
RS-vHEAT: HEAT CONDUCTION GUIDED EFFICIENT REMOTE SENSING FOUNDATION MODEL

A PREPRINT

Huiyang Hu^{1,2,3,4}, Peijin Wang^{1,2,3,4}, Hanbo Bi^{1,2,3,4}, Boyuan Tong^{1,2,3,4}, Zhaozhi Wang^{5,6}, Wenhui Diao^{1,2,3,4}, Hao Chang^{1,2,3,4}, Yingchao Feng^{1,2,3,4}, Ziqi Zhang⁷, Yaowei Wang^{5,7}, Qixiang Ye^{5,6}, Kun Fu^{1,2,3,4}, and Xian Sun^{1,2,3,4}

¹Aerospace Information Research Institute, Chinese Academy of Sciences

²School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences

³University of Chinese Academy of Sciences

⁴Key Laboratory of Target Cognition and Application Technology

⁵Peng Cheng Laboratory

⁶School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences

⁷State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences

⁸Harbin Institute of Technology (Shenzhen)

ABSTRACT

Remote sensing foundation models largely break away from the traditional paradigm of designing task-specific models, offering greater scalability across multiple tasks. However, they face challenges such as low computational efficiency and limited interpretability, especially when dealing with large-scale remote sensing images. To overcome these, we draw inspiration from heat conduction, a physical process modeling local heat diffusion. Building on this idea, we are the first to explore the potential of using the parallel computing model of heat conduction to simulate the local region correlations in high-resolution remote sensing images, and introduce RS-vHeat, an efficient multi-modal remote sensing foundation model. Specifically, RS-vHeat 1) applies the Heat Conduction Operator (HCO) with a complexity of $O(N^{1.5})$ and a global receptive field, reducing computational overhead while capturing remote sensing object structure information to guide heat diffusion; 2) learns the frequency distribution representations of various scenes through a self-supervised strategy based on frequency domain hierarchical masking and multi-domain reconstruction; 3) significantly improves efficiency and performance over state-of-the-art techniques across 4 tasks and 10 datasets. Compared to attention-based remote sensing foundation models, we reduce memory usage by 84%, FLOPs by 24% and improves throughput by 2.7 times. The code will be made publicly available.

Keywords Remote sensing foundation model · Self-supervised learning · Heat conduction · Remote sensing

1 Introduction

Recently, remote sensing (RS) technology has become a vital data source for scientific research, resource management, and environmental monitoring Sherrah [2016], Zhang et al. [2021], Sun et al. [2022a], Li et al. [2020a] by capturing surface information via satellites. Traditional models, designed as single-task networks for specific RS tasks Lu et al. [2021], Sun et al. [2021], struggle with multi-payload, multi-resolution, multi-temporal, and multi-feature RS data FU et al. [2021]. However, the emergence of remote sensing foundation models (RSFMs) has overcome these limitations, enabling unified handling of multiple tasks and diverse scenarios, significantly enhancing the scalability and versatility of the models Li et al. [2021], Manas et al. [2021], Mall et al. [2023], Ayush et al. [2021], Cong et al. [2022], Wang et al. [2022a], Tao et al. [2023], Reed et al. [2023], Mendieta et al. [2023], Bastani et al. [2023]. By constructing visual encoders, RSFMs can automatically extract and learn features from remote sensing imagery (RSI), providing a robust foundation for various real-world RS tasks.

RS-VHEAT: HEAT CONDUCTION GUIDED EFFICIENT REMOTE SENSING FOUNDATION MODEL

一个预印本

Huiyang Hu^{1,2,3,4}, Peijin Wang^{1,2,3,4}, Hanbo Bi^{1,2,3,4}, Boyuan Tong^{1,2,3,4}, Zhaozhi Wang^{5,6}, Wenhui Diao^{1,2,3,4}, Hao Chang^{1,2,3,4}, Yingchao Feng^{1,2,3,4}, Ziqi Zhang⁷, Yaowei Wang^{5,7}, Qixiang Ye^{5,6}, Kun Fu^{1,2,3,4}, 和 Xian Sun^{1,2,3,4}

1中国科学院航空航天信息研究所 2中国科学院大学电子、电气与通信工程学院 3中国科学院大学 4目标认知与应用技术重点实验室 5鹏城实验室 6中国科学院大学电子、电气与通信工程学院 7多模态人工智能系统国家重点实验室 中国科学院自动化研究所 8哈尔滨工业大学（深圳）

5
2
0
2
r
a

M

7

V
C

s
c
l

2
v
4
8
9
7
1
1
4
2
:
v
i

X
r
a

摘要

遥感基础模型在很大程度上脱离了传统为特定任务设计模型的范式，提供了在多个任务上的更大可扩展性。然而，它们面临着计算效率低和可解释性有限等挑战，尤其是在处理大规模遥感图像时。为克服这些挑战，我们从热传导这一物理过程汲取灵感，该过程用于模拟局部热扩散。基于这一想法，我们首次探索了使用热传导的并行计算模型来模拟高分辨率遥感图像中的局部区域相关性，并引入了RS-vHeat，一种高效的多模态遥感基础模型。具体来说，RS-vHeat 1) 应用了复杂度为 $O(N^{1.5})$ 且具有全局感受野的热传导算子 (HCO)，在减少计算开销的同时捕获遥感对象结构信息以指导热扩散；2) 通过基于频域分层遮罩和多域重建的自监督策略学习各种场景的频率分布表示；3) 在4个任务和10个数据集上显著提高了效率和性能。与基于注意力的遥感基础模型相比，我们减少了84%的内存使用、24%的FLOPs，并将吞吐量提高了2.7倍。代码将公开发布。

Keywords 遥感 · 基础模型 · 自监督学习 · 热传导 · 在 · 遥感

1 介绍

最近，遥感 (RS) 技术已成为科学研究、资源管理及环境监测的重要数据来源 Sherrah [2016], Zhang 等 [2021], Sun 等 [2022a], Li 等 [2020a]，通过卫星捕获地表信息。传统模型是为特定 RS 任务设计的单任务网络 Lu 等 [2021], Sun 等 [2021]，难以处理多载荷、多分辨率、多时相和多特征 RS 数据 FU 等 [2021]。然而，遥感基础模型 (RSFMs) 的出现克服了这些限制，使模型能够统一处理多种任务和多种场景，显著增强了模型的可扩展性和通用性 Li 等 [2021], Manas 等 [2021], Mall 等 [2023], Ayush 等 [2021], Cong 等 [2022], Wang 等 [2022a], Tao 等 [2023], Reed 等 [2023], Mendieta 等 [2023], Bastani 等 [2023]。通过构建视觉编码器，RSFMs 可以自动从遥感影像 (RSI) 中提取和学习特征，为各种实际 RS 任务提供坚实的基础。

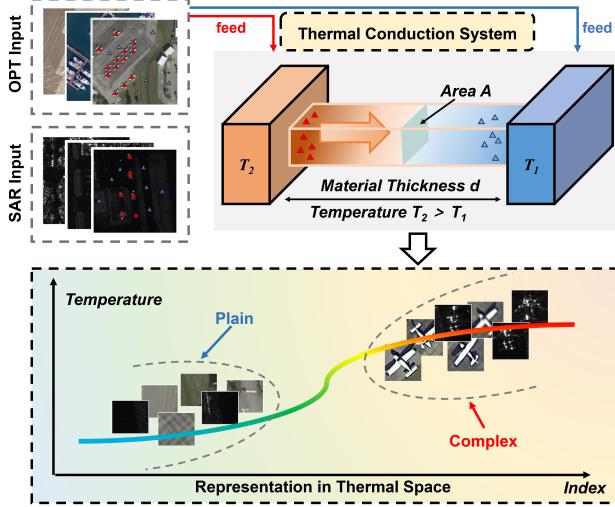


Figure 1: Calculation of RS images in the heat conduction system. In heat conduction theory, different materials exhibit varying diffusion rates. Inspired by this, the heat-conduction-based encoder maps optical (OPT) and synthetic aperture radar (SAR) into a unified thermal space, enhancing sensitivity to shape features by simulating heat flow and accumulation within the object region.

Prior RSFMs typically use visual encoders and decoders, leveraging large-scale RS datasets for pre-training while incorporating self-supervised learning strategies. Specifically, RSFMs can be categorized based on the type of backbone used: CNN-based methods often use ResNet18/50 He et al. [2016] to extract features Li et al. [2021], Manas et al. [2021], Mall et al. [2023], Ayush et al. [2021], while attention-based methods mainly use ViT Dosovitskiy et al. [2020] or Swin transformer Liu et al. [2021] with attention mechanisms Vaswani et al. [2017] to capture global dependencies Cong et al. [2022], Wang et al. [2022a], Tao et al. [2023], Reed et al. [2023], Mendieta et al. [2023], Bastani et al. [2023]. Both approaches use strategies such as masked reconstruction, knowledge distillation, or contrastive learning during pre-training to enhance model robustness (a detailed comparison is available in the supplementary materials). Despite the significant advancements achieved by these RSFMs, they still face two limitations:

Balancing Efficiency and Receptive Field. To accurately capture information about the large objects in RSI, model outputs must be responsive to sufficiently large regions Luo et al. [2016]. However, this need significantly increases computational complexity Christophe et al. [2011], Ma et al. [2015]. CNN-based networks lack a global receptive field due to their reliance on sliding computations with fixed-size convolutional kernels. While attention-based models achieve global modeling, their attention mechanisms incur quadratic computational complexity. Therefore, existing RSFMs struggle to deliver both fast and high-accuracy inference in practical applications.

Weak Physical Interpretability. RS objects often exhibit irregular polygonal shapes Li and Narayanan [2003], and current RSFMs struggle to integrate physical principles to explain how object features propagate. This deficiency makes it challenging for researchers to adjust learning strategies effectively Shen et al. [2022], Temenos et al. [2023], Pérez-Suay et al. [2020]. In the long term, RSFMs need to possess a certain degree of information interpretability.

To address these, this paper introduces RS-vHeat, a heat-conduction-based RSFM that supports multi-modal inputs inspired by the idea of vHeat Wang et al. [2024a]. **First**, heat conduction represents the natural process of energy diffusion from high- to low-temperature regions, as shown in the conceptual physical model in fig. 1. This process transitions from an unsteady to a steady state based on the material. Since its computational process resembles feature extraction in neural networks, it can be applied to RS image processing. **Second**, we hypothesize that object types correspond to special feature distributions, with the model predicting diffusion rates based on RS-specific properties, simulating parameter computation through heat flow. This approach projects all modalities into a common thermal space, following the constraint that complex areas containing RS objects are high-temperature regions where heat accumulates, while sparse regions are low-temperature areas where heat diffuses easily, as shown in the lower part of fig. 1. **Third**, building on this theory, we further design a RSFM with physical interpretability using 3 million optical and SAR data for pre-training, as shown in fig. 2. The heat conduction network simulates the diffusion process of heat across large-scale multi-modal RS data, facilitating the alignment of feature propagation with the structural characteristics of the objects. Moreover, its computational approach provides guidance for the efficient operation of the network.

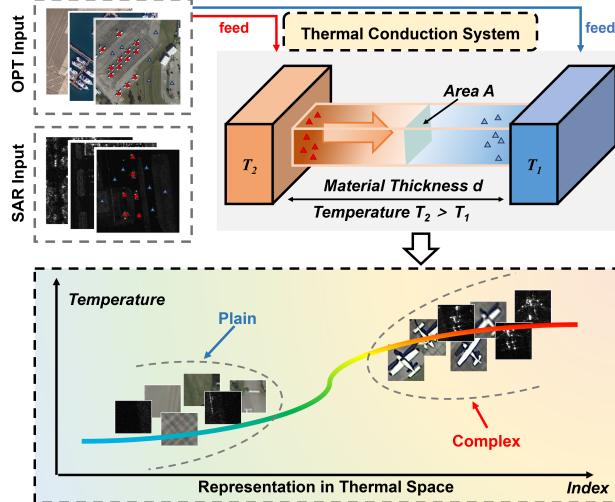


图1：热传导系统中 RS 图像的计算。在热传导理论中，不同材料表现出不同的扩散率。受此启发，基于热传导的编码器将光学（OPT）和合成孔径雷达（SAR）映射到统一的热空间，通过模拟对象区域内热量的流动和积累来增强对形状特征的敏感性。

Prior RSFMs 通常使用视觉编码器和解码器，利用大规模 RS 数据集进行预训练，同时结合自监督学习策略。具体而言，RSFMs 可以根据所使用的骨干网络类型进行分类：基于 CNN 的方法通常使用 ResNet18/50 (He et al. [2016]) 提取特征 (Li et al. [2021], Manas et al. [2021], Mall et al. [2023], Ayush et al. [2021]), 而基于注意力的方法主要使用 ViT (Dosovitskiy et al. [2020]) 或 Swin 变形器 (Liu et al. [2021]) 结合注意力机制 (Vaswani et al. [2017]) 来捕获全局依赖关系 (Cong et al. [2022], Wang et al. [2022a], Tao et al. [2023], Reed et al. [2023], Mendieta et al. [2023], Bastani et al. [2023])。两种方法在预训练过程中都使用了掩码重建、知识蒸馏或对比学习等策略来增强模型的鲁棒性（详细比较请参见补充材料）。尽管这些 RSFMs 取得了显著的进步，但它们仍然面临两个限制：

平衡效率与感受野。为了准确捕捉RSI中大型对象的信息，模型输出必须对足够大的区域作出响应，Luo等[2016]。然而，这一需求显著增加了计算复杂性，Christophe等[2011]，Ma等[2015]。基于CNN的网络由于依赖于固定大小卷积核的滑动计算缺乏全局感受野。虽然基于注意力的模型能够实现全局建模，但其注意力机制带来了二次计算复杂性。因此，现有的RSFMs在实际应用中难以同时提供快速且高精度的推理。

弱物理可解释性。RS对象经常表现出不规则的多边形形状Li和Narayanan [2003]，当前的RSFMs难以整合物理原理来解释对象特征如何传播。这一缺陷使得研究人员难以有效地调整学习策略Shen等 [2022]、Temenos 等 [2023]、Pérez-Suay等 [2020]。从长远来看，RSFMs需要具备一定的信息可解释性。

为了应对这些挑战，本文引入了RS-vHeat，这是一种受vHeat Wang等人的[2024a]启发、支持多模态输入的基于热传导的RSFM。首先，热传导代表了能量从高温区域向低温区域扩散的自然过程，如图1中的概念物理模型所示。这一过程基于材料的不同，从非稳态过渡到稳态。由于其计算过程类似于神经网络中的特征提取，因此可以应用于RS图像处理。其次，我们假设对象类型对应于特殊的特征分布，模型根据RS特定属性预测扩散率，通过热流模拟参数计算。这种方法将所有模态投影到一个共同的热空间中，遵循复杂区域包含RS对象时为高温区域，热量积聚，而稀疏区域为低温区域，热量容易扩散的约束，如图1下半部分所示。第三，基于这一理论，我们进一步设计了一种具有物理可解释性的RSFM，使用300万张光学和SAR数据进行预训练，如图2所示。热传导网络模拟了热在大规模多模态RS数据中的扩散过程，有助于特征传播与对象结构特征的对齐。此外，其计算方法为网络的高效运行提供了指导。

To summarize, our contributions are as follows:

1. We introduce RS-vHeat, a RSFM designed based on the heat conduction differential equation to process RS data. It conceptualizes the semantic relationships between pixels in RS images as the propagation of heat.
2. We propose a self-supervised strategy based on frequency domain hierarchical masking and multi-domain reconstruction that preserves small objects, driving the model to reconstruct fine and coarse frequency signals.
3. We design spatial correction embeddings, which operate directly on the global features to capture local details, assisting in simulating the rate of thermal diffusion.
4. We evaluate RS-vHeat on 10 datasets, showing it outperforms advanced RSFMs in accuracy while maintaining lower computational complexity. When processing large-scale images, RS-vHeat reduces memory usage by 84%, decreases FLOPs by 24%, and improves throughput by 2.7 times compared to attention-based RSFMs.

