Remote Sensing Image Change Detection with Transformers

摘要

本文探讨了遥感变化检测技术在地理信息处理中的重要性，特别是在土地资源管理、环境保护与监测、灾害应急响应以及城市规划与发展等方面的应用。针对高分辨率遥感图像中物体复杂性和光谱特征的多变性带来的挑战，提出了一种创新的双时间图像转换器（BIT）方法。BIT 方法通过引入语义令牌和 Transformer 编码器，实现了在时空域中对上下文的有效建模，显著提高了计算效率。与纯卷积方法相比，BIT 在计算成本和模型参数上显著降低，同时保持了较高的准确性。此外，BIT 方法还能够处理高分辨率遥感图像中的复杂变化，通过上下文丰富的令牌反馈到像素空间，对原始特征进行细化，从而提高了变化检测的准确性。本研究还创新地引入 Segment Anything Model（SAM）进行特征提取，进一步提升了模型的运行效率和准确性。本研究为未来遥感变化检测领域的技术创新提供了新的思路，并验证了所提方法的有效性和潜力。

关键词：遥感图像；变化检测；BIT；Transformer；SAM

1 引言

遥感变化检测（CD）作为地理信息处理的关键技术，对于国家与政府管理具有不可忽视的重要性。随着深度学习技术的不断进步，特别是卷积神经网络（CNN）和自注意力机制的引入，遥感变化检测任务取得了显著的成果。然而，面对高分辨率遥感图像中物体复杂性和光谱特征的多变性，现有的方法仍面临诸多挑战。如下图 1所示。在此背景下，我们提出复现一篇介绍双时间图像转换器（BIT）的论文，主要基于以下几个理由：

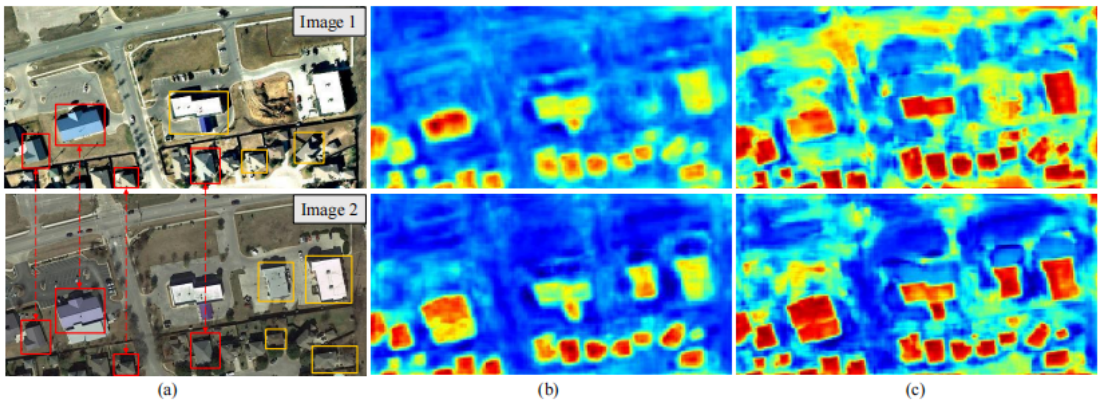

图 1. 本实验 BIT 模块效果的示意图。

首先，土地资源管理：遥感变化检测能够实时、准确地监测土地资源的变化，包括农田侵占、城市扩张 [2]、森林砍伐 [5] 等，为政府提供及时、准确的土地资源管理信息，助力制定科学的土地政策。其次，环境保护与监测：通过监测湖泊、河流、森林等自然环境的变化，遥感变化检测有助于及时发现环境破坏和污染问题，为环境保护和治理提供科学依据。灾害应急响应：自然灾害如洪水、地震后，快速准确地识别受影响区域对于紧急救援行动极为关键。变化检测技术能够迅速对比灾前后的影像资料，帮助相关部门确定受损范围并优先安排援助资源。城市规划与发展：遥感变化检测能够监测城市基础设施的建设和变化，如道路、桥梁、建筑等，为城市规划和发展提供基础数据支持。

