告每个配置的F1/IOU分数
LEVIR -CD、WHU-CD和DSIFN-CD测试集。

EDDD		莱维尔-CD F1	莱维尔-CD IOU	武汉大学CD F1	武汉大学CD IOU	DSIFN-CD F1	DSIFN-CD IOU
1	1	88.93	80.07	82.34	70.00	67.38	50.80
2	1	89.13	80.39	81.83	69.24	66.96	50.34
4	1	88.97	80.13	82.15	69.70	66.95	50.32
8	1	88.93	80.06	80.73	67.68	67.11	50.50
1	2	88.91	80.03	82.99	70.92	67.17	50.57
1	4	89.26	80.59	83.69	71.95	69.05	52.73
1	8	89.31	80.68	83.98	72.39	69.26	52.97

比 Non-local 更好，从而有利于有效提取
的关系。

标记器上的消融。我们通过将 to-kenizer 从 BIT 中删除来对它进行消融。得到的模型
可以考虑使用密集标记，它们是序列
CNN 主干提取的特征。如图所示
标签。 III，基于 BIT 的模型（wo tokenizer）的 F1 分数显着下降。它表明标记器

模块在我们基于变压器的框架中至关重要。我们
可以看到该模型（wo tokenizer）仅稍微好一点
比基础S4。这可能是因为密集的特征包含
过多的冗余信息使得训练
基于变压器的模型是一项艰巨的任务。相反，
我们提出的标记器在空间上汇集密集的特征
聚合语义信息，从而得到紧凑的
概念的标记。

变压器解码器上的烧蚀。为了验证我们的 Transformer Decoder (TD) 的有效性，我们将其替换为
一个简单的模块来融合代币Ti 新的 来自 TE 和
来自 CNN 主干的原始特征Xi 。在简单的
模块中，我们扩展了每个标记的空间维度
的 新的 （包含 L 代币）到 R 形状
扩展的令牌和Xi相加以产生更新的
然后将特征馈送到预测头。标签。三、
表明 BIT 模型的性能持续下降
硬件 。还有L

Machine Translated by Google

在三个测试集上没有 TD。这可能是因为交叉注意力（TD 的核心部分)提供了一种优雅的方式,通过建模它们的关系来增强原始特征与上下文丰富的标记。此外,BIT（包括 TE 和 TD)比正常的 BIT 模型要差得多。

位置嵌入的效果。 Transformer 架构是排列不变的,而 CD 任务需要空间和时间位置信息。为此,我们将学习到的位置嵌入（PE)添加到馈送到变压器的特征序列中。我们在 TE 和 TD 中对 PE 进行消融。我们设置不包含PE的BIT模型作为基线。如表所示。 IV,当将 PE 添加到馈入 TE 的代币中时,我们的 BIT 模型在三个测试集上实现了 F1 分数的一致改进。它表明双时态标记集中的位置信息对于 TE 中的上下文建模至关重要。与基线相比,将 PE 添加到馈入 TD 的查询时,BIT 模型的 F1 分数没有显着改进。位置信息对于TD查询来说可能是不必要的,因为TD中的键（即,令牌)是高度抽象的并且不包含空间结构。因此,我们在BIT模型中只在TE中添加PE,而在TD中不添加PE。

E、参数分析

令牌长度。我们的标记器在空间上将图像的密集特征汇集到一个紧凑的标记集中。我们的直觉是,双时态图像中兴趣的变化可以通过一些视觉概念（即语义标记)来描述。

令牌集合L的长度是一个重要的超参数。
我们分别测试不同的 L ∈ {2, 4, 8, 16, 32} 来分析其对模型在 LEVIR-CD、WHU-CD 和 DSIFN-CD 数据集上的性能的影响。标签。 V 显示当将 token 长度从 32 减少到 4 时,模型的 F1 分数有了显着改善。这表明紧凑的 token 集足以表示兴趣变化的语义概念,冗余 token 可能会影响模型性能。我们还可以观察到,当 L 从 4 进一步减小到 2 时,F1 分数略有下降。这是因为当 L 太短时,模型可能会丢失一些与变化概念相关的有用信息。因此,我们将L设置为4。