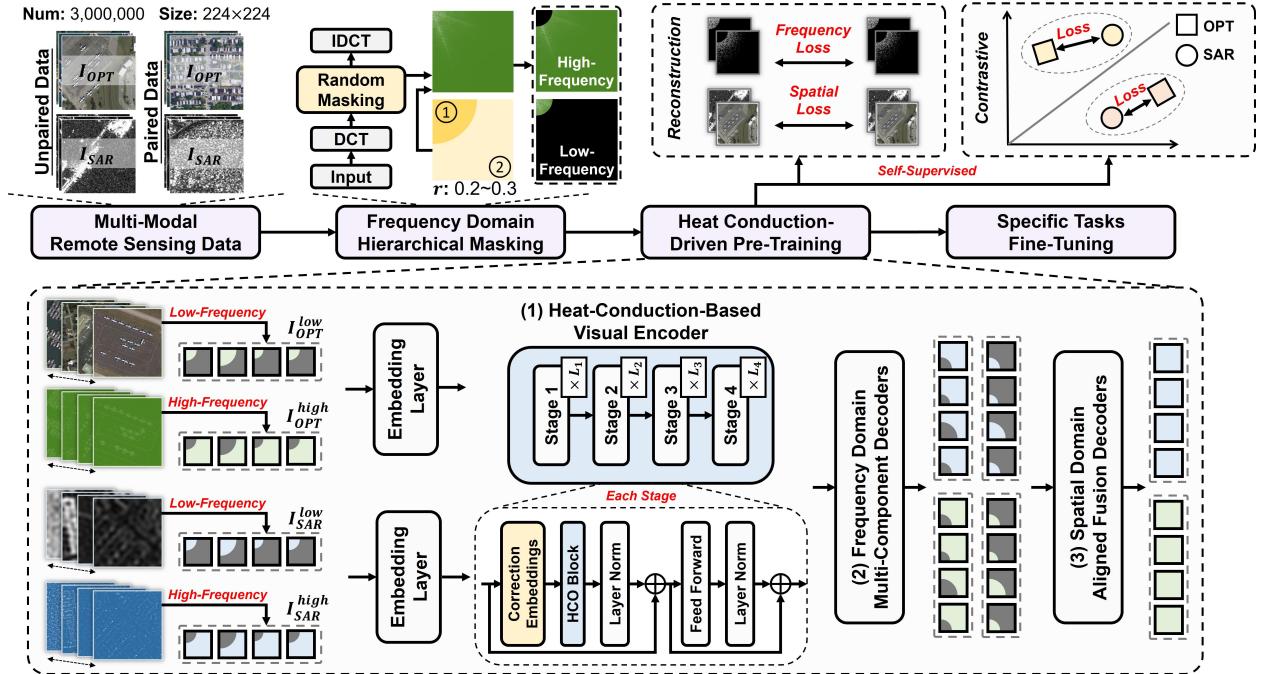


Figure 2: The pre-training process of RS-vHeat. It performs frequency domain hierarchical masking based on randomly generated sector regions to separate each image into high- and low-frequency components. These component images are fed into the network and projected into the thermal space via embedding layers. The heat diffusion is computed within the heat-conduction-based visual encoder to simulate complex RS objects. The deep features undergo multi-domain reconstruction loss and contrastive loss computation via decoders.

2 Related Work

2.1 Mainstream Network Architectures in RS

Convolutional Neural Networks. CNNs have been adapted for RS to tackle challenges in satellite imagery Zhang et al. [2019], Li et al. [2020b]. However, local receptive fields, mainly limited by the kernel size, restrict the capture of long-range dependencies Dosovitskiy et al. [2020]. This drawback is especially critical in RS, where large-scale patterns and complex spatial relationships are prevalent, making the balance between local feature extraction and global context a key research focus Dong et al. [2021], Chen et al. [2020].

Transformers. Self-attention mechanisms Vaswani et al. [2017] empower networks to capture long-range dependencies Dosovitskiy et al. [2020], Liu et al. [2021]. Research studies have shown that plain ViTs performs better than traditional CNN models in RS applications Wang et al. [2022a], Yan et al. [2022]. However, the limitation stems from the ViTs' restricted ability to effectively handle long sequences. When processing large-scale RS images, the computational load of networks grows quadratically, leading to significant overhead.

总结来说，我们的贡献如下：

1. 我们引入了基于热传导微分方程设计的RS-vHeat，这是一种用于处理RS数据的RSFM。它将RS图像中像素之间的语义关系概念化为热量的传播。
2. 我们提出了一种基于频域分层掩蔽和多域重建的自监督策略，该策略保留了小对象，促使模型重建细粒度和粗粒度的频率信号。
3. 我们设计了空间校正嵌入，这些嵌入直接作用于全局特征以捕获局部细节，有助于模拟热扩散速率。
4. 我们在10个数据集上评估了RS-vHeat，结果显示它在准确率上优于先进的RSFM，同时保持较低的计算复杂度。在处理大规模图像时，RS-vHeat将内存使用量减少了84%，减少了24%的FLOPs，并将吞吐量提高了2.7倍，相比基于注意力的RSFM。

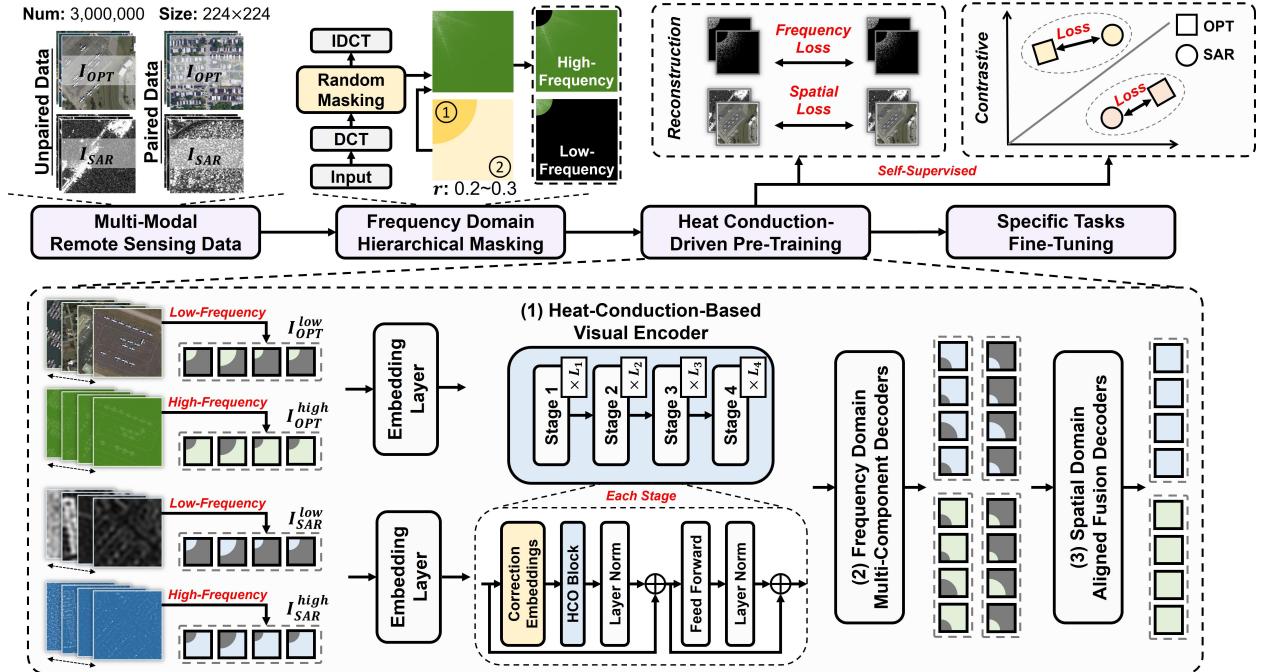


Figure 2: RS-vHeat 的预训练过程。它基于随机生成的扇区区域进行频域分层掩码，将每张图像分离成高频和低频分量。这些分量图像被输入网络并通过嵌入层投影到热空间中。在基于热传导的视觉编码器内计算热扩散，以模拟复杂的RS对象。深层特征通过解码器进行多域重建损失和对比损失计算。

2 相关工作

2.1 RS 中的主要网络架构

卷积神经网络。CNNs 已被adapted 用于RS 以应对卫星影像中的挑战 Zhang 等人 [2019], Li 等人 [2020b]。然而，局部感受野，主要受限于核大小，限制了对长距离依赖性的捕获 Dosovitskiy 等人 [2020]。这一缺点在RS中尤为关键，因为大规模模式和复杂的空间关系普遍存在，使得局部特征提取与全局上下文之间的平衡成为研究重点 Dong 等人 [2021], Chen 等人 [2020]。

变换器。自我注意机制 Vaswani 等人 [2017] 使网络能够捕获长范围依赖 Dosovitskiy 等人 [2020]、Liu 等人 [2021]。研究发现，在 RS 应用中，简单的 ViTs 模型比传统的 CNN 模型表现更好 Wang 等人 [2022a]、Yan 等人 [2022]。然而，这一局限来自于 ViTs 有限的能力，无法有效处理长序列。在处理大规模 RS 图像时，网络的计算负载会呈二次增长，导致显著的开销。

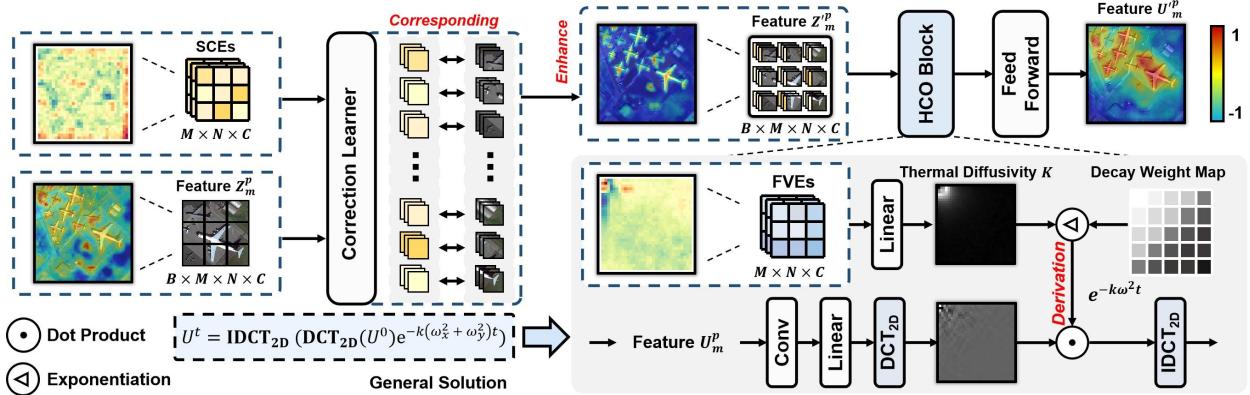


Figure 3: The overall structure of the heat-conduction-based visual encoder, simulates the general solution of heat conduction. SCEs first dynamically perform spatial domain correction, adjusting heat distribution based on the RS scene. The enhanced image undergoes a 2D DCT in the HCO block, interacting with the heat diffusion rate predicted by the FVEs to complete a heat conduction calculation, which is then transformed back into a visual representation via 2D IDCT.

Mamba-Based Models. Mamba, an efficient implementation of State Space Models (SSMs) Gu et al. [2021], leverages selective scan mechanisms for long-sequence processing Liu et al. [2024a], Zhu et al. [2024a], balancing computational efficiency and high accuracy. Gu and Dao [2023], Xu et al. [2024]. It has been rapidly adopted in RS Chen et al. [2024a], Zhu et al. [2024b], Chen et al. [2024b], with its ability to process long sequences, although Mamba, as an innovative architecture, suffers from limited interpretability.

2.2 Self-Supervised Learning Strategies for RSFMs

Contrastive Learning. By establishing rules to distinguish between positive and negative samples, contrastive learning aims to fully understand the relationships between these samples. GASSL Ayush et al. [2021] treats RS images of the same scene captured at different times as positive pairs, using them as self-supervised signals. Based on seasonal contrast method, SeCo Manas et al. [2021] and CACo Mall et al. [2023] effectively leverage temporal information within the networks by comparing scene variations across different years or seasons as perceptual signals. Skysense Guo et al. [2024] utilizes multi-modal RS images as input and implements a multi-granularity contrastive learning framework. Motivated by the aforementioned methods, we apply contrastive constraints in the thermal space to optical and SAR data, encouraging the model to focus on deep, fine-grained semantic relationships within the RS images.