本文提出的双时间图像转换器 [8]（BIT）方法通过引入语义令牌和变压器编码器，实现了在时空域中对上下文的有效建模，显著提高了计算效率。与纯卷积方法 [6] 相比，BIT 在计算成本和模型参数上显著降低，同时保持了较高的准确性。同时，它还能够处理高分辨率遥感图像中物体复杂性和光谱特征的多变性，通过上下文丰富的令牌反馈到像素空间，对原始特征进行细化，从而提高了变化检测的准确性。另外，BIT 的设计基于朴素的主干网络（如 ResNet18），没有依赖复杂的结构（如 FPN、UNet 等），因此具有较高的可扩展性和普适性。这意味着 BIT 可以方便地应用于不同的遥感变化检测任务和数据集上。我们还希望通过复现 BIT 论文来验证其有效性和准确性，为遥感变化检测领域的技术创新和发展提供新的思路和方法。通过复现过程，可以深入理解 BIT 的工作原理和关键技术，为未来的研究和应用提供有益的参考。

2 相关工作

2.1 基于深度学习的遥感图像变化检测

随着深度学习的不断发展，许多研究致力于提高网络的特征判别能力，从而提出了多层特征融合结构、基于生成对抗网络（GAN）的优化目标以及增加接收场 [22] [17]（Receptive Field, RF）等方法来更好地进行时空上下文建模。特别是在高分辨率遥感图像的变化检测中，上下文建模对于识别场景中的兴趣变化至关重要。为了扩展接收场的大小，现有方法主要使用更深层次的 CNN 模型（如 ResNet）提取特征，并通过扩张卷积来扩大接收场。除此之外，还会引入注意力机制例如通道注意、空间注意和自注意力等来进一步扩大模型的接收场。例如，一些研究者采用深度 CNN 骨干（ResNet101）提取图像特征，并通过扩张卷积来增强接收场的范围。然而，尽管这些方法可以有效扩大接收场，但仍然存在一定的局限性，在处理大量像素时计算效率较低，且随着像素数量增加，计算复杂度呈二次增长。自注意力机制 [3] [4] 因其能够有效建模时空像素之间的全局关系，在变化检测任务中表现出较好的性能。然而，这些方法大多仍然面临着如何高效利用时间信息和上下文信息的挑战。多数研究要么将注意力机制单独视为每个时间点图像的特征增强模块，要么简单地在通道或空间维度重新加权融合双时特征。

本文的主要创新点在于通过高效和有效的方式学习和利用双时相图像中的全局语义信息来增强变化检测性能。不同于现有的基于注意力机制的 CD 方法，我们提出了一种新的方法通过提取图像中的语义标记，并在基于标记的时空上下文中进行建模，从而避免了直接在像素空间中建模密集的时空关系。通过利用生成的丰富上下文信息，我们能够在像素空间中增强原始特征，提升变化检测的准确性和效率。我们认为场景中的变化可以通过一些视觉词（标

记）来描述，而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，我们的方法不仅提升了计算效率，而且在处理高分辨率遥感图像中的复杂变化时展现出较强的性能。

2.2 基于 SAM 的模型优化

在深度神经网络的实际应用中，遇见的一个主要瓶颈是对大量高质量标注训练数据 [1] 的需求，特别是在语义分割和变化检测（CD）等密集预测任务中。为了应对这一挑战，近年来计算机视觉领域涌现出一种新兴趋势，即通过指定用户提示来泛化基础视觉大模型来应对常见的视觉任务。该趋势核心源自 SAM [16] 的诞生。它是一种分割大模型，在数百万个带注释的图像上进行训练能够实现对“未见过”图像和物体的 zero-shot 泛化。通过提供用户提示例如位置指示或者文本描述等，SAM 能够在推理过程中分割出感兴趣的物体。与 SAM 类似，SegGPT [18] 也声称具备 zero-shot 识别能力可以对常见视觉图像进行高效识别。此外，SEEM 则进一步扩展了用户提示的形式，其中包括点、文本、音频等多种灵活的提示方式。尽管 SAM 展现了优异的泛化能力，但其对计算资源的需求较高。因此，FastSAM [14] 应运而生在以实时速度进行物体分割的同时保持与 SAM 相当的泛化性能，但是其推理速度是 SAM 的 50 倍。如下图 2所示 FastSAM 特性的适配器网络。