变压器的深度。变压器层数是一个重要的超参数。我们测试了 BIT 模型的不同配置,其中包含不同数量的 TE 和 TD 变压器层。标签。 VI 显示,当增加 Transformer 编码器的深度时,三个数据集上的 BIT 的 F1/IoU 分数没有显着改善。这表明单层 TE 可以很好地学习双时态 token 之间的关系。标签。 VI 还表明模型性能与解码器深度大致正相关。这可能是因为在 Transformer 解码器的每一层之后,通过考虑上下文丰富的标记来细化图像特征。当解码器深度为8时获得最佳结果。虽然进一步增加解码器深度可能会有性能增益,但为了效率和精度之间的权衡,我们将编码器深度设置为1,解码器深度设置为8。

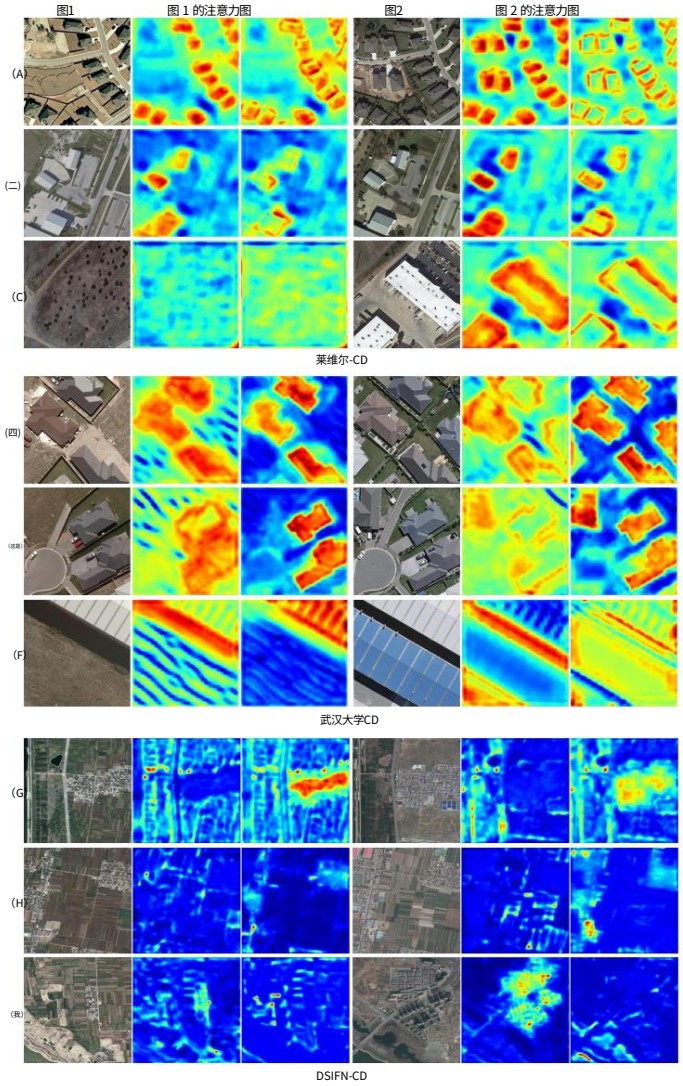


图 7.LEVIR-CD、WHU-CD 和 DSIFN-CD 测试集上的代币可视化。红色表示较高的关注值,蓝色表示较低的值。

F. 代币可视化

我们假设我们的分词器可以提取揭示兴趣变化的高级语义概念。为了更好地理解语义标记,我们将标记器从双时态特征图中提取的注意力图Ai ∈ R 可视化。标记集合Ti中的每个标记Ti对应一个注意力图Ai 。图 7 显示了来自 LEVIR-CD、WHU-CD 和 DSIFN-CD 数据集的一些双时图像的标记可视化结果。我们为每个输入图像显示从Ti中选择的两个标记的注意力图。红色表示较高的关注值,蓝色表示较低的值。 HW ∈ R