Masked Image Modeling. By involving masking parts of the image and predicting the missing information, networks can understand the details of RSI. RingMo Sun et al. [2022b] attempts to solve the issue that directly masking image patches can easily lead to the loss of small objects by designing an incomplete masking strategy that is implemented proportionally within the patches. SpectralGPT Hong et al. [2024] models multi-spectral data as three-dimensional data and applies masking in three-dimensional space. Scale-MAE Reed et al. [2023] masks pixels in the spatial domain and reconstructs high- and low-frequency images to learn image representations at different scales. Inspired by these, to preserve and learn the information of complex RS scenes, we employ dual reconstruction in both the spatial and frequency domains as constraints.

3 Proposed RS-vHeat

We introduce RS-vHeat, a multi-modal remote sensing foundation model with three key components: a frequency domain hierarchical masking strategy for multi-modal RS data, a visual encoder for modeling internal heat flow within RS images, and decoders for multi-domain reconstruction. fig. 2 illustrates the RS-vHeat structure during pre-training, enabling shared representation of multi-modal data through frequency domain masking and projection into a thermal space to simulate heat propagation. Preliminary details on vHeat Wang et al. [2024a] are included in the supplementary materials.

Frequency Domain Hierarchical Masking Strategy. Traditional spatial masking Xie et al. [2022] on RS images often obscures smaller objects, hindering reconstruction. Inspired by the frequency-aware dynamic network Xie et al.

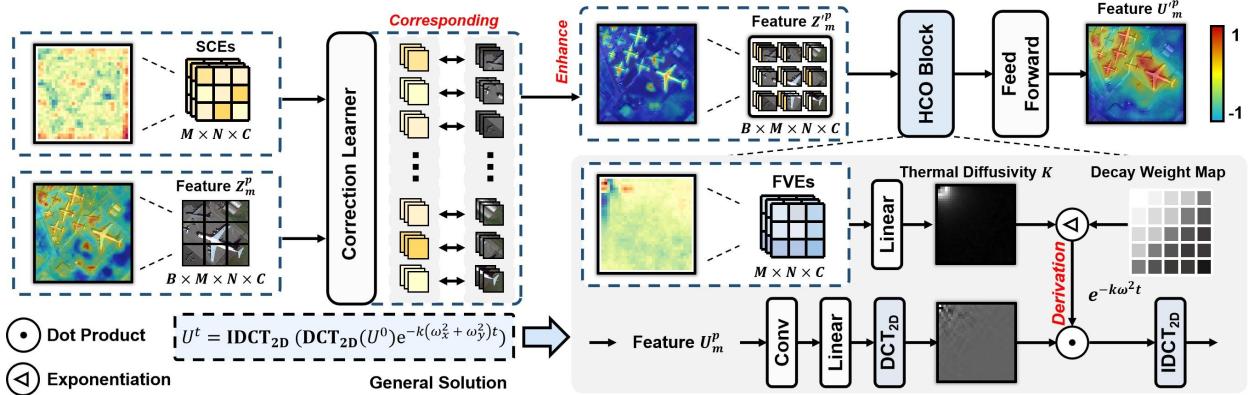


图3：基于热传导的视觉编码器的整体结构模拟了热传导的一般解。SCEs 首先动态地在空间域内进行校正，根据RS场景调整热分布。增强后的图像在HCO块中进行2D DCT变换，与FVEs预测的热扩散速率相互作用以完成热传导计算，然后通过2D IDCT变换重新转换为视觉表示。

Mamba 基础模型。Mamba 是 Gu 等人 [2021] 对 State Space Models (SSMs) 的高效实现，利用选择性扫描机制进行长序列处理，Liu 等人 [2024a]，Zhu 等人 [2024a]，在保持计算效率和高精度之间取得平衡。Gu 和 Dao [2023]，Xu 等人 [2024]。Mamba 由于能够处理长序列，已被迅速采纳到推荐系统中，Chen 等人 [2024a]，Zh u 等人 [2024b]，Chen 等人 [2024b]，尽管作为一项创新架构，Mamba 在可解释性方面存在局限。

2.2 自监督学习策略 for RSFMs

对比学习。通过建立规则来区分正样本和负样本，对比学习旨在全面理解这些样本之间的关系。GASSL A yush 等人 [2021] 将同一场景在不同时间拍摄的 RS 图像视为正样本对，并使用它们作为自我监督信号。基于季节对比方法，SeCo Manas 等人 [2021] 和 CACo Mall 等人 [2023] 通过比较不同年份或季节的场景变化作为感知信号，有效地在网络中利用时间信息。Skysense Guo 等人 [2024] 利用多模态 RS 图像作为输入，并实现了一个多粒度对比学习框架。受上述方法的启发，我们在热红外空间中应用对比约束，鼓励模型关注 RS 图像中的深层次、细粒度语义关系。

Masked Image Modeling. 通过遮掩图像的一部分并预测缺失的信息，网络可以理解 RSI 的细节。RingMo Sun 等 [2022b] 试图通过设计一种在 patches 内部按比例实现的不完全遮掩策略来解决直接遮掩图像 patches 容易导致小对象丢失的问题。SpectralGPT Hong 等 [2024] 将多光谱数据建模为三维数据，并在三维空间中应用遮掩。Scale-MAE Reed 等 [2023] 在空间域中遮掩像素并重建高频和低频图像以在不同尺度上学习图像表示。受这些方法的启发，为了保留和学习复杂 RS 场景的信息，我们在空间域和频率域中同时采用双重重建作为约束。

3 提出的 RS-vHeat

我们介绍 RS-vHeat，这是一种多模态遥感基础模型，包含三个关键组件：一种多模态遥感数据的频域分层掩码策略、一种视觉编码器用于建模遥感图像内的热流，以及用于多域重建的解码器。图 2 展示了预训练期间 RS-vHeat 的结构，通过频域掩码和投影到热空间来实现多模态数据的共享表示，模拟热传播。关于 vHeat 的初步细节参见 Wang 等人 [2024a] 的补充材料。

频域分层掩码策略。传统的基于空间的掩码 Xie 等人 [2022] 在 RS 图像中往往会使较小的对象被遮挡，妨碍重建。受频率意识动态网络 Xie 等人的启发。

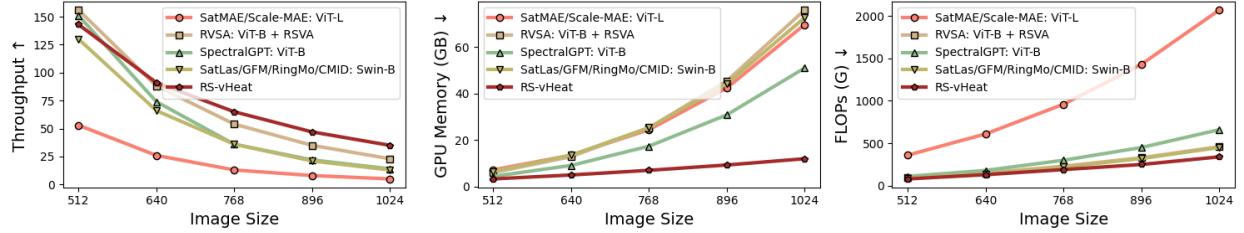


Figure 4: Comparison of throughput (left), GPU memory (middle), and FLOPs (right) across image sizes for RSFMs. RS-vHeat shows significantly higher throughput, lower memory usage and lower FLOPs than current attention-based RSFMs, especially for large-scale images. All tests were performed in a consistent environment using a single A100 (80G) GPU with a batch size of 32.

[2021a], we implement a frequency domain hierarchical masking strategy to preserve object structure, enabling more accurate heat propagation by focusing on frequency domain signals.

Specifically, given the input (optical and SAR), denoted as $I(x, y, c) = \{I_o, I_s\}, I_o \in \mathbb{R}^{H \times W \times 3}, I_s \in \mathbb{R}^{H \times W \times 1}$, two parallel streams process paired and unpaired data. The Discrete Cosine Transform (DCT) is applied along each image dimension, converting $I(x, y)$ to its frequency representation $\tilde{I}(u, v)$, where low-frequency components are concentrated in the top-left of the spectrum. A sector mask separates high-frequency $\tilde{I}^{high}(u, v)$ and low-frequency $\tilde{I}^{low}(u, v)$ regions, using a random masking rate of 20%-30%. Next, the Inverse Discrete Cosine Transform (IDCT) is applied along the image dimensions to revert the data back to the spatial domain, obtaining $I^{low}(x, y)$ and $I^{high}(x, y)$. Notably, the DCT and IDCT operations are differentiable, enabling efficient computation and seamless integration into network training on both CPUs and GPUs. Details can be found in the supplementary material.

Heat-Conduction-Based Visual Encoder. High- and low-frequency information from multi-modal data is mapped to a shared thermal space and then fed into the heat-conduction-based visual encoder for thermal simulation. In most RS downstream tasks, areas containing RS objects tend to appear as high-temperature zones, while sparse or empty regions display as low-temperature zones.

The Heat Conduction Operator (HCO) simulates the process of visual information transmission as thermal conduction. The two-dimensional temperature distribution at time t , $u(x, y, t)$ is extended to a multi-dimensional feature distribution $U(x, y, c, t)$. The HCO block specifically models the general solution of physical heat conduction:

$$U_m^t = \mathcal{F}^{-1} \left(\mathcal{F}(U_m^0) e^{-k(\omega_x^2 + \omega_y^2)t} \right) \quad (1)$$

where U^0 and U^t represents the input $U(x, y, c, 0)$ and output $U(x, y, c, t)$. We denote the Discrete Fourier Transform (DFT) and its inverse (IDFT) as \mathcal{F} and \mathcal{F}^{-1} . $m \in \{o, s\}$ refers to the modality (optical or SAR). Based on the Neumann boundary condition Cheng and Cheng [2005], we replace the 2D DFT and IDFT with the 2D DCT (DCT_{2D}) and IDCT (IDCT_{2D}) Strang [1999]:

$$U_m^t = IDCT_{2D} \left(DCT_{2D}(U_m^0) e^{-k(\omega_x^2 + \omega_y^2)t} \right) \quad (2)$$

where $e^{-k(\omega_x^2 + \omega_y^2)t}$ functions as an adaptive filter in the frequency domain, executing heat conduction.

A set of learnable frequency value embeddings (FVEs) is predicted to estimate the heat diffusion coefficient k ($k := k(\omega_x, \omega_y)$), facilitating adaptive heat transfer, as shown in fig. 3. The weights are denoted as $W_{FVEs} \in \mathbb{R}^{M \times N \times C}$:

$$U_m^p = HCO(U_m^p, W_{FVEs}) \quad (3)$$

where U and U' represent the temperature states of the features from different modalities before and after passing through the HCO block, respectively, with $p \in \{\text{high}, \text{low}\}$ indicating the frequency level.

To enhance the spatial representation of different structural components, we introduce a lightweight correction learner. By predicting a set of spatial correction embeddings (SCEs) from large-scale pre-trained images, it interacts with the existing temperature field and perform activation, as shown in fig. 3. The weight $W_{SCEs} \in \mathbb{R}^{M \times N \times C}$ adaptively adjusts the object boundaries in the spatial domain based on content by adding them to the original image $I(x, y)$. Given the upper-layer temperature feature Z_m^p , the correction learner can be expressed as CL :

$$Z_m'^p = CL(Z_m^p, W_{SCEs}) \quad (4)$$

Where $Z_m'^p$ represents the temperature features after adaptive adjustment.

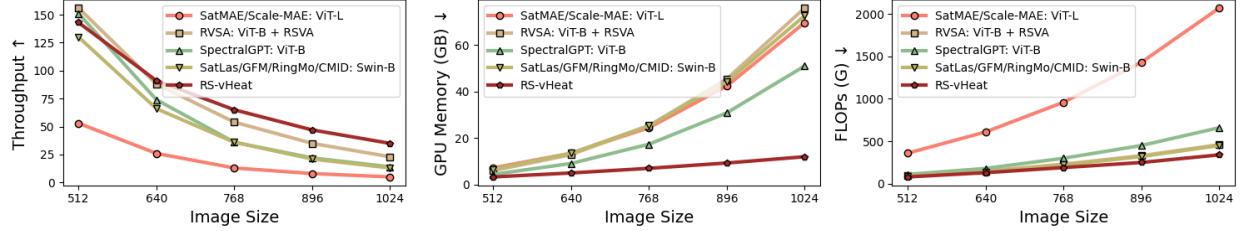


Figure 4: RSFsMs 不同图像大小下的吞吐量（左）、GPU 内存（中）和 FLOPs（右）比较。RS-vHeat 的吞吐量显著更高，内存使用量更低，FLOPs 更低，尤其是在大规模图像方面，与当前基于注意力的 RSFsMs 相比更为明显。所有测试均在一致的环境中使用单个 A100 (80G) GPU 和批量大小为 32 的条件下进行。

[2021a]，我们实现了一种频域分层掩蔽策略以保留对象结构，从而通过专注于频域信号来实现更准确的热量传播。

具体来说，给定输入（光学和SAR），表示为 $I(x, y, c) = \{I_o, I_s\}, I_o \in \mathbb{R}^{H \times W \times 3}, I_s \in \mathbb{R}^{H \times W \times 1}$ ，两个并行流处理配对和未配对的数据。离散余弦变换 (DCT) 沿每个图像维度应用，将 $I(x, y)$ 转换为其频率表示 $\tilde{I}(u, v)$ ，其中低频分量集中在频谱的左上角。扇区掩码将高频 $\tilde{I}^{high}(u, v)$ 和低频 $\tilde{I}^{low}(u, v)$ 区域分开，使用 20%-30% 的随机遮罩率。接下来，沿图像维度应用逆离散余弦变换 (IDCT) 将数据重新转换回空间域，获得 $I^{low}(x, y)$ 和 $I^{high}(x, y)$ 。值得注意的是，DCT 和 IDCT 操作是可微的，这使得高效的计算和在 CPU 和 GPU 上无缝集成到网络训练成为可能。更多细节可以在补充材料中找到。

基于热传导的视觉编码器。多模态数据中的高频和低频信息被映射到共享的热空间，然后输入基于热传导的视觉编码器进行热模拟。在大多数 RS 下游任务中，包含 RS 对象的区域往往表现为高温区，而稀疏或空旷的区域则表现为低温区。

The Heat Conduction Operator (HCO) 模拟视觉信息传输过程中的热传导。时间 t 时的二维温度分布 $u(x, y, t)$ 扩展到多维特征分布 $U(x, y, c, t)$ 。HCO 块具体模型物理热传导的一般解：

$$U_m^t = \mathcal{F}^{-1} \left(\mathcal{F}(U_m^0) e^{-k(\omega_x^2 + \omega_y^2)t} \right) \quad (1)$$