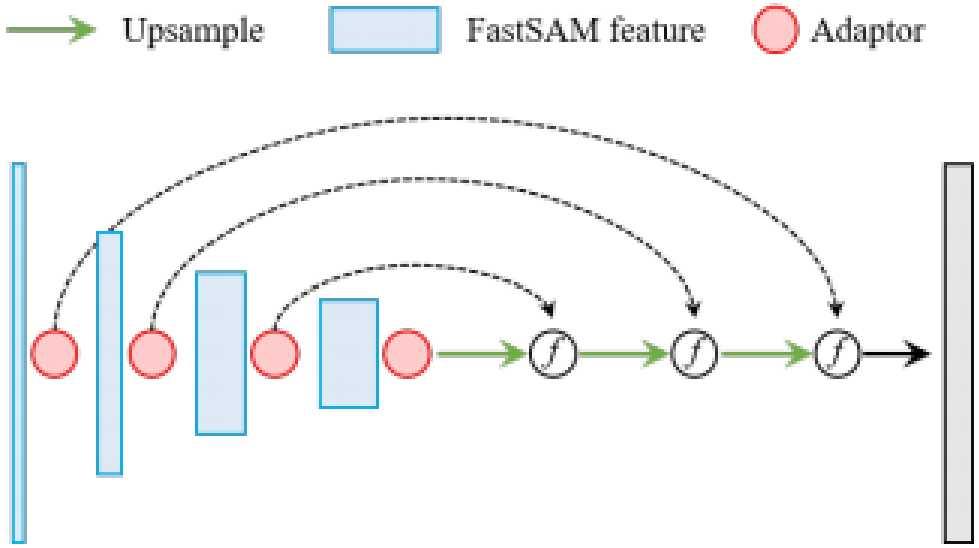

图 2. FastSAM 特性的适配器网络

尽管视觉基础模型（VFMs）宣称能够分割任何物体。可是，它们在某些特定领域仍表现出一定的局限性，特别是在医学图像、制造场景和遥感影像（RSIs）[12] 等应用中。由于 VFMs 主要在自然图像上进行训练，它们往往对前景物体具有较强的聚焦能力，而在处理小型或不规则物体时表现不佳。另一个重要的限制是，VFMs 通常不能提供与分割掩膜相关联的语义类别，这使得其在一些特定应用中缺乏足够的解释能力。针对这一问题，一种可行的解决方案是通过文本提示生成分割样本，并逐步映射出所有类别。在本研究中，我们通过自适应调整方法对 VFM 进行微调，使其能够学习高分辨率遥感影像中的语义潜在特征，从而有效提升模型在遥感影像语义分割任务中的表现。

3 本文方法

3.1 本文方法概述

本文针对高分辨率遥感图像变化检测（CD）中存在的挑战，特别是由于场景中物体的复杂性导致相同语义概念在不同时间和空间位置表现出不同的光谱特征的问题提出了解决方法。尽管基于深度学习的卷积方法已经在这一领域取得了一定进展，但它们在处理时空远程概念关联时仍然存在不足。同时，虽然非局部自注意力机制能够通过建模像素间密集关系来改善性能，但是其计算效率较低。

为解决上述问题，本研究提出了一种新的双时间图像转换器（BIT），它旨在更有效地对时空上下文进行建模。该模型的核心思想是利用少量视觉词，或称作“语义令牌”来表达兴趣变化中的高级概念。具体实现方法是，首先将两时相图像表示为一系列符号，并运用 transformer 编码器在紧凑的基于符号的时空域内建立上下文联系；其次，通过 transformer 解码器将学到的富含上下文信息的令牌反馈到原始像素级特征图中，从而细化特征表达，如图 3所示。此外，该方法被整合进一个基于深层特征差异的变化检测框架中。