从图7中我们可以看到,提取的标记可以关注属于兴趣变化的语义概念的区域。不同的标记可能涉及不同语义的对象。例如,由于 LEVIR-CD 和 WHU-CD 数据集仅描述建筑物变化,因此这些数据集中的学习标记主要关注属于建筑物的像素。而因为 DSIFN-CD 数据集

Machine Translated by Google

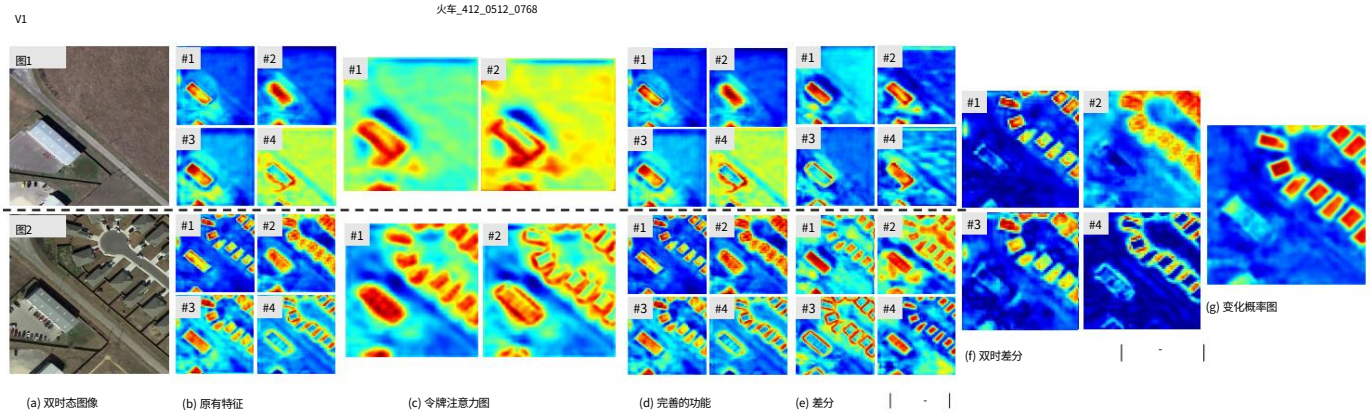


图 8. 网络可视化示例。 (a) 输入图像,(b) 选择的高级特征图Xi ,(c) 通过分词器选择的注意力图Ai ,(d) 精炼特征图X和Xi ,(f) 双时态特征差分图像,(g) 变化概率地图P。样本来自LEVIR-CD数据集。我们使用相同的图例进行可视化（最小，最大）来可视化每个激活图。

包含各种变化,这些标记可以突出不同的语义区域,例如建筑物、农田和水体。有趣的是,如图 7 (c) 和 (f) 所示,我们的分词器还可以突出显示建筑物周围的像素（例如阴影）,即使在训练我们的模型时没有提供对这些区域的明确监督。这并不奇怪,因为建筑物周围的环境是物体识别的关键线索。这表明我们的模型可以隐式学习一些额外的概念来促进变化识别。

无关的改变。时空上下文建模对于增强特征辨别能力至关重要。我们提出的 BIT 模块可以有效地对基于 token 的时空中的上下文信息进行建模,并使用上下文丰富的 token 来增强原始特征。与 Base 模型相比,我们的 BIT-base 模型可以生成更准确的预测,误报更少,召回率更高（见图 5 和表 I）。此外,BIT还可以提高模型训练的效率和稳定性（见图6）。这是因为我们的BIT将图像表达为少量的视觉词（token向量）,这样的高密度信息可以提高训练效率。我们的 BIT 也可以被视为一种有效的基于注意力的方式,以增加模型的接收范围,从而有利于变化识别的特征表示能力。