在哪里， U^0 和 U^t 表示输入 $U(x, y, c, 0)$ 和输出 $U(x, y, c, t)$ 。我们用离散傅里叶变换 (DFT) 及其逆变换 (IDFT) 表示为 \mathcal{F} 和 \mathcal{F}^{-1} 。 $m \in \{o, s\}$ 指的是模态（光学或SAR）。基于 Cheng 和 Cheng [2005] 的 Neumann 边界条件，我们将 2D DFT 和 IDFT 替换为 2D DCT (DCT_{2D}) 和 IDCT ($IDCT_{2D}$) Strang [1999]：

$$U_m^t = IDCT_{2D} \left(DCT_{2D}(U_m^0) e^{-k(\omega_x^2 + \omega_y^2)t} \right) \quad (2)$$

其中 $e^{-k(\omega_x^2 + \omega_y^2)t}$ 作为频域中的自适应滤波器，执行热传导。

一组可学习的频率值嵌入 (FVEs) 被预测用于估计热扩散系数 k ($k := k(\omega_x, \omega_y)$)，促进自适应热传递，如图3所示。权重表示为 $W_{FVEs} \in \mathbb{R}^{M \times N \times C}$ ：

$$U_m'^p = HCO(U_m^p, W_{FVEs}) \quad (3)$$

其中 U 和 U' 分别表示在经过 HCO 块之前和之后的不同模态特征的温度状态， $p \in \{high, low\}$ 表示频率级别。

为了增强不同结构组件的空间表示，我们引入了一个轻量级的校正学习器。通过从大规模预训练图像中预测一组空间校正嵌入 (SCEs)，它与现有的温度场进行交互并执行激活操作，如图3所示。权重 $W_{SCEs} \in \mathbb{R}^{M \times N \times C}$ 根据内容在空间域中自适应地调整对象边界，并将其添加到原始图像 $I(x, y)$ 中。给定高层温度特征 Z_m^p ，校正学习器可以表示为 CL ：

$$Z_m'^p = CL(Z_m^p, W_{SCEs}) \quad (4)$$

Where $Z_m'^p$ 表示经过自适应调整后的温度特征。

Multi-Domain Reconstruction Decoders. Considering that the restoration of heat distribution is modality-specific, the optical and SAR decoders (fig. 2 (2)), D_o^δ and D_s^δ concurrently operate on the encoded outputs ($F_o'^p, F_s'^p \in \mathbb{R}^{M \times N \times C}$). By employing convolution and pixel shuffling, decoders effectively upsample feature maps. The reconstruction outputs I' for both modalities— I_o^{high}, I_o^{low} for optical, and I_s^{high}, I_s^{low} for SAR—are as follows:

$$I_o'^p = D_o^\delta(F_o'^p), I_s'^p = D_s^\delta(F_s'^p) \quad (5)$$

The reconstruction is guided by the \mathcal{L}_1 loss in the frequency domain. After applying DCT, the loss \mathcal{L}_{Fre} is computed by measuring the difference between transformed features $\tilde{I}_m^p(u, v)$ and $\tilde{I}_m^p(u, v)$.

To learn more robust higher-level features, spatial decoders (fig. 2 (3)) are responsible for fusion and reconstruction. Specifically, outputs, $F_m'^p$ and F_m^p , from the third and fourth stages extract modality-specific frequency features. The fusion layer integrates the frequency components into higher-level image features, F'_m and F_m :

$$F'_m = \text{Concat}[\text{CONV}(F_m^{high}), \text{CONV}(F_m^{low})], F_m = \text{Concat}[\text{CONV}(F_m^{high}), \text{CONV}(F_m^{low})] \quad (6)$$

Furthermore, the integration of multi-stage features is achieved through the function $P_m = g(F_m, F'_m)$, where g combines convolution, ReLU activation and element-wise summation operations to collectively produce the final high-level output $P_m = \{P_o, P_s\}$. Subsequently, these features are fed into modality-specific spatial domain decoders D_o^ϕ and D_s^ϕ , they are upsampled to the original image size to restore the high-level semantics of the original image:

$$I'_o = D_o^\phi(P_o), I'_s = D_s^\phi(P_s) \quad (7)$$

Where I'_o and I'_s represent the reconstructed high-level optical and SAR data. The reconstruction loss in the spatial domain \mathcal{L}_{Spa} is then computed using the \mathcal{L}_1 loss.

To explore the network's understanding of fine-grained semantic information, we employ contrastive loss to compute the semantic discrepancy between different modalities. The loss operates on embeddings F_m^p obtained from two different preprocessing methods applied to the same image, representing low- and high-frequency features. The contrastive loss \mathcal{L}_{Con} is calculated leveraging cosine similarity, following the computational methodology outlined in Atito et al. [2021].

The overall loss function is shown in eq. (8). The combined approach encourages the model to fully capture both global structures and fine-grained details:

$$\mathcal{L}_{total} = \mathcal{L}_{Con} + \mathcal{L}_{Spa} + \mathcal{L}_{Fre} \quad (8)$$

4 Experiments

4.1 Training Implementation

The training of RS-vHeat consists of self-supervised pre-training and downstream fine-tuning. The visual encoder follows the Swin-B Liu et al. [2021] configuration with four stages of 2, 2, 18, and 2 blocks. During pre-training, we use a large-scale multi-modal dataset with 450k matched optical and SAR image pairs, totaling over 3 million entries, following the methodology of RingMo Sun et al. [2022b]. The model is trained on eight A100 (80G) GPUs for 200 epochs with images of size 224. The training starts with a learning rate of 1e-6 for 10 warm-up epochs, gradually increasing to 2e-4, with a cosine annealing schedule and a minimum learning rate of 1e-5. For fine-tuning, we transfer the pre-trained embedding layers and visual encoder structure, adjusting the fixed-size FVEs and SCEs to match the image dimensions.

4.2 Performance

We comprehensively evaluate the performance of RS-vHeat compared to other representative RSFMs. Primarily, as shown in fig. 4, we analyze throughput, memory usage, and FLOPs across various image sizes. RS-vHeat (depicted by the red line) demonstrates significantly higher throughput, lower memory consumption and FLOPs, with these advantages becoming more pronounced as image sizes increase. For instance, when processing 1024×1024 resolution images, RS-vHeat achieves 2.7 times the throughput of Swin-B-based models such as SatLas Bastani et al. [2023], GFM Mendieta et al. [2023], RingMo Sun et al. [2022b], and CMID Muhtar et al. [2023], while reducing memory usage by 84%. Similarly, when compared to the improved ViT-B-based SpectralGPT Hong et al. [2024], RS-vHeat achieves 2.5 times the throughput while reducing memory usage by 77%.

Next, we evaluate the accuracy and computational efficiency of fine-tuning RS-vHeat on four tasks—semantic segmentation, object detection, classification and change detection—to demonstrate the effectiveness, along with extensive

多域重建解码器。由于热分布的恢复具有模态特异性，光学和SAR解码器（图2（2））， D_o^δ 和 D_s^δ 同时作用于编码输出 ($F_o'^p$, $F_s'^p \in \mathbb{R}^{M \times N \times C}$)。通过使用卷积和像素混洗，解码器有效地进行上采样。两种模态的重建输出 I' 如下所示—— $I_o'^{high}$ 为光学， $I_o'^{low}$ 为SAR， $I_s'^{high}$ 和 $I_s'^{low}$ 为SAR：

$$I_o'^p = D_o^\delta(F_o'^p), I_s'^p = D_s^\delta(F_s'^p) \quad (5)$$

重建由频域中的 \mathcal{L}_1 损失引导。应用DCT后，通过测量变换特征 $\tilde{I}_m'^p(u, v)$ 和 $\tilde{I}_m^p(u, v)$ 之间的差异来计算损失 \mathcal{L}_{Fre} 。

为了学习更健壮的高级特征，空间解码器（图2（3））负责融合和重构。具体来说，第三和第四阶段的输出 $F_m'^p$ 和 F_m^p 提取了模态特定的频率特征。融合层将这些频率成分整合到更高层次的图像特征 F'_m 和 F_m 中：

$$F'_m = \text{Concat}[\text{CONV}(F_m'^{high}), \text{CONV}(F_m'^{low})], F_m = \text{Concat}[\text{CONV}(F_m^{high}), \text{CONV}(F_m^{low})] \quad (6)$$

此外，多阶段特征的集成通过函数 $P_m = g(F_m, F'_m)$ 实现，其中 g 结合了卷积、ReLU激活和元素级求和操作，共同生成最终的高层输出 $P_m = \{P_o, P_s\}$ 。随后，这些特征被输入到模态特定的空间域解码器 D_o^ϕ 和 D_s^ϕ 中，并被上采样到原始图像大小以恢复原始图像的高层语义：

$$I'_o = D_o^\phi(P_o), I'_s = D_s^\phi(P_s) \quad (7)$$

Where I'_o 和 I'_s 表示重构的高光谱和SAR数据。然后在空间域中使用 \mathcal{L}_1 损失计算重构损失 \mathcal{L}_{Spa} 。

为了探索网络对细粒度语义信息的理解，我们采用对比损失来计算不同模态之间的语义差异。该损失作用于同一图像应用两种不同预处理方法后得到的嵌入 F_m^p ，分别表示低频和高频特征。对比损失 \mathcal{L}_{Con} 利用余弦相似度进行计算，遵循Atito等[2021]提出的计算方法。

整体损失函数如式(8)所示。结合方法鼓励模型全面捕捉全局结构和细粒度细节：

$$\mathcal{L}_{total} = \mathcal{L}_{Con} + \mathcal{L}_{Spa} + \mathcal{L}_{Fre} \quad (8)$$

4 实验

4.1 训练实施

RS-vHeat 的训练包括自我监督的预训练和下游微调。视觉编码器遵循 Liu 等人 [2021] 提出的 Swin-B 配置，包含四个阶段，分别为 2、2、18 和 2 个块。在预训练过程中，我们使用包含 450 万对匹配的光学和 SAR 图像的大型多模态数据集，总计超过 300 万条记录，遵循 Sun 等人 [2022b] 的方法。该模型在八块 A100 (80G) GPU 上训练 200 个周期，图像大小为 224。训练从学习率 1e-6 开始，持续 10 个预热周期，然后逐渐增加到 2e-4，采用余弦退火计划，最小学习率为 1e-5。在微调过程中，我们转移预训练的嵌入层和视觉编码器结构，调整固定大小的 FVEs 和 SCEs 以匹配图像尺寸。

4.2 性能

我们全面评估了 RS-vHeat 与其他代表性RSFMs 的性能。如图4所示，我们分析了不同图像大小下的吞吐量、内存使用和FLOPs。RS-vHeat（用红色线条表示）显示出显著更高的吞吐量、更低的内存消耗和FLOPs，随着图像大小的增加，这些优势更加明显。例如，处理 1024×1024 分辨率的图像时，RS-vHeat 的吞吐量是基于 Swin-B 的模型 SatLas Bastani et al. [2023]、GFM Mendieta et al. [2023]、RingMo Sun et al. [2022b] 和 CMID Muhtar et al. [2023] 的 2.7 倍，同时内存使用减少了 84%。同样，与改进的ViT-B基SpectralGPT Hong et al. [2024]相比，RS-vHeat 的吞吐量是其 2.5 倍，内存使用减少了 77%。

接下来，我们评估在语义分割、对象检测、分类和变化检测四项任务上对 RS-vHeat 进行微调的准确性和计算效率，以展示其有效性，并结合广泛的

ablation studies. Many RSFMs primarily focus on optical data and do not support SAR input, so we compare RS-vHeat with other RSFMs using optical datasets, while SAR-specific tasks are compared against specialized models. RSFMs are categorized into three main groups based on the backbone: CNN-based, ViT-based, and Swin-based. All datasets follow official partitioning methods for training and testing. More details on datasets and visualizations are provided in the supplementary materials.

Single- and Multi-Modal Semantic Segmentation. We evaluate our model on two optical datasets (Potsdam Sherrah [2016], iSAID Waqas Zamir et al. [2019]), one SAR dataset (Air-PoSAR-Seg Wang et al. [2022b]), and one multi-modal dataset (Li et al. [2022]), using UPerNet Xiao et al. [2018] with cross-entropy loss for the output head. table 1 compares RS-vHeat with 14 other RSFMs on iSAID and Potsdam. RS-vHeat demonstrates superior accuracy compared to CNN-based models, and achieves lower FLOPs than ViT-based models, exemplified by its 1.28% and 2.95% improvement over Scale-MAE Reed et al. [2023]. It also exhibits the lowest FLOPs among Swin-based models, with over a fourfold reduction in parameter count compared to SkySense Guo et al. [2024], with only 1.17% and 2.19% decreases in accuracy. Besides, on AIR-PoSAR-Seg (table 3) and WHU-OPT-SAR (table 2), RS-vHeat outperforms other specialized segmentation models, demonstrating the applicability of the heat conduction model in SAR and multi-modal segmentation tasks.

Table 1: Comparison of mF1, mIoU, parameters, and FLOPs on iSAID and Potsdam with other RSFMs.

Backbone Type	Method	Backbone	Potsdam mF1 ↑	iSAID mIoU ↑	Image Size	Params	FLOPs ↓
CNN-Based	GASSL (ICCV'2021) Ayush et al. [2021]	ResNet-50	91.27	65.95	896 ²	64M	722G
	SeCo (ICCV'2021) Manas et al. [2021]	ResNet-50	89.03	57.20	896 ²	64M	722G
	SSL4EO (GRSM'2023) Wang et al. [2023]	ResNet-50	91.54	64.01	896 ²	64M	722G
	CACo (CVPR'2023) Mall et al. [2023]	ResNet-18	91.35	64.32	896 ²	41M	671G
	TOV (JSTARS'2023) Tao et al. [2023]	ResNet-50	92.03	66.24	896 ²	64M	722G
	SAMRS (NeurIPS'2023) Wang et al. [2024b]	ResNet-50	91.43	66.26	896 ²	64M	722G
	RS-vHeat (Ours)	vHeat-B + SCEs	92.82	68.72	896 ²	148M	921G
ViT-Based	RVSA (TGRS'2022) Wang et al. [2022a]	ViT-B + RVSA	-	64.49	896 ²	128M	1043G
	SatMAE (NeurIPS'2022) Cong et al. [2022]	ViT-L	90.63	62.97	896 ²	341M	1536G
	Scale-MAE (ICCV'2023) Reed et al. [2023]	ViT-L	91.54	65.77	896 ²	341M	1536G
	RS-vHeat (Ours)	vHeat-B + SCEs	92.82	68.72	896 ²	148M	921G
Swin-Based	RingMo (TGRS'2022) Sun et al. [2022b]	Swin-B	91.27	67.20	896 ²	121M	968G
	SatLas (ICCV'2023) Bastani et al. [2023]	Swin-B	91.28	68.71	896 ²	121M	968G
	GFM (ICCV'2023) Mendieta et al. [2023]	Swin-B	91.85	66.62	896 ²	121M	968G
	CMID (TGRS'2023) Muhtar et al. [2023]	Swin-B	91.86	66.21	896 ²	121M	968G
	SkySense (CVPR'2024) Guo et al. [2024]	Swin-H	93.99	70.91	896 ²	>702M	>2708G
	RS-vHeat (Ours)	vHeat-B + SCEs	92.82	68.72	896 ²	148M	921G

Table 2: Comparison of OA and user's accuracy on WHU-OPT-SAR with other specialized models.