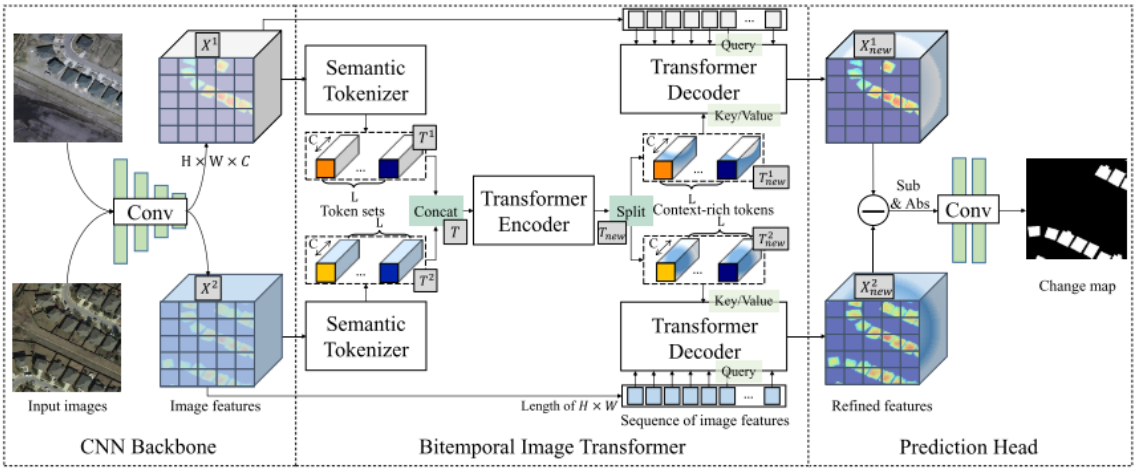

图 3. 语义标记器将由 CNN 主干提取的图像特征汇聚成一个紧凑的标记词汇集（L « HW）。随后，将双时间标记连接并输入至 transformer 编码器中，以在基于标记的时空上下文中建立概念间的联系。由此产生的每个时间点图像的上下文标记被投影回像素空间，并通过 transformer 解码器来精细化原始信号特征。最终，在预测阶段，通过将计算得到的特征差异图像送入浅 CNN 中，从而生成像素级的变化检测结果。

3.2 语义标记器

本实验认为输入图像中的变化可以通过一些高级概念来描述，并将这些概念称为语义标记。这些语义概念可以在双时序图像之间共享。因此，我们采用了一个 Siamese 结构的标记器从每个时序图像的特征图中提取紧凑的语义标记。它有些类似于自然语言处理中的标记器可以将输入句子分割成多个元素例如单词或短语，并用标记向量表示每个元素。本实验的语义标记器将整个图像分割成多个视觉单词，每个视觉单词对应一个标记向量。如图 4所示，为了获得紧凑的标记，我们的语义标记器学习了一组空间注意力图用于空间池化特征图，从而生成一组语义特征，即语义标记集。