G. 网络可视化

为了更好地理解我们的模型,我们提供了一个示例来可视化 BIT 模型不同阶段的激活图。给定双时图像（图8（a））,连体FCN生成高级特征图Xi （图 8（b））。

然后,标记器使用学习到的注意力图Ai将特征图在空间上池化为几个标记向量（图 8（c））。
然后,由 Transformer 编码器生成的上下文丰富的标记通过 Transformer 解码器投影回像素空间,从而产生细化的特征图Xi （图 8（d））。我们展示了原始特征X和细化特征Xi的四个相应的代表性特征图。从图 8 (b)和（d）中,我们可以观察到我们的模型可以提取与每个特征的兴趣变化相关的高级特征。时间图像,例如建筑物及其边缘的概念。为了更好地说明BIT模块的效果,精炼后的特征与原始特征之间的差异图像如图8（e)所示。这表明我们的BIT可以进一步突出与变化类别相关的语义概念区域。最后,预测头计算Xi之间的特征差异图像（图8（f）)并生成变化概率图P（图8（g））。

锯。结论

在本文中,我们提出了一种有效的基于变压器的遥感图像变化检测模型。我们的 BIT 学习一组紧凑的标记来表示高级概念,这些概念揭示了双时态图像中存在的兴趣变化。我们利用转换器来关联基于标记的时空中的语义概念。大量的实验验证了我们方法的有效性。我们用 BIT 取代 ResNet18 的最后一个卷积阶段,获得了显着的精度提升（LEVIR-CD/WHU-CD/DSIFN-CD 测试集上的 F1 分数为 1.7/2.4/10.8 分）,计算复杂度降低了 3 倍,模型参数小 3 倍。我们的经验证据表明 BIT 比纯卷积模块更高效、更有效。仅使用简单的 CNN 主干 (ResNet18),我们的方法优于其他几种采用更复杂结构的 CD 方法,例如 FPN 和 UNet。我们还在三个 CD 数据集上展示了比最近四种基于注意力的方法在效率和准确性方面更好的性能。

和习

五、讨论

我们提供了一种高效且有效的方法来执行高分辨率遥感图像的变化检测。

整个时空中同一类别像素的高反射率变化给模型识别感兴趣的物体和区分真实变化带来了困难

参考

[1] A. SINGH，“使用遥感数据回顾数字变化检测技术”,国际遥感杂志,卷。10、不。 6,第 989–1003 页,1989 年。

Machine Translated by Google

[2] H. Chen 和 Z. Shi，“基于时空注意力的方法和用于遥感图像变化检测的新数据集”，
偏僻的。感觉，卷。 12、没有。 10，p。 1662, 2020.
[3] PP de Bem、OA de Carvalho Junior、RF Guimaraes 和 RAT Gomes，“使用陆地卫星数据和卷积神经网络对巴西亚马逊森林砍伐的变化检测”，遥感，卷。 12、没有。 6，p。 901，2020。

[4] JZ Xu、W. Lu、Z. Li、P. Khaitan 和 V. Zaytseva，“使用卷积神经网络在卫星图像中构建损伤检测”，2019 年。

[5] 石伟、张明、张荣、陈世、詹子，“基于人工智能的变化检测:现状与挑战”，遥感，2017年第1期。 12，p。 1688, 2020.

[6] J. Chen、Z. Yuan、J. Peng、L. Chen、H. Huang、J. Zhu、T. Lin 和 H. Li，“Dasnet:用于高通量变化检测的双注意力全卷积暹罗网络”分辨率卫星图像。”

[7] M. 张，G. Xu，K. Chen，M. Yan，X. Sun，“基于三元组的航空遥感图像变化检测语义关系学习”，IEEE Geosci。偏僻的。 Sens. Lett.，卷。 16、没有。 2，第 266-270 页，2019 年。

[8] M. 张和 W. Shi，“一种基于特征差异卷积神经网络的变化检测方法”，TGRS，第 1-15 页，2020 年。

[9] Y. Liu、C. Pang、Z. Zhan、X. Zhang 和 X. Yang，“使用双任务约束深度暹罗卷积网络模型构建遥感图像变化检测”，IEEE 地球科学与遥感《信件》，第 1-5 页，2020 年。