Method	Publication	User's Accuracy ↑							OA ↑
		Farmland	City	Village	Water	Forest	Road	Others	
SegFormer Xie et al. [2021b]	NeurIPS'2021	79.1	72.9	38.0	64.7	88.1	0.3	0.4	75.5
Segmenter Strudel et al. [2021]	ICCV'2021	82.3	75.2	51.7	74.4	89.4	16.0	12.0	79.9
MCANet Li et al. [2022]	JAG'2022	74.3	62.2	53.1	65.7	95.5	31.0	9.8	82.9
VMamba Liu et al. [2024a]	NeurIPS'2024	82.1	74.1	57.8	81.4	89.1	41.5	18.4	81.4
MMOKD Liu et al. [2024b]	TGRS'2024	70.0	58.1	50.0	69.7	80.1	40.1	25.0	82.5
RS-vHeat (Ours)	-	81.1	67.3	67.5	79.0	90.2	54.9	55.3	83.9

Table 3: Comparison of mIoU, OA, and AA on AIR-PoSAR-Seg with other specialized models.

Method	Publication	mIoU ↑	OA ↑	AA ↑
DeepLab V3+ Chen et al. [2018]	ECCV'2018	48.21	76.81	63.55
EncNet Zhang et al. [2018]	CVPR'2018	47.75	75.67	57.51
PSANet Zhao et al. [2018]	ECCV'2018	47.14	76.21	62.92
CCNet Huang et al. [2019]	ICCV'2019	46.46	75.53	55.83
DANet Fu et al. [2019]	CVPR'2019	51.93	76.91	62.79
GCNet Cao et al. [2019]	ICCV'2019	47.56	76.75	57.10
RS-vHeat (Ours)	-	57.46	81.46	65.92

Object Detection. We conduct coarse- and fine-grained experiments on two optical datasets (FAIR1M Sun et al. [2022a], DIOR Li et al. [2020a]) and one SAR dataset (SAR-AIRCRAFT-1.0 Zhirui et al. [2023]), using YOLOX Ge et al. [2021] as the output head. On the DIOR dataset, table 4 shows that RS-vHeat demonstrates not only lower FLOPs but also surpasses the results of SkySense Guo et al. [2024] by 3.57%. As shown in table 6 and table 7, RS-vHeat also achieves strong performance and high computational efficiency on two additional fine-grained datasets.

Change Detection. We train and test on the LEVIR-CD dataset Chen and Shi [2020], employing the BIT architecture Chen et al. [2021] with cross-entropy loss. RS-vHeat demonstrates superior adaptability, achieving an F1 score of

消融研究。许多RSFMs主要关注光学数据而不支持SAR输入，因此我们将RS-vHeat与其他RSFMs使用光学数据集进行比较，而SAR特定任务则与专门模型进行比较。基于骨干网络，RSFMs被分为三类：CNN基、ViT基和Swin基。所有数据集均遵循官方的训练和测试分割方法。有关数据集和可视化方面的更多细节请参见补充材料。

单模态和多模态语义分割。我们在两个光学数据集（Potsdam Sherrah [2016]，iSAID Waqas Zamir et al. [2019]），一个SAR数据集（Air-PolSAR-Seg Wang et al. [2022b]）和一个多模态数据集（Li et al. [2022]）上评估了我们的模型，使用UPerNet Xiao et al. [2018]和交叉熵损失作为输出头。表1将RS-vHeat与其他14种RSFMs在iSAID和Potsdam上的表现进行了比较。RS-vHeat在准确度方面优于基于CNN的模型，并且其FLOPs低于基于ViT的模型，例如，其在Scale-MAE Reed et al. [2023]上的改进分别为1.28%和2.95%。此外，它在基于Swin的模型中拥有最低的FLOPs，与SkySense Guo et al. [2024]相比，参数量减少了四倍以上，准确度仅下降1.17%和2.19%。此外，在AIR-PolSAR-Seg（表3）和WHU-OPT-SAR（表2）上，RS-vHeat也优于其他专门的分割模型，展示了热传导模型在SAR和多模态分割任务中的适用性。

表1：与其他RSFMs在iSAID和Potsdam上的mF1、mIoU、参数和FLOPs比较。

Backbone Type	Method	Backbone	Potsdam mF1 ↑	iSAID mIoU ↑	Image Size	Params	FLOPs ↓
CNN-Based	GASSL (ICCV'2021) Ayush et al. [2021]	ResNet-50	91.27	65.95	896 ²	64M	722G
	SeCo (ICCV'2021) Manas et al. [2021]	ResNet-50	89.03	57.20	896 ²	64M	722G
	SSL4EO (GRSM'2023) Wang et al. [2023]	ResNet-50	91.54	64.01	896 ²	64M	722G
	CACo (CVPR'2023) Mall et al. [2023]	ResNet-18	91.35	64.32	896 ²	41M	671G
	TOV (JSTARS'2023) Tao et al. [2023]	ResNet-50	92.03	66.24	896 ²	64M	722G
	SAMRS (NeurIPS'2023) Wang et al. [2024b]	ResNet-50	91.43	66.26	896 ²	64M	722G
	RS-vHeat (Ours)	vHeat-B + SCEs	92.82	68.72	896 ²	148M	921G
ViT-Based	RVSA (TGRS'2022) Wang et al. [2022a]	ViT-B + RVSA	-	64.49	896 ²	128M	1043G
	SatMAE (NeurIPS'2022) Cong et al. [2022]	ViT-L	90.63	62.97	896 ²	341M	1536G
	Scale-MAE (ICCV'2023) Reed et al. [2023]	ViT-L	91.54	65.77	896 ²	341M	1536G
	RS-vHeat (Ours)	vHeat-B + SCEs	92.82	68.72	896 ²	148M	921G
Swin-Based	RingMo (TGRS'2022) Sun et al. [2022b]	Swin-B	91.27	67.20	896 ²	121M	968G
	SatLas (ICCV'2023) Bastani et al. [2023]	Swin-B	91.28	68.71	896 ²	121M	968G
	GFM (ICCV'2023) Mendieta et al. [2023]	Swin-B	91.85	66.62	896 ²	121M	968G
	CMID (TGRS'2023) Muhtar et al. [2023]	Swin-B	91.86	66.21	896 ²	121M	968G
	SkySense (CVPR'2024) Gue et al. [2024]	Swin-H	93.99	70.91	896 ²	>702M	>2708G
	RS-vHeat (Ours)	vHeat-B + SCEs	92.82	68.72	896 ²	148M	921G

表2：OA 和用户准确率在 WHU-OPT-SAR 上与其他专门模型的比较

Method	Publication	User's Accuracy ↑							OA ↑
		Farmland	City	Village	Water	Forest	Road	Others	
SegFormer Xie et al. [2021b]	NeurIPS'2021	79.1	72.9	38.0	64.7	88.1	0.3	0.4	75.5
Segmenter Strudel et al. [2021]	ICCV'2021	82.3	75.2	51.7	74.4	89.4	16.0	12.0	79.9
MCANet Li et al. [2022]	JAG'2022	74.3	62.2	53.1	65.7	95.5	31.0	9.8	82.9
VMamba Liu et al. [2024a]	NeurIPS'2024	82.1	74.1	57.8	81.4	89.1	41.5	18.4	81.4
MMOKD Liu et al. [2024b]	TGRS'2024	70.0	58.1	50.0	69.7	80.1	40.1	25.0	82.5
RS-vHeat (Ours)	-	81.1	67.3	67.5	79.0	90.2	54.9	55.3	83.9

表3：AIR-PolSAR-Seg 上 mIoU、OA 和 AA 与其他专门模型的比较

Method	Publication	mIoU ↑	OA ↑	AA ↑
DeepLab V3+ Chen et al. [2018]	ECCV'2018	48.21	76.81	63.55
EncNet Zhang et al. [2018]	CVPR'2018	47.75	75.67	57.51
PSANet Zhao et al. [2018]	ECCV'2018	47.14	76.21	62.92
CCNet Huang et al. [2019]	ICCV'2019	46.46	75.53	55.83
DANet Fu et al. [2019]	CVPR'2019	51.93	76.91	62.79
GCNet Cao et al. [2019]	ICCV'2019	47.56	76.75	57.10
RS-vHeat (Ours)	-	57.46	81.46	65.92

目标检测。我们在两个光学数据集（FAIR1M Sun et al. [2022a]，DIOR Li et al. [2020a]）和一个SAR数据集（SAR-AIRCRAFT-1.0 Zhirui et al. [2023]）上进行了粗粒度和细粒度实验，使用YOLOX Ge et al. [2021]作为输出头。在DIOR数据集上，表4显示RS-vHeat不仅具有更低的FLOPs，还在SkySense Guo et al. [2024]的基础上提高了3.57%。如表6和表7所示，RS-vHeat在两个额外的细粒度数据集上也实现了强大的性能和高计算效率。

Change Detection. 我们在 LEVIR-CD 数据集 Chen 和 Shi [2020] 上进行训练和测试，采用 Chen 等人 [2021] 的 BIT 架构并使用交叉熵损失。RS-vHeat 展示出更强的适应性，达到的 F1 分数为

Table 4: Comparison of mAP₅₀, parameters, and FLOPs on DIOR with other RSFMs.

Backbone Type	Method	Backbone	Image Size	mAP ₅₀ ↑	Params	FLOPs ↓
CNN-Based	GASSL (ICCV'2021) Ayush et al. [2021]	ResNet-50	800 ²	67.40	41M	134G
	CACo (CVPR'2023) Mall et al. [2023]	ResNet-18	800 ²	66.91	28M	101G
	TOV (JSTARS'2023) Tao et al. [2023]	ResNet-50	800 ²	70.16	41M	134G
	SSL4EO (GRSM'2023) Wang et al. [2023]	ResNet-50	800 ²	64.82	41M	134G
	RS-vHeat (Ours)	vHeat-B + SCEs	800 ²	82.30	128M	266G
ViT-Based	RVSA (TGRS'2022) Wang et al. [2022a]	ViT-B + RVSA	800 ²	73.22	113M	378G
	SatMAE (NeurIPS'2022) Cong et al. [2022]	ViT-L	800 ²	70.89	324M	1094G
	Scale-MAE (ICCV'2023) Reed et al. [2023]	ViT-L	800 ²	73.81	324M	1094G
	RS-vHeat (Ours)	vHeat-B + SCEs	800 ²	82.30	128M	266G
Swin-Based	RingMo (TGRS'2022) Sun et al. [2022b]	Swin-B	800 ²	75.90	105M	322G
	SatLas (ICCV'2023) Bastani et al. [2023]	Swin-B	800 ²	74.10	105M	322G
	GFM (ICCV'2023) Mendieta et al. [2023]	Swin-B	800 ²	72.84	105M	322G
	CMID (TGRS'2023) Muhtar et al. [2023]	Swin-B	800 ²	75.11	105M	322G
	SkySense (CVPR'2024) Guo et al. [2024]	Swin-H	800 ²	78.73	>674M	>1679G
	RS-vHeat (Ours)	vHeat-B + SCEs	800 ²	82.30	128M	266G

Table 5: Comparison of OA, parameters, and FLOPs on AID and NWPU-RESISC45 with other RSFMs.

Backbone Type	Method	Backbone	AID		NWPU-RESISC45		Image Size	Params	FLOPs ↓
			TR=20%	TR=50%	TR=10%	TR=20%			
CNN-Based	GASSL (ICCV'2021) Ayush et al. [2021]	ResNet-50	93.55	95.92	90.86	93.06	1024 ²	24M	87G
	SeCo (ICCV'2021) Manas et al. [2021]	ResNet-50	93.47	95.99	89.64	92.91	1024 ²	24M	87G
	CACo (CVPR'2023) Mall et al. [2023]	ResNet-18	90.88	95.05	88.28	91.94	1024 ²	11M	38G
	TOV (JSTARS'2023) Tao et al. [2023]	ResNet-50	95.16	97.09	90.97	93.79	1024 ²	24M	87G
	SSL4EO (GRSM'2023) Wang et al. [2023]	ResNet-50	91.06	94.74	87.60	91.27	1024 ²	24M	87G
	RS-vHeat (Ours)	vHeat-B + SCEs	96.81	97.58	92.01	95.66	1024 ²	150M	340G
ViT-Based	RVSA (TGRS'2022) Wang et al. [2022a]	ViT-B + RVSA	97.03	98.50	93.93	95.69	1024 ²	89M	460G
	SatMAE (NeurIPS'2022) Cong et al. [2022]	ViT-L	95.02	96.94	91.72	94.10	1024 ²	310M	2070G
	Scale-MAE (ICCV'2023) Reed et al. [2023]	ViT-L	96.44	97.58	92.63	95.04	1024 ²	310M	2070G
	RS-vHeat (Ours)	vHeat-B + SCEs	96.81	97.58	92.01	95.66	1024 ²	150M	340G
Swin-Based	RingMo (TGRS'2022) Sun et al. [2022b]	Swin-B	96.90	98.34	94.25	95.67	1024 ²	87M	450G
	SatLas (ICCV'2023) Bastani et al. [2023]	Swin-B	94.96	97.38	92.16	94.70	1024 ²	87M	450G
	GFM (ICCV'2023) Mendieta et al. [2023]	Swin-B	95.47	97.09	92.73	94.64	1024 ²	87M	450G
	CMID (TGRS'2023) Muhtar et al. [2023]	Swin-B	96.11	97.79	94.05	95.53	1024 ²	87M	450G
	SkySense (CVPR'2024) Guo et al. [2024]	Swin-H	97.68	98.60	94.85	96.32	1024 ²	>660M	>2760G
	RS-vHeat (Ours)	vHeat-B + SCEs	96.81	97.58	92.01	95.66	1024 ²	150M	340G

93.48%, outperforming existing methods and surpassing SkySense Guo et al. [2024] by 0.9 points, as shown in table 8. Despite a 256-pixel input size, RS-vHeat maintains FLOPs comparable to the Swin-B-based baseline while delivering higher precision.