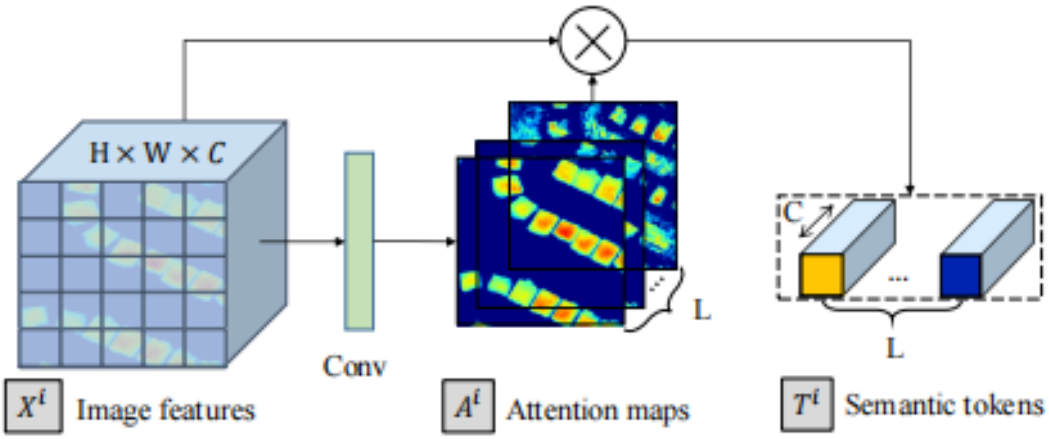

图 4. 语义标记器的图解

具体来说，假设 $X_1, X_2 \in \mathbb{R}^{HW \times C}$ 是输入的双时序特征图，其中 H、W 和 C 分别表示特征图的高度、宽度和通道维度。而 $T_1, T_2 \in \mathbb{R}^{L \times C}$ 则表示两个标记集，其中 L 是标记集的大小，即标记的词汇表大小。对于每个时序特征图 X_i 中的像素 $X_i^p (i = 1, 2)$, 我们使用逐点卷积操作来生成 L 个语义组，每个语义组对应一个语义概念。然后，通过对每个语义组在 HW 维度上应用 softmax 函数，计算出空间注意力图。最终，我们使用这些注意力图对像素进行加权平均，从而得到一个紧凑的语义标记集 T_i, 其大小为 L。

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

形式上，假设 $\phi(\cdot)$ 表示带有学习核 $W \in \mathbb{R}^{HW \times L}$。语义标记 T_i 通过将注意力图 A_i 与特征图 X_i 相乘得到。通过这种方式，语义标记器可以有效地提取图像中的高层语义特征，并将其表示为紧凑的标记集。这种方法为后续的双时序图像处理和变化检测任务提供了有效的语义表示。

4 复现细节

4.1 与已有开源代码对比

本次复现工作的基础内容，我们再次实现双时间图像转换器（BIT）来有效建模双时间图像中的远程上下文关系。BIT 的核心思想是利用少量语义令牌（即视觉词）来表示高级概念的变化，而非在像素空间中进行密集像素关系建模。具体而言，我们将输入图像压缩为高级语义标记，通过基于标记的紧凑时空建模上下文，同时增强每个像素和语义标记之间的关系来提升原始像素空间的特征表达。在该复现过程中，我们将 BIT 集成到基于深度特征差异的变化检测（CD）框架中。

模型总体架构流程如下：首先，通过 CNN 例如 ResNet 从输入图像对中提取高级语义特征，然后使用空间注意机制将每个时间的特征映射转化为紧凑的语义令牌。接着，利用 Transformer 编码器对两组语义令牌间的上下文信息进行建模。生成的上下文丰富的语义令牌再通过 Siamese Transformer 解码器回投影至像素空间来增强原始的像素级特征。最终，通过计算两个增强特征映射的特征差异图（FDI）并输入浅层 CNN [13]，以获得像素级的变化预测。

本次复现工作的创新内容，我们利用 Segment Anything Model（SAM）的强大视觉识别能力，改进高分辨率遥感图像（VHR RSI）的变化检测（CD）。具体而言，我们计划将 BIT 前的卷积神经网络替换为 SAM，以期提升模型的运行效率和准确性。通过这一改进，能够更快速地在多种视觉场景中应用变化检测任务。

在本次研究中，我们致力于复现先前的工作，并在此过程中，我们参考了本论文的开源代码作为我们复现的基础，链接如下https://github.com/sstary/SSRS/tree/main/SAM_RS

4.2 实验环境搭建

个人的电脑配置和服务器配置如下表 1 和表 2 所示：

操作系统	Windows 10
CPU	Intel(R) Core(TM) i7-7700 CPU @ 3.60GHz
机带 RAM	16GB
内存	32GB
存储器容量	1TB

表 1. 个人电脑系统配置

CPU	12 x Xeon Gold 6271
GPU	NVIDIA Tesla P100-16GB
内存	16GB
内存	48GB
存储	1.7TB

表 2. 服务器配置

4.3 创新点

- 引入 transformer 架构进行变化检测：本文提出了一种创新的方法，将 Transformer 架构 [21] [7] 应用于遥感图像的变化检测任务 [9] 中。这种方法能够更有效地对双时间图像中的上下文信息进行建模，从而提高识别图像变化的能力，并有效排除不相关的变化。基于标记的紧凑时空上下文建模：不同于传统的像素级密集关系建模方式，我们提出了一种新的方法——双时间图像转换器（BIT）。该方法通过将输入图像抽象为一系列视觉单词或标记，在一个更加紧凑的基于标记的空间内建模时空上下文。这样不仅减少了计算复杂度，还提高了模型处理远程依赖关系的能力。