[10] C. 张，P. Yue，D. Tapete，L. Jiang，B. Shangguan，L. Huang，and G. Liu，“一种用于高分辨率双时态遥感变化检测的深度监督图像融合网络”图像”，ISPRS，卷。 166，第 183-200 页，2020 年。

[11] X. Peng，R. Zhu，Z. Li，Q. Li，“基于注意力机制和图像差异的光学遥感图像变化检测”，IEEE 地球科学与遥感学报，第1页- 2020 年 12 月。

[12] H. Jiang，X. Hu，K. Li，J. Zhang，J. Kong，and M.Zhang，“Pga-siamnet:基于金字塔特征的注意力引导暹罗网络，用于遥感正射影像建筑变化保护”，遥感，卷。 12、没有。 3，第 3 页。 484，2020。

[13] FI Diakogiannis、F. Waldner 和 P. Caccetta，“寻求改变?掷骰子并寻求关注。”

[14] S. Fang、K. Li、J. Shao 和 Z. Li，“Snunet-cd:用于 vhr 图像变化检测的密集连接暹罗网络”，
IEEE 地球科学与遥感快报，第 1-5 页，2021 年。
[15] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、AN Gomez、L. Kaiser 和 I. Polosukhin，“Attention is all you need”，《神经信息处理系统进展 30:2017 年神经信息处理系统年会》，2017 年 12 月 4-9 日，美国加利福尼亚州长滩，I. Guyon, U. von Luxburg, S. Bengio, HM Wallach, R. Fergus, SVN Vishwanathan 和 R. Garnett，编辑，2017 年，第 14 页。 5998–6008。
[16] K. Nemoto、T. Imaizumi、S. Hikosaka、R. Hamaguchi、M. Sato 和 A. Fujita，“仅使用 RGB 航空图像通过 CNN 组合进行建筑物变化检测”，2017 年 10 月。

[17] S. Ji，Y. Shen，M. Lu，Y. Zhang，“利用卷积神经网络和模拟样本构建大规模航空图像实例变化检测”，遥感，第 1 卷。 11、没有。 11，p。 1343, 2019.

[18] R. Liu、M. Kuffer 和 C. Persello，“采用基于CNN的变化检测方法的贫民窟的时间动态”，
偏僻的。感觉，卷。 11、没有。 23，p。 2844, 2019.
[19] RC Daudt、BL Saux、A. Boulch 和 Y. Gousseau，“使用卷积神经网络进行多光谱地球观测的城市变化检测”，IGARSS，2018 年。

[20] FU Rahman、B. Vasu、JV Cor、J. Kerekes 和 AE Savakis，“具有多级特征的 Siamese 网络，用于卫星图像中基于补丁的变化检测”，2018 年 IEEE 全球信号和信息处理会议，全球 SIP 2018 年，美国加利福尼亚州阿纳海姆，2018 年 11 月 26 日至 29 日。IEEE，2018 年，第 27 页。 958–962。
[21] M. Wang、K. Tan、X. Jia、X. Wang 和 Y. Chen，“基于多传感器遥感图像的变化检测的具有混合卷积特征提取模块的深度连体网络”，遥感，卷。 12、没有。 2，第 14 页。 205, 2020。

[22] RC Daudt、BL Saux 和 A. Boulch，“用于变化检测的全卷积暹罗网络”，ICIP，2018 年。

[23] MA Lebedev、YV Vizilter、OV Vygolov、VA Knyaz 和 AY Rubis，“使用条件对抗网络进行遥感图像变化检测”，卷。 XLII-2，2018 年，第 565–571 页。

[24] D. Peng、Y.Zhang 和 H.Guan，“使用改进的unet++对高分辨率卫星图像进行端到端变化检测”，
遥感，卷。 11、没有。 11，p。 1382, 2019.
[25] T. Bao，C. Fu，T. Fang，H. Huo，“Ppcnet:一种用于高分辨率遥感图像变化检测的斑块级和像素级端到端深度网络”，卷。 PP，第 1-5 页，2020 年。