Image Classification. We validate our model on two benchmark datasets (AID Xia et al. [2017], NWPU-RESISC45 Cheng et al. [2017]) by attaching a classification head designed, and employ cross-entropy loss for computation. As shown in table 5, RS-vHeat surpasses CNN-based models like CACo Mall et al. [2023] in classification accuracy. Compared to ViT- and Swin-based models, RS-vHeat demonstrates superior computational efficiency while also achieving competitive accuracy.

4.3 Ablations

To validate the effectiveness of the features learned through pre-training based on heat conduction theory, we conduct ablation studies on key components.

Effectiveness of Masked Training. We visualize the accuracy curves of the visual encoder before and after introducing new structures across downstream tasks, as shown in fig. 5. All three networks are pre-trained on the same multi-modal RS data. In (a), we use vHeat-B with the SimMIM Xie et al. [2022] pixel mask training scheme; in (b), we add frequency domain masking and multi-domain reconstruction; and in (c), we integrate the innovative RS-vHeat structure, which includes SCEs. For optical tasks, RS-vHeat shows faster accuracy improvement, stabilizing at a higher ceiling. It

Table 6: Comparison of mAP₅₀ and mAP₇₅ on SAR-AIRcraft-1.0 with other specialized models.

Method	Publication	mAP ₅₀ ↑	mAP ₇₅ ↑
Faster R-CNN Ren et al. [2016]	TPAMI'2016	76.1	62.2
Cascade R-CNN Cai and Vasconcelos [2018]	CVPR'2018	75.7	58.9
RepPoints Yang et al. [2019]	ICCV'2019	72.6	53.3
SKG-Net Fu et al. [2021]	JSTARS'2021	70.7	46.4
SA-Net Zhirui et al. [2023]	RADARS'2023	77.7	62.8
RS-vHeat (Ours)	-	87.1	67.4

表4：与其他RSFMs相比，mAP₅₀、参数和FLOPs的比较。

Backbone Type	Method	Backbone	Image Size	mAP ₅₀ ↑	Params	FLOPs ↓
CNN-Based	GASSL (ICCV'2021) Ayush et al. [2021]	ResNet-50	800 ²	67.40	41M	134G
	CACo (CVPR'2023) Mall et al. [2023]	ResNet-18	800 ²	66.91	28M	101G
	TOV (JSTARS'2023) Tao et al. [2023]	ResNet-50	800 ²	70.16	41M	134G
	SSL4EO (GRSM'2023) Wang et al. [2023]	ResNet-50	800 ²	64.82	41M	134G
	RS-vHeat (Ours)	vHeat-B + SCEs	800 ²	82.30	128M	266G
ViT-Based	RVSA (TGRS'2022) Wang et al. [2022a]	ViT-B + RVSA	800 ²	73.22	113M	378G
	SatMAE (NeurIPS'2022) Cong et al. [2022]	ViT-L	800 ²	70.89	324M	1094G
	Scale-MAE (ICCV'2023) Reed et al. [2023]	ViT-L	800 ²	73.81	324M	1094G
	RS-vHeat (Ours)	vHeat-B + SCEs	800 ²	82.30	128M	266G
Swin-Based	RingMo (TGRS'2022) Sun et al. [2022b]	Swin-B	800 ²	75.90	105M	322G
	SatLas (ICCV'2023) Bastani et al. [2023]	Swin-B	800 ²	74.10	105M	322G
	GFM (ICCV'2023) Mendieta et al. [2023]	Swin-B	800 ²	72.84	105M	322G
	CMID (TGRS'2023) Muhtar et al. [2023]	Swin-B	800 ²	75.11	105M	322G
	SkySense (CVPR'2024) Guo et al. [2024]	Swin-H	800 ²	78.73	>674M	>1679G
	RS-vHeat (Ours)	vHeat-B + SCEs	800 ²	82.30	128M	266G

表5：与其他RSFMs在AID和NWPU-RESISC45上的OA、参数和FLOPs比较。

Backbone Type	Method	Backbone	AID		NWPU-RESISC45		Image Size	Params	FLOPs ↓
			TR=20%	TR=50%	TR=10%	TR=20%			
CNN-Based	GASSL (ICCV'2021) Ayush et al. [2021]	ResNet-50	93.55	95.92	90.86	93.06	1024 ²	24M	87G
	SeCo (ICCV'2021) Manas et al. [2021]	ResNet-50	93.47	95.99	89.64	92.91	1024 ²	24M	87G
	CACo (CVPR'2023) Mall et al. [2023]	ResNet-18	90.88	95.05	88.28	91.94	1024 ²	11M	38G
	TOV (JSTARS'2023) Tao et al. [2023]	ResNet-50	95.16	97.09	90.97	93.79	1024 ²	24M	87G
	SSL4EO (GRSM'2023) Wang et al. [2023]	ResNet-50	91.06	94.74	87.60	91.27	1024 ²	24M	87G
	RS-vHeat (Ours)	vHeat-B + SCEs	96.81	97.58	92.01	95.66	1024 ²	150M	340G
ViT-Based	RVSA (TGRS'2022) Wang et al. [2022a]	ViT-B + RVSA	97.03	98.50	93.93	95.69	1024 ²	89M	460G
	SatMAE (NeurIPS'2022) Cong et al. [2022]	ViT-L	95.02	96.94	91.72	94.10	1024 ²	310M	2070G
	Scale-MAE (ICCV'2023) Reed et al. [2023]	ViT-L	96.44	97.58	92.63	95.04	1024 ²	310M	2070G
	RS-vHeat (Ours)	vHeat-B + SCEs	96.81	97.58	92.01	95.66	1024 ²	150M	340G
Swin-Based	RingMo (TGRS'2022) Sun et al. [2022b]	Swin-B	96.90	98.34	94.25	95.67	1024 ²	87M	450G
	SatLas (ICCV'2023) Bastani et al. [2023]	Swin-B	94.96	97.38	92.16	94.70	1024 ²	87M	450G
	GFM (ICCV'2023) Mendieta et al. [2023]	Swin-B	95.47	97.09	92.73	94.64	1024 ²	87M	450G
	CMID (TGRS'2023) Muhtar et al. [2023]	Swin-B	96.11	97.79	94.05	95.53	1024 ²	87M	450G
	SkySense (CVPR'2024) Guo et al. [2024]	Swin-H	97.68	98.60	94.85	96.32	1024 ²	>660M	>2760G
	RS-vHeat (Ours)	vHeat-B + SCEs	96.81	97.58	92.01	95.66	1024 ²	150M	340G

93.48%，超过现有方法并比 SkySense Guo 等人 [2024] 高出 0.9 分，如表 8 所示。尽管输入大小为 256 像素，RS-vHeat 保持与基于 Swin-B 的基线相当的 FLOPs，同时提供更高的精度。

图像分类。我们通过附加一个设计好的分类头，在两个基准数据集（AID Xia et al. [2017], NWPU-RESISC45 Cheng et al. [2017]）上验证了我们的模型，并采用交叉熵损失进行计算。如表5所示，RS-vHeat 在分类精度上超过了像 CACo Mall et al. [2023] 这样的基于CNN的模型。与基于ViT-和Swin的模型相比，RS-vHeat 在计算效率上表现出优越性，同时也能达到竞争性的精度。

4.3 削减实验

为了验证基于热传导理论预训练学到的特征的有效性，我们在关键组件上进行了消融研究。

Masked训练的有效性。我们可视化了在下游任务中引入新结构前后视觉编码器的准确率曲线，如图5所示。所有三个网络都在相同的多模态RS数据上进行预训练。在(a)中，我们使用vHeat-B和SimMIM Xie et al. [2022]的像素掩蔽训练方案；在(b)中，我们增加了频域掩蔽和多域重建；在(c)中，我们整合了创新的RS-vHeat结构，其中包括SCEs。对于光学任务，RS-vHeat显示出更快的准确率提升，并稳定在更高的上限。

表6：SAR-AIRcraft-1.0 上 mAP₅₀ 和 mAP₇₅ 与其他专门模型的比较。

Method	Publication	mAP ₅₀ ↑	mAP ₇₅ ↑
Faster R-CNN Ren et al. [2016]	TPAMI'2016	76.1	62.2
Cascade R-CNN Cai and Vasconcelos [2018]	CVPR'2018	75.7	58.9
RepPoints Yang et al. [2019]	ICCV'2019	72.6	53.3
SKG-Net Fu et al. [2021]	JSTARS'2021	70.7	46.4
SA-Net Zhirui et al. [2023]	RADARS'2023	77.7	62.8
RS-vHeat (Ours)	-	87.1	67.4

Table 7: Comparison of mAP, parameters, and FLOPs on FAIR1M-2.0 with other RSFMs.

Method	Backbone	mAP \uparrow	Params	FLOPs \downarrow
CACo Mall et al. [2023]	ResNet-18	47.83	42M	64G
GASSL Ayush et al. [2021]	ResNet-50	48.15	55M	77G
TOV Tao et al. [2023]	ResNet-50	49.62	55M	77G
SSL4EO Wang et al. [2023]	ResNet-50	49.37	55M	77G
RVSA Wang et al. [2022a]	ViT-B + RVSA	47.04	126M	160G
SatMAE Cong et al. [2022]	ViT-L	46.55	336M	241G
Scale-MAE Reed et al. [2023]	ViT-L	48.31	336M	241G
SatLas Bastani et al. [2023]	Swin-B	46.19	119M	167G
GFM Mendieta et al. [2023]	Swin-B	49.69	119M	167G
RingMo Sun et al. [2022b]	Swin-B	46.21	119M	167G
CMID Muhtar et al. [2023]	Swin-B	50.58	119M	167G
SkySense Guo et al. [2024]	Swin-H	54.57	>688M	>900G
RS-vHeat	vHeat-B + SCEs	48.29	130M	137G

Table 8: Comparison of F1 on LEVIR-CD with other RSFMs.

Method	Backbone	F1 \uparrow	Params	FLOPs \downarrow
CACo Mall et al. [2023]	ResNet-18	81.04	12M	11G
GASSL Ayush et al. [2021]	ResNet-50	78.19	27M	25G
SeCo Manas et al. [2021]	ResNet-50	90.14	27M	25G
SSL4EO Wang et al. [2023]	ResNet-50	89.05	27M	25G
RVSA Wang et al. [2022a]	ViT-B + RVSA	90.86	94M	57G
SatMAE Cong et al. [2022]	ViT-L	87.65	304M	162G
Scale-MAE Reed et al. [2023]	ViT-L	92.07	304M	162G
SatLas Bastani et al. [2023]	Swin-B	90.62	88M	45G
GFM Mendieta et al. [2023]	Swin-B	91.73	88M	45G
RingMo Sun et al. [2022b]	Swin-B	91.86	88M	45G
CMID Muhtar et al. [2023]	Swin-B	91.72	88M	45G
SkySense Guo et al. [2024]	Swin-H	92.58	>656M	>307G
RS-vHeat (Ours)	vHeat-B + SCEs	93.48	93M	46G

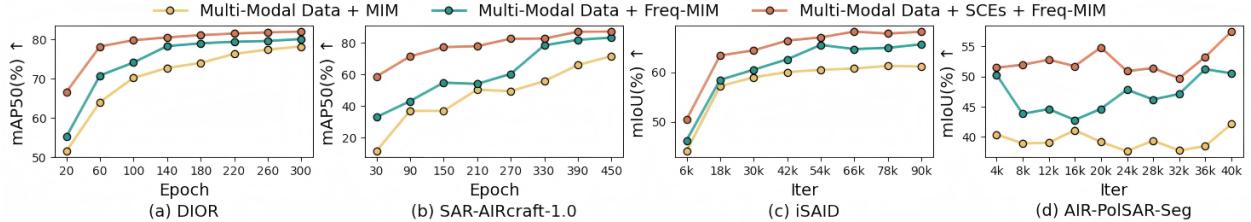


Figure 5: Accuracy curves of various methods across modalities. (a) optical object detection on DIOR, (b) SAR object detection on SAR-AIRcraft-1.0, (c) optical semantic segmentation on iSAID, and (d) SAR semantic segmentation on AIR-PolSAR-Seg.

surpasses 80% mAP₅₀ on the DIOR dataset by the 120th epoch. For SAR tasks, while there are fluctuations, RS-vHeat demonstrates better adaptability and prediction accuracy. Overall, the improvements in pre-training and SCEs lead to advanced feature extraction performance.

Reconstruction Learning Strategy. table 9 demonstrates the effectiveness of different loss components—spatial domain reconstruction loss (SDR), frequency domain reconstruction loss (FDR), and contrastive loss (CL). When only SDR is applied (row a), performance drops slightly on both optical and SAR datasets. Adding FDR (row b) improves performance by leveraging frequency domain information. Introducing CL (row c) further enhances performance by enforcing similarity constraints in the multi-modal feature space, improving feature learning. This leads to an 83.9% OA on WHU-OPT-SAR, surpassing rows (a) and (b) by 3.4% and 1.6%, respectively. These results highlight the importance of combining these three losses for improved predictions across tasks and modalities.

Table 9: Results of RS-vHeat under various constraints of self-supervised learning strategies.