- 引入 SAM 架构进行特征提取。利用 Segment Anything Model（SAM）的强大视觉识别能力来改进高分辨率遥感图像的变化检测。我们将 BIT 前的卷积神经网络替换为 SAM 来提升模型的运行效率和准确性。令本实验能够更快速地在多种视觉场景中应用变化检测任务。

5 实验结果分析

本实验主要利用 LEVIR-CD [20] 公开数据集进行训练测试。首先复现了源论文中的代码，验证了 transformer 架构对于遥感图像的变化检测任务帮助，有效地对双时间图像中的上下文信息进行建模。如下图 5所示为实验预测图像的对比图，2012 年到 2016 年的变化检测图像。可以清晰地观察到预测的图像和标签的变化检测图非常接近，很好的反映了图片中 2012 年到 2016 年的建筑和森林的变化信息。

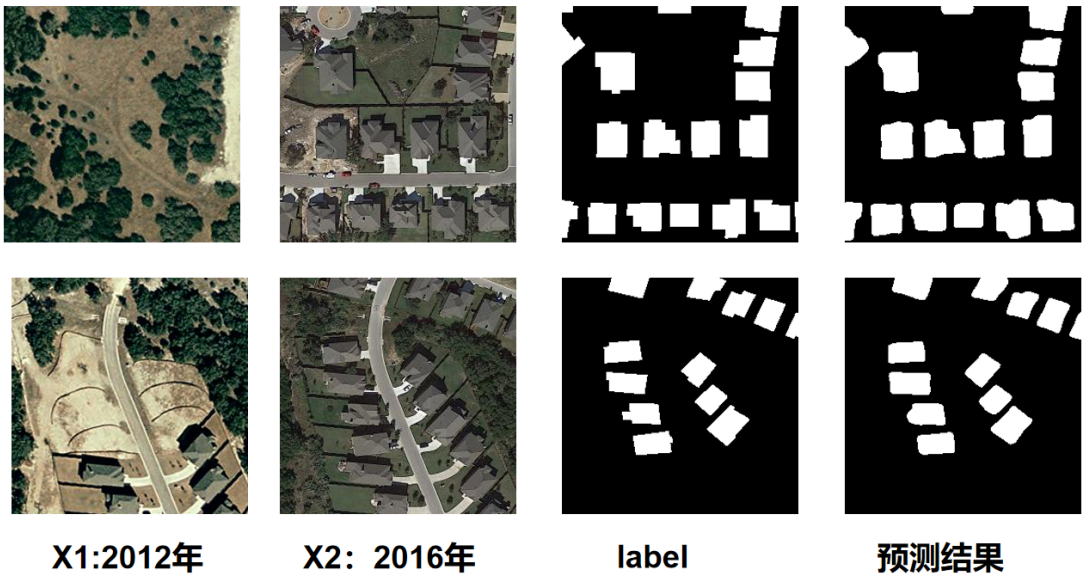

图 5. 实验预测图像的对比图

还显示 BIT 方法的效率会比卷积模型更加高效，准确率和 f1 分数都相对增加。如下表 3 所示为各模型的数据对比。

表 3. Performance comparison of Base and BIT models.

Model	Acc	IoU 2	F1	Pre	Rec
Base	94.42	78.91	82.35	83.36	85.58
BIT	**94.99**	**79.87**	**86.15**	**84.82**	**86.20**

同时，本实验的创新通过修改模型中的卷积部分，引入 SAM 来提高的图像的特征提取能力 [15]。SAM 优良的分割能力为后续的图像特征的处理打下了坚实的基础。本实验创新之后的实验数据结果与复现源代码的结果的还有论文中的实验结果的对比如下表 4。观察对比表可以发现，本实验的创新相对于源代码复现的准确率，F1 分数等都微高，但是对比原论文的实验结果有一些差距。说明本实验还是需要改进，更深入的查询其中的问题优化模型，提高效率。