[26] B. Hou、Q. Liu、H. Wang 和 Y. Wang，“从 w-net 到 cdgan:通过深度学习技术进行双时态变化检测”，
IEEE 地球科学与遥感学报，卷。 58，没有。 3，第 1790–1802 页，2020 年。

[27] Y. Zhan、K. Fu、M. Yan、X. Sun、H. Wang 和 X. Qiu，“基于光学航空图像深度暹罗卷积网络的变化检测”，IEEE 地球科学与遥感快报，卷。 14，第 1845–1849 页，2017 年。

[28] B. Fang，L. Pan，R. Kou，“基于双时态 VHR 光学遥感图像变化检测的双学习暹罗框架”，遥感，卷。 11、没有。 11，p。 1292, 2019.

[29] W.Zhao、X.Chen、X.Ge 和 J.Chen，“使用对抗网络进行双时态遥感图像中的多重变化检测”，IEEE 地球科学与遥感快报，第 1-5 页， 2020.

[30] H. Chen、W. Li 和 Z. Shi，“用于在遥感图像中构建变化检测的对抗性实例增强”，IEEE 地球科学与遥感学报，第 1-16 页，2021 年。

[31] W. Zhao、L. Mou、J. Chen、Y. Bo 和 WJ Emery，“将度量学习和对抗网络用于季节不变变化检测”，IEEE Trans。地理学。偏僻的。感觉，卷。 58，没有。 4，第 2720–2731 页，2020 年。

[32] K. He、X. Zhang、S. Ren 和 J. Sun，“图像识别的深度残差学习”，2016 年 IEEE 计算机视觉和模式识别会议，CVPR 2016，美国内华达州拉斯维加斯， 2016 年 6 月 27-30 日。IEEE 计算机协会，2016 年，第 770-778 页。

[33] A. Dosovitskiy、L. Beyer、A. Kolesnikov、D. Weissenborn、X. Zhai、T. Unterthiner、M. Dehghani、M. Minderer、G. Heigold、S. Gelly、J. Uszkoreit 和 N. Houlsby，“一张图像相当于 16x16 个单词:用于大规模图像识别的 Transformer。”

[34] H. Touvron、M. Cord、M. Douze、F. Massa、A. Sablayrolles 和 H. Jegou，“通过注意力训练数据高效的图像转换器和蒸馏。”

[35] B. Wu、C. Xu、X. Dai、A. Wan、P. Zhang、M. Tomizuka、K. Keutzer 和 P. Vajda，“视觉变换器:基于标记的图像表示和处理计算机视觉，”CoRR，卷。绝对/2006.03677，2020。

[36] D. 张、H. 张、J. Tang、M. Wang、X. Hua 和 Q. Sun，“特征金字塔变换器”，计算机视觉 – ECCV 2020，A. Vedaldi、H. Bischof、T . 布洛克斯和 J.-M.弗拉姆，编辑。
Cham:施普林格国际出版社，2020 年，第 323-339 页。
[37]郑S.郑，J.卢，H.赵，X.朱，Z.罗，Y.王，Y.Fu，J.Feng，T.Xiang，PHS Torr，和L.Zhang，“重新思考语义使用 Transformer 从序列到序列的角度进行分割。”

[38] N. Carion、F. Massa、G. Synnaeve、N. Usunier、A. Kirillov 和 S. Zagoruyko，“使用 Transformer 进行端到端对象检测”

Machine Translated by Google

计算机视觉 - ECCV 2020 - 第 16 届欧洲会议,英国格拉斯哥,2020 年 8 月 23-28 日,会议记录,第一部分,系列。
计算机科学讲义,A. Vedaldi、H. Bischof、T. Brox 和 J. Frahm,编辑,卷。 12346。施普林格,2020 年,第 213-229 页。

[39] X. Zhu、W. Su、L. Lu、B. Li、X. Wang 和 J. Dai，“Deformable {detr}:用于端到端物体检测的可变形变压器”，国际会议学习表示,2021。[在线]。可用：https://openreview。 net/forum?id=gZ9hCDWe6ke [40] M. Chen、A. Radford、R. Child、J. Wu、H. Jun、D. Luan 和 I. Sutskever，“从像素生成预训练”,论文集第 37 届机器学习国
际会议,ICML 2020,2020 年 7 月 13-18 日,虚拟活动,系列。机器学习研究论文集，卷。 119.PMLR,2020 年,第 1691-1703 页。