#	Loss			Object Detection		Semantic Segmentation		
	SDR	FDR	CL	DIOR (Optical) mAP ₅₀ \uparrow	SAR-AIRcraft-1.0 (SAR) mAP ₅₀ \uparrow	iSAID (Optical) mIoU \uparrow	AIR-PolSAR-Seg (SAR) mIoU \uparrow	WHU-OPT-SAR (Optical+SAR) OA \uparrow
(a)	✓	✗	✗	78.2	84.6	66.1	54.7	80.5
(b)	✓	✓	✗	79.5	86.4	67.2	55.1	82.3
(c)	✓	✓	✓	82.3	87.1	68.7	57.5	83.9

表7：与其他RSFMs在FAIR1M-2.0上的mAP、参数和FLOPs比较

Method	Backbone	mAP↑	Params	FLOPs↓
CACo Mall et al. [2023]	ResNet-18	47.83	42M	64G
GASSL Ayush et al. [2021]	ResNet-50	48.15	55M	77G
TOV Tao et al. [2023]	ResNet-50	49.62	55M	77G
SSL4EO Wang et al. [2023]	ResNet-50	49.37	55M	77G
RVSA Wang et al. [2022a]	ViT-B + RVSA	47.04	126M	160G
SatMAE Cong et al. [2022]	ViT-L	46.55	336M	241G
Scale-MAE Reed et al. [2023]	ViT-L	48.31	336M	241G
SatLas Bastani et al. [2023]	Swin-B	46.19	119M	167G
GFM Mendieta et al. [2023]	Swin-B	49.69	119M	167G
RingMo Sun et al. [2022b]	Swin-B	46.21	119M	167G
CMID Muhtar et al. [2023]	Swin-B	50.58	119M	167G
SkySense Guo et al. [2024]	Swin-H	54.57	>688M	>900G
RS-vHeat	vHeat-B + SCEs	48.29	130M	137G

表8：与其他RSFMs在LEVIR-CD上的F1比较。

Method	Backbone	F1↑	Params	FLOPs↓
CACo Mall et al. [2023]	ResNet-18	81.04	12M	11G
GASSL Ayush et al. [2021]	ResNet-50	78.19	27M	25G
SeCo Manas et al. [2021]	ResNet-50	90.14	27M	25G
SSL4EO Wang et al. [2023]	ResNet-50	89.05	27M	25G
RVSA Wang et al. [2022a]	ViT-B + RVSA	90.86	94M	57G
SatMAE Cong et al. [2022]	ViT-L	87.65	304M	162G
Scale-MAE Reed et al. [2023]	ViT-L	92.07	304M	162G
SatLas Bastani et al. [2023]	Swin-B	90.62	88M	45G
GFM Mendieta et al. [2023]	Swin-B	91.73	88M	45G
RingMo Sun et al. [2022b]	Swin-B	91.86	88M	45G
CMID Muhtar et al. [2023]	Swin-B	91.72	88M	45G
SkySense Guo et al. [2024]	Swin-H	92.58	>656M	>307G
RS-vHeat (Ours)	vHeat-B + SCEs	93.48	93M	46G

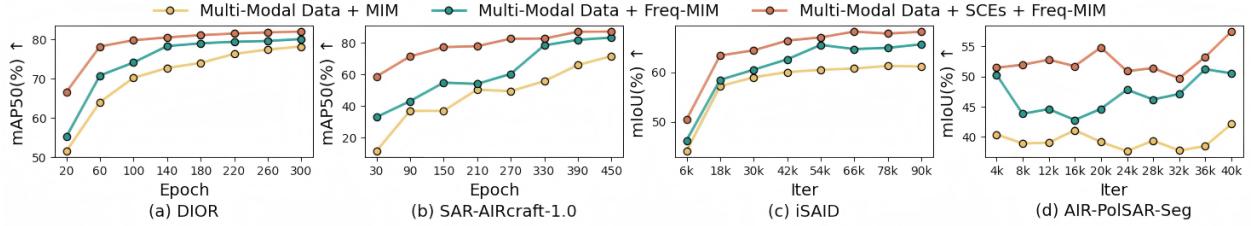


Figure 5: 各种方法在不同模态下的准确性曲线。(a) 光学目标检测在DIOR, (b) SAR目标检测在SAR-AIRcraft-1.0, (c) 光学语义分割在iSAID, 和 (d) SAR语义分割在AIR-PolSAR-Seg。

超越DIOR数据集上80% mAP₅₀的性能，在第120个epoch。对于SAR任务，尽管存在波动，RS-vHeat展示了更好的适应性和预测准确性。总体而言，预训练改进和SCEs导致了更先进的特征提取性能。

重建学习策略。表9展示了不同损失组件的有效性——空间域重建损失（SDR）、频率域重建损失（FDR）和对比损失（CL）。仅应用SDR（行a）时，性能在光学和SAR数据集上略有下降。添加FDR（行b）通过利用频率域信息提高了性能。引入CL（行c）进一步通过在多模态特征空间中施加相似性约束来增强性能，改善了特征学习。这导致在WHU-OPT-SAR上的OA达到83.9%，分别超过行a和行b的3.4%和1.6%。这些结果突显了结合这三种损失的重要性，以提高跨任务和模态的预测性能。

表9：自监督学习策略在不同约束条件下的RS-vHeat结果。

#	Loss			Object Detection		Semantic Segmentation		
	SDR	FDR	CL	DIOR (Optical) mAP ₅₀ ↑	SAR-AIRcraft-1.0 (SAR) mAP ₅₀ ↑	iSAID (Optical) mIoU↑	AIR-PolSAR-Seg (SAR) mIoU↑	WHU-OPT-SAR (Optical+SAR) OA↑
(a)	✓	✗	✗	78.2	84.6	66.1	54.7	80.5
(b)	✓	✓	✗	79.5	86.4	67.2	55.1	82.3
(c)	✓	✓	✓	82.3	87.1	68.7	57.5	83.9

5 Conclusion

In this work, we introduce the concept of heat conduction into RS tasks for the first time and establish a multi-modal RSFM, RS-vHeat. By employing a self-supervised learning strategy that integrates frequency domain masking and multi-domain reconstruction, along with a heat conduction operator incorporating spatial correction embeddings, we propose an approach that balances computational complexity and global receptive field coverage in RS. Our method captures global details from both spatial and frequency domains, significantly reducing the issue of small object omission, and achieves consistent multi-modal feature representation by mapping images into a high-dimensional thermal space. In future work, we plan to propose new solutions that offer fresh insights and methods for tackling visual modeling challenges across various industries.

References

- Jamie Sherrah. Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. *arXiv preprint arXiv:1606.02585*, 2016.
- Yuxiang Zhang, Wei Li, Ran Tao, Jiangtao Peng, Qian Du, and Zhaoquan Cai. Cross-scene hyperspectral image classification with discriminative cooperative alignment. *IEEE Transactions on Geoscience and Remote Sensing*, 59(11):9646–9660, 2021.
- Xian Sun, Peijin Wang, Zhiyuan Yan, Feng Xu, Ruiping Wang, Wenhui Diao, Jin Chen, Jihao Li, Yingchao Feng, Tao Xu, et al. Fair1m: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 184:116–130, 2022a.
- Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159:296–307, 2020a.
- Xiaonan Lu, Xian Sun, Wenhui Diao, Yingchao Feng, Peijin Wang, and Kun Fu. Lil: Lightweight incremental learning approach through feature transfer for remote sensing image scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–20, 2021.
- Xian Sun, Peijin Wang, Cheng Wang, Yingfei Liu, and Kun Fu. Pbnet: Part-based convolutional neural network for complex composite object detection in remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 173:50–65, 2021.
- Kun FU, Xian SUN, Xiaolan QIU, Wenhui DIAO, Zhiyuan YAN, Lijia HUANG, and Hongfeng YU. Multi-satellite integrated processing and analysis method under remote sensing big data. *National Remote Sensing Bulletin*, 25(3):691–707, 2021. doi:10.11834/jrs.20211058.
- Wenyuan Li, Keyan Chen, Hao Chen, and Zhenwei Shi. Geographical knowledge-driven representation learning for remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2021.
- Oscar Manas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodriguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9414–9423, 2021.
- Utkarsh Mall, Bharath Hariharan, and Kavita Bala. Change-aware sampling and contrastive learning for satellite images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5261–5270, 2023.
- Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10181–10190, 2021.
- Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022.
- Di Wang, Qiming Zhang, Yufei Xu, Jing Zhang, Bo Du, Dacheng Tao, and Liangpei Zhang. Advancing plain vision transformer toward remote sensing foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2022a.
- Chao Tao, Ji Qi, Guo Zhang, Qing Zhu, Weipeng Lu, and Haifeng Li. Tov: The original vision model for optical remote sensing image understanding via self-supervised learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:4916–4930, 2023.
- Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4088–4099, 2023.

5 结论

在本工作中，我们首次将热传导的概念引入到RS任务中，并建立了多模态RSFM, RS-vHeat。通过结合频域掩码和多域重建的自监督学习策略，以及包含空间校正嵌入的热传导算子，我们提出了一种在RS中平衡计算复杂性和全局感受野覆盖的方法。我们的方法从空间域和频域中捕获全局细节，显著减少了小对象遗漏的问题，并通过将图像映射到高维热空间中实现了多模态特征表示的一致性。在未来的工作中，我们计划提出新的解决方案，为跨不同行业的视觉建模挑战提供新的见解和方法。

参考文献

- Jamie Sherrah. 全卷积网络在高分辨率航空影像密集语义标注中的应用。 *arXiv preprint arXiv:1606.02585*, 2016。
- 张玉翔, 李卫, 貂然, 彭江涛, 杜千, 以及 蔡钊全. 跨场景高光谱图像分类的判别性协同对齐。*IEEE Transactions on Geoscience and Remote Sensing*, 59(11):9646–9660, 2021.
- 孙贤, 王培金, 闫志远, 徐峰, 王瑞萍, 董文慧, 陈进, 李际浩, 邢颖超, 徐涛, 等. Fair1m: 高分辨率遥感影像中细粒度目标识别的基准数据集。*ISPRS Journal of Photogrammetry and Remote Sensing*, 184:116–130, 2022a.
- 李凯, 汪刚, 陈工, 孙立秋, 韩军伟. 光学遥感图像中的目标检测: 一项综述和一个新的基准。*ISPRS journal of photogrammetry and remote sensing*, 159: 296–307, 2020a.
- 陆晓南, 孙贤, 董文慧, 邢颖超, 王培金, 付坤. Lil: 通过特征转移实现轻量级增量学习的遥感图像场景分类方法。*IEEE Transactions on Geoscience and Remote Sensing*, 60:1–20, 2021.
- 孙贤, 王培金, 王成, 刘Yi ngfei, 付坤. Pbnet: 遥感影像中复杂复合对象检测的部分卷积神经网络。*ISPRS Journal of Photogrammetry and Remote Sensing*, 173:50–65, 2021.
- 付坤, 孙贤, 邱晓兰, 董文慧, 闫志远, 黄利佳, 于洪峰. 遥感大数据下的多卫星综合处理与分析方法。*National Remote Sensing Bulletin*, 25(3): 691–707, 2021. doi:10.11834/jrs.20211058.
- 李文远, 陈凯yan, 陈浩, 石振伟. 地理知识驱动的遥感图像表示学习。*IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2021.
- Oscar Manas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, 和 Pau Rodriguez. Seasonal contrast: 从未整理的遥感数据中进行无监督预训练. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 页码 9414–9423, 2021.
- Utkarsh Mall, Bharath Hariharan, 和 Kavita Bala. 为卫星图像进行变化感知采样和对比学习. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 页码 5261–5270, 2023.
- Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, 和 Stefano Ermon. 地理意识的自监督学习. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 页码 10181–10190, 2021.
- 康泽臻, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, 赫玉彤, Marshall Burke, David Lobell, 和 Stefano Ermon. Satmae: 为时空多光谱卫星图像预训练的变压器。*Advances in Neural Information Processing Systems*, 35:197–211, 2022.
- 王迪, 张启明, 徐宇飞, 张静, Du Bo, Tao Dacheng, 和 张良培. 使普通视觉变压器成为遥感基础模型。*IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–15, 2022a.
- 唐超, 祁济, 张郭, 朱青, 陆卫朋, 李海峰. Tov: 通过自监督学习实现光学遥感图像理解的原始视觉模型。*IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:4916–4930, 2023.
- Colorado J Reed, Ritwik Gupta, 李书帆, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, 和 Trevor Darrell. Scale-mae: 多尺度地理空间表示学习的尺度感知掩码自编码器. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 页码 4088–4099, 2023.

- Matías Mendieta, Boran Han, Xingjian Shi, Yi Zhu, and Chen Chen. Towards geospatial foundation models via continual pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16806–16816, 2023.
- Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. Satlaspretrain: A large-scale dataset for remote sensing image understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16772–16782, 2023.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_2016/file/c8067ad1937f728f51288b3eb986afaa-Paper.pdf.
- Emmanuel Christophe, Julien Michel, and Jordi Ingla. Remote sensing processing: From multicore to gpu. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 4(3):643–652, 2011.
- Yan Ma, Haiping Wu, Lizhe Wang, Bormin Huang, Rajiv Ranjan, Albert Zomaya, and Wei Jie. Remote sensing big data computing: Challenges and opportunities. *Future Generation Computer Systems*, 51:47–60, 2015. ISSN 0167-739X. doi:<https://doi.org/10.1016/j.future.2014.10.029>. URL <https://www.sciencedirect.com/science/article/pii/S0167739X14002234>. Special Section: A Note on New Trends in Data-Aware Scheduling and Resource Provisioning in Modern HPC Systems.
- Jiang Li and Ram M Narayanan. A shape-based approach to change detection of lakes using time series remote sensing images. *IEEE transactions on geoscience and remote sensing*, 41(11):2466–2477, 2003.
- Huanfeng Shen, Menghui Jiang, Jie Li, Chenxia Zhou, Qiangqiang Yuan, and Liangpei Zhang. Coupling model- and data-driven methods for remote sensing image restoration and fusion: Improving physical interpretability. *IEEE Geoscience and Remote Sensing Magazine*, 10(2):231–249, 2022. doi:10.1109/MGRS.2021.3135954.
- Anastasios Temenos, Nikos Temenos, Maria Kaselimi, Anastasios Doulamis, and Nikolaos Doulamis. Interpretable deep learning framework for land use and land cover classification in remote sensing using shap. *IEEE Geoscience and Remote Sensing Letters*, 20:1–5, 2023. doi:10.1109/LGRS.2023.3251652.
- Adrián Pérez-Suay, Jose E. Adsuara, María Piles, Laura Martínez-Ferrer, Emiliiano Díaz, Alvaro Moreno-Martínez, and Gustau Camps-Valls. Interpretability of recurrent neural networks in remote sensing. In *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, pages 3991–3994, 2020. doi:10.1109/IGARSS39084.2020.9323898.
- Zhaozhi Wang, Yue Liu, Yunfan Liu, Hongtian Yu, Yaowei Wang, Qixiang Ye, and Yunjie Tian. vheat: Building vision models upon heat conduction. *arXiv preprint arXiv:2405.16555*, 2024a.
- Chi Zhang, Shiqing Wei, Shunping Ji, and Meng Lu. Detecting large-scale urban land cover changes from very high resolution remote sensing images using cnn-based classification. *ISPRS International Journal of Geo-Information*, 8(4):189, 2019.
- Haifeng Li, Haikuo Huang, Li Chen, Jian Peng, Haozhe Huang, Zhenqi Cui, Xiaoming Mei, and Guohua Wu. Adversarial examples for cnn-based sar image classification: An experience study. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:1333–1347, 2020b.
- Xiaohu Dong, Ruigang Fu, Yinghui Gao, Yao Qin, Yuanxin Ye, and Biao Li. Remote sensing object detection based on receptive field expansion block. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021.
- Xi Chen, Zhiqiang Li, Jie Jiang, Zhen Han, Shiyi Deng, Zhihong Li, Tao Fang, Hong Huo, Qingli Li, and Min Liu. Adaptive effective receptive field convolution for semantic segmentation of vhr remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 59(4):3532–3546, 2020.