表 4. Comparison Table of BIT, Innovation, and Original Paper

Model	Acc	IoU 2	F1	Pre
BIT	**94.99**	**79.87**	**86.15**	**84.82**
Innovation	95.42	79.91	86.35	84.36
Original Paper	97.42	80.91	87.35	86.09

6 总结与展望

本实验验证了双时间图像转换器（BIT）在遥感变化检测任务中的应用和效果。通过 BIT 方法引入语义令牌和 tranSformer 编码器 [19] [10] [11]，在时空域中对上下文进行有效建模。证明了这种方法可以提高计算效率和准确性。本研究的创新之处：利用 Segment Anything Model（SAM）进行特征提取。该方法提升模型对遥感图像的细节分割的准确率，给遥感图像变化监测领率带来新的思路。但是由于遥感图像不同于图像，存在一定研究领域关注误差，需要我们努力提升它在遥感图像的适配性来优化模型。我们将探索 SAM 在其他遥感任务中的应用，研究如何将这些技术应用于更广泛的领域。

参考文献

[1] W. G. C. Bandara, N. G. Nair, and V. M. Patel. Ddpm-cd: Remote sensing change detection using denoising diffusion probabilistic models. *arXiv preprint arXiv:2206.11892*, 2022.

[2] H. Chen and Z. Shi. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sensing*, 12(10):1662, 2020.

[3] H. Chen and Z. Shi. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sensing*, 12(10):1662, 2020.

[4] J. Chen, Z. Yuan, J. Peng, L. Chen, H. Huang, J. Zhu, T. Lin, and H. Li. Dasnet: Dual attentive fully convolutional siamese networks for change detection of high resolution satellite images. *Remote Sensing*.

[5] P. P. de Bem, O. A. de Carvalho Junior, R. F. Guimaraes, and R. A. T. Gomes. Change detection of deforestation in the brazilian amazon using landsat data and convolutional neural networks. *Remote Sensing*, 12(6):901, 2020.

[6] F. I. Diakogiannis, F. Waldner, and P. Caccetta. Looking for change? roll the dice and demand attention. *arXiv preprint arXiv:2107.12728*, 2021.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 100–110.

[8] S. Fang, K. Li, J. Shao, and Z. Li. Snunet-cd: A densely connected siamese network for change detection of vhr images. *IEEE Geoscience and Remote Sensing Letters*, pages 1–5, 2021.

[9] J. He, L. Zhao, H. Yang, M. Zhang, and W. Li. Hsi-bert: Hyperspectral image classification using the bidirectional encoder representation from transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 58(1):165–178, 2020.

[10] D. Hendrycks and K. Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2013.

[11] S. Ji, S. Wei, and M. Lu. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1):574–586, 2019.

[12] W. Ji, J. Li, Q. Bi, W. Li, and L. Cheng. Segment anything is not always perfect: A study on different real-world applications. *arXiv preprint arXiv:2305.02450*, 2023.

[13] H. Jiang, X. Hu, K. Li, J. Zhang, J. Gong, and M. Zhang. Pgasiamnet: Pyramid feature-based attention-guided siamese network for remote sensing orthoimagery building change detection. *IEEE Transactions on Geoscience and Remote Sensing*.

[14] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W-Y. Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[15] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W-Y. Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023.

[17] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang. Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model. *IEEE Geoscience and Remote Sensing Letters*, pages 1–5, 2020.

[18] L. Mou, L. Bruzzone, and X. X. Zhu. Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 57(2):924–935, 2018.

[19] T. Q. Nguyen and J. Salazar. Transformers without tears: Improving the normalization of self-attention. *CoRR*, abs/1910.05895, 2019.

[20] A. Singh. Review article: Digital change detection techniques using remotely-sensed data. *International Journal of Remote Sensing*, 10(6):989–1003, 1989.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 5998–6008, Long Beach, CA, USA, December 4-9, 2017 2017.

[22] M. Zhang and W. Shi. A feature difference convolutional neural network-based change detection method. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, pages 1–15, 2020.