[41] P. Esser、R. Rombach 和 B. Ommer，“驯服变压器以实现高分辨率图像合成。”
[42] W. Liu、S. Chen、L.Guo、X. Zhu 和 J. Liu，“Cptr:用于图像字幕的完整变压器网络”。
[43] F. Yang、H. Yang、J. Fu、H. Lu 和 B.Guo，“用于图像超分辨率的学习纹理变换网络”,2020 年 6 月。
[44] H. Chen,Y. Wang,T.Guo,C. Xu,Y. Deng,Z. Liu,S. Ma,C. Xu,C. Xu,and W. Gau，“预训练图像处理变压器，”2020。

[45] Y. Yuan 和 L. Lin，“用于卫星图像时间序列分类的变压器的自监督预训练”，IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,第 1-1 页,2020 年。

[46] Z. Li、G. Chen 和 T. Zhang，“使用多时相多传感器图像进行农作物分类的 cnn-transformer 混合方法”,IEEE 应用地球观测和遥感主题精选杂志,卷。 13,第 847–858 页,2020 年。

[47] J.He、L.Zhao、H.Yang、M.Zhang 和 W.Li，“Hsi-bert:使用 Transformer 双向编码器表示的高光谱图像分类”,IEEE 地球科学与遥感学报,卷。 58,没有。 1,第 165–178 页,2020 年。

[48] Y. Bazi、L. Bashmal、MMA Rahhal、RA Dayil 和 NA Ajlan，“用于遥感图像分类的视觉转换器”,遥感,卷。 13、没有。 2021 年 3 日。

[49] X. Shen、B. Liu、Y. Zhou 和 J. Zhao，“通过变压器和强化学习生成遥感图像标题”，
多米。工具应用,卷。 79,没有。 35-36,第 26 661–26 682 页,2020 年。

[50] Q. Wang、W. Huang、X. Zhang 和 X. Li，“遥感图像字幕的词句框架”,IEEE Trans-actions on Geoscience and Remote Sensing,第 1-12 页,2020 年。
[51] J.德夫林,M.-W。 Chang、K. Lee 和 K. Toutanova，“BERT:用于语言理解的深度双向转换器的预训练”,载于计算语言学协会北美分会 2019 年会议记录：人类语言技术,第 1 卷（长论文和短论文）。明尼苏达州明尼阿波利斯:计算语言学协会,2019 年 6 月,第 4171–4186 页。

[52] TQ Nguyen 和 J. Salazar，“没有眼泪的变形金刚:改善自我注意力的标准化”,CoRR,卷。绝对/1910.05895,2019。

[53] D. Hendrycks 和 K. Gimpel，“高斯误差线性单位
（霜）。”
[54] S. Ji、S. Wei 和 M. Lu，“从开放航空和卫星图像数据集中提取多源建筑的全卷积网络”,IEEE Trans。地球科学与遥感,卷。 57、没有。 1,第 574-586 页，2019 年。

[55] Z. Zhou、MMR Siddiquee、N. Tajbakhsh 和 J. Liang，“Unet++:用于医学图像分割的嵌套 u-net 架构”,《医学图像分析中的深度学习 - 和 - 用于临床决策支持的多模态学习》 - 第四届国际研讨会,DLMIA 2018,和第八届
国际研讨会,ML-CDS 2018,与 MICCAI 2018 联合举行,西班牙格拉纳达,2018 年 9 月 20 日,会议记录,系列。计算机科学讲义,卷。
11045。施普林格,2018 年,第 3-11 页。 [在线的]。可用:https: //doi.org/10.1007/978-3-030-00889-5 1 [56] X. Wang、R. _
Girshick、A. Gupta 和 K. He，“非局部神经网络”， 2018 年 6 月。