Matías Mendieta, Boran Han, Xingjian Shi, Yi Zhu, 和 Chen Chen. 通过持续预训练构建地理空间基础模型. 在 *Proceedings of the IEEE/CVF International Conference on Computer Vision* 中, 第 16806–16816 页, 2023.

Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, 和 Aniruddha Kembhavi. Satlaspretrain: 用于遥感图像理解的大规模数据集. 在 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 第 16772–16782 页, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, 和 Jian Sun. 深度残差学习在图像识别中的应用. 在 *Proceedings of the IEEE conference on computer vision and pattern recognition* 中, 第 770–778 页, 2016年.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, 等. 一张图片值16x16个词: 面向大规模图像识别的变压器模型. *arXiv preprint arXiv:2010.11929*, 2020.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, 和 Baining Guo. Swin 变former: 使用移动窗口的分层视觉变形式. 在 *Proceedings of the IEEE/CVF international conference on computer vision*, 第 10012–10022 页, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 注意力是你需要的一切. *Advances in neural information processing systems*, 30, 2017.

Wenjie Luo, Yujia Li, Raquel Urtasun, 和 Richard Zemel. 理解深度卷积神经网络中的有效感受野. 在 D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, 和 R. Garnett 编辑的 *Advances in Neural Information Processing Systems*, 卷 29 。 Curran Associates, Inc., 2016。 URL https://proceedings.neurips.cc/paper_files/paper/2016/file/c8067ad1937f728f51288b3eb986afaa-Paper.pdf.

Emmanuel Christophe, Julien Michel, 和 Jordi Ingla. 远程 sensing 处理: 从多核到 GPU. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 4(3):643–652, 2011。

颜马, 吴海平, 王立哲, 黄伯民, 刘ajay, 亚伯特·佐马亚, 和 贾ie. 远程 sensing 大数据计算: 挑战与机遇.

Future Generation Computer Systems, 51:47–60, 2015. ISSN 0167-739X. doi:<https://doi.org/10.1016/j.future.2014.10.029>. URL <https://www.sciencedirect.com/science/article/pii/S0167739X14002234>. 特别部分: 关于现代HPC系统中数据感知调度和资源分配新趋势的附注。

江丽和拉姆·M·纳拉扬an. 一种基于形状的方法用于湖泊时间序列遥感图像变化检测.
IEEE transactions on geoscience and remote sensing, 41(11):2466–2477, 2003.

胡风沈, 江梦辉, 李杰, 周晨霞, 元强强, 和 张梁佩. 结合模型驱动和数据驱动方法的遥感图像恢复与融合: 提高物理可解释性. *IEEE Geoscience and Remote Sensing Magazine*, 10(2):231–249, 2022. doi:10.1109/MGRS.2021.3135954.

Anastasios Temenos, Nikos Temenos, Maria Kaselimi, Anastasios Doulamis, 和 Nikolaos Doulamis. 使用 SHAP 的可解释深度学习框架用于遥感的土地利用和土地覆盖分类. *IEEE Geoscience and Remote Sensing Letters*, 20:1–5, 2023. doi:10.1109/LGRS.2023.3251652.

Adrián Pérez-Suay, Jose E. Adsuara, María Piles, Laura Martínez-Ferrer, Emiliano Díaz, Alvaro Moreno-Martínez, 和 Gustau Camps-Valls. 远程 sensing 中循环神经网络的可解释性. 在 *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, 第 3991–3994 页, 2020 。 doi:10.1109/IGARSS39084.2020.9323898。

赵智王, 刘越, 刘云帆, 于洪天, 王 Yaowei, 叶启祥, 和 田云洁. vheat: 基于热传导构建视觉模型.
arXiv preprint arXiv:2405.16555, 2024a.

Chi Zhang, Shiqing Wei, Shunping Ji, 和 Meng Lu. 使用基于CNN的分类方法从非常高分辨率的遥感图像中检测大规模城市土地覆盖变化. *ISPRS International Journal of Geo-Information*, 8 (4):189, 2019.

李海峰, 黄海国, 陈利, 彭健, 黄浩哲, 崔振奇, 梅晓明, 和吴国华. 基于CNN的SAR图像分类的对抗样本 : 一种经验研究.*IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:1333–1347, 2020b.

徐浩东, 傅瑞刚, 高英辉, 秦瑶, 叶远欣, 和 李标. 基于感受野扩展块的遥感目标检测.
IEEE Geoscience and Remote Sensing Letters, 19:1–5, 2021.

Xi Chen, Zhiqiang Li, Jie Jiang, Zhen Han, Shiyi Deng, Zhihong Li, Tao Fang, Hong Huo, Qingli Li, and Min Liu. 适应性有效感受野卷积用于高分辨率遥感图像的语义分割.*IEEE Transactions on Geoscience and Remote Sensing*, 59(4):3532–3546, 2020.

- Tianyu Yan, Zifu Wan, and Pingping Zhang. Fully transformer network for change detection of remote sensing images. In *Proceedings of the Asian Conference on Computer Vision*, pages 1691–1708, 2022.
- Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.
- Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model, 2024a. URL <https://arxiv.org/abs/2401.10166>.
- Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024a.
- Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.
- Rui Xu, Shu Yang, Yihui Wang, Bo Du, and Hao Chen. A survey on vision mamba: Models, applications and challenges. *arXiv preprint arXiv:2404.18861*, 2024.
- Keyan Chen, Bowen Chen, Chenyang Liu, Wenyuan Li, Zhengxia Zou, and Zhenwei Shi. Rsmamba: Remote sensing image classification with state space model. *IEEE Geoscience and Remote Sensing Letters*, 2024a.
- Qinfeng Zhu, Yuanzhi Cai, Yuan Fang, Yihan Yang, Cheng Chen, Lei Fan, and Anh Nguyen. Samba: Semantic segmentation of remotely sensed images with state space model. *arXiv preprint arXiv:2404.01705*, 2024b.
- Hongruixuan Chen, Jian Song, Chengxi Han, Junshi Xia, and Naoto Yokoya. Changemamba: Remote sensing change detection with spatio-temporal state space model. *arXiv preprint arXiv:2404.03425*, 2024b.
- Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, et al. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27672–27683, 2024.
- Xian Sun, Peijin Wang, Wanxuan Lu, Zicong Zhu, Xiaonan Lu, Qibin He, Junxi Li, Xuee Rong, Zhujun Yang, Hao Chang, et al. Ringmo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 2022b.
- Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, et al. Spectralgpt: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9653–9663, 2022.
- Wenbin Xie, Dehua Song, Chang Xu, Chunjing Xu, Hui Zhang, and Yunhe Wang. Learning frequency-aware dynamic network for efficient super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4308–4317, 2021a.
- Alexander H-D Cheng and Daisy T Cheng. Heritage and early history of the boundary element method. *Engineering analysis with boundary elements*, 29(3):268–302, 2005.
- Gilbert Strang. The discrete cosine transform. *SIAM review*, 41(1):135–147, 1999.
- Sara Atito, Muhammad Awais, and Josef Kittler. Sit: Self-supervised vision transformer. *arXiv preprint arXiv:2104.03602*, 2021.
- Dilxat Muhtar, Xueliang Zhang, Pengfeng Xiao, Zhenshi Li, and Feng Gu. Cmid: A unified self-supervised learning framework for remote sensing image understanding. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–17, 2023.
- Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–37, 2019.
- Zhirui Wang, Xuan Zeng, Zhiyuan Yan, Jian Kang, and Xian Sun. Air-polsar-seg: A large-scale data set for terrain segmentation in complex-scene polsar images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:3830–3841, 2022b.
- Xue Li, Guo Zhang, Hao Cui, Shasha Hou, Shunyao Wang, Xin Li, Yujia Chen, Zhijiang Li, and Li Zhang. Mcanet: A joint semantic segmentation framework of optical and sar images for land use classification. *International Journal of Applied Earth Observation and Geoinformation*, 106:102638, 2022.

Tianyu Yan, Zifu Wan, and Pingping Zhang. 使用完全变压器网络进行遥感图像变化检测. 在 *Proceedings of the Asian Conference on Computer Vision* 中, 第 1691–1708 页, 2022 年. Albert Gu, Karan Goel, 和 Christopher Ré. 使用结构化状态空间高效建模长序列. *arXiv preprint arXiv:2111.00396*, 2021 年. Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, 和 Yunfan Liu. Vmamba: 视觉状态空间模型, 2024a. URL <https://arxiv.org/abs/2401.10166>. Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, 和 Xinggang Wang. Vision mamba: 使用双向状态空间模型的高效视觉表示学习. *arXiv preprint arXiv:2401.09417*, 2024a. Albert Gu 和 Tri Dao. Mamba: 使用选择性状态空间的状态空间模型的线性时间序列建模. *arXiv preprint arXiv:2312.00752*, 2023 年. Rui Xu, Shu Yang, Yihui Wang, Bo Du, 和 Hao Chen. 关于 vision mamba 的综述: 模型、应用和挑战. *arXiv preprint arXiv:2404.18861*, 2024 年.

陈键, 陈 Bowen, 刘晨阳, 李文远, 邹正霞, 和 石振威. Rsmamba: 基于状态空间模型的遥感图像分类. *IEEE Geoscience and Remote Sensing Letters*, 2024a.

Qinfeng Zhu, Yuanzhi Cai, Yuan Fang, Yihan Yang, Cheng Chen, Lei Fan, 和 Anh Nguyen. Samba: 遥感图像的语义分割与状态空间模型. *arXiv preprint arXiv:2404.01705*, 2024b.

陈 Hongruixuan, 宋 Jian, 韩 Chengxi, 夏 Junshi, 和 伊藤 Naoto. Changemamba: 时空状态空间模型下的遥感变化检测. *arXiv preprint arXiv:2404.03425*, 2024b.

Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, 等. Skysense: 一种面向地球观测图像通用解释的多模态遥感基础模型. 在 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 页码 27672–27683, 2024.

先Sun, 贺佩jin王, 陆wanxuan, 朱zicong, 陆xiaonan, 何qibin, 李junxi, 宋xuee, 杨zhujun, 常hao, 等. Ringmo: 一种基于掩码图像建模的遥感基础模型. *IEEE Transactions on Geoscience and Remote Sensing*, 2022b.

Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, 等. Spectralgpt: 光谱遥感基础模型. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, 和 Han Hu. Simmim: 一个简单的masked image modeling框架. 在 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 页码 9653–9663, 2022.

Wenbin Xie, Dehua Song, Chang Xu, Chunjing Xu, Hui Zhang, 和 Yunhe Wang. 学习频率感知动态网络以实现高效的超分辨率. 在 *Proceedings of the IEEE/CVF International Conference on Computer Vision* 中, 第 4308–4317 页, 2021a.

Alexander H-D Cheng 和 Daisy T Cheng. 遗产和边界元方法的早期历史. *Engineering analysis with boundary elements*, 29(3):268–302, 2005.

Gilbert Strang. 离散余弦变换. *SIAM review*, 41(1):135–147, 1999.

Sara Atito, Muhammad Awais, 和 Josef Kittler. 坐: 自我监督的视觉变换器. *arXiv preprint arXiv:2104.03602*, 2021.

Dilxat Muhtar, Xueliang Zhang, Pengfeng Xiao, Zhenshi Li, 和 Feng Gu. Cmid: 一种统一的遥感图像理解自监督学习框架. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–17, 2023.

Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, 和 Xiang Bai. iSAID: 一个用于航空图像实例分割的大规模数据集. 在 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* 中, 第28-37页, 2019年.

王志瑞, 曾轩, 闫志远, 康健, 和 孙贤. Air-polsar-seg: 复杂场景polsar图像地形分割的大规模数据集.

IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15:3830–3841, 2022b.

Xue Li, Guo Zhang, Hao Cui, Shasha Hou, Shunyao Wang, Xin Li, Yujia Chen, Zhijiang Li, 和 Li Zhang. Mcanet: 光学和sar图像联合语义分割的用地分类框架. *International Journal of Applied Earth Observation and Geoinformation*, 106:102638, 2022.

- Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.
- Yi Wang, Nassim Ait Ali Braham, Zhitong Xiong, Chenying Liu, Conrad M Albrecht, and Xiao Xiang Zhu. Ssl4eo-s12: A large-scale multimodal, multitemporal dataset for self-supervised learning in earth observation [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 11(3):98–106, 2023.
- Di Wang, Jing Zhang, Bo Du, Minqiang Xu, Lin Liu, Dacheng Tao, and Liangpei Zhang. Samrs: Scaling-up remote sensing segmentation dataset with segment anything model. *Advances in Neural Information Processing Systems*, 36, 2024b.
- Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021b.
- Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021.
- Xiao Liu, Fei Jin, Shuxiang Wang, Jie Rui, Xibing Zuo, Xiaobing Yang, and Chuanxiang Cheng. Multimodal online knowledge distillation framework for land use/cover classification using full or missing modalities. *IEEE Transactions on Geoscience and Remote Sensing*, 2024b.
- Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.
- Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7151–7160, 2018.
- Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *Proceedings of the European conference on computer vision (ECCV)*, pages 267–283, 2018.
- Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 603–612, 2019.
- Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019.
- Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.
- Wang Zhirui, Kang Yuzhuo, Zeng Xuan, WANG Yuelei, ZHANG Ting, and SUN Xian. Sar-aircraft-1.0: High-resolution sar aircraft detection and recognition dataset. *Journal of Radars*, 12(4):906–922, 2023.
- Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.
- Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.
- Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. Reppoints: Point set representation for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9657–9666, 2019.
- Kun Fu, Jiamei Fu, Zhirui Wang, and Xian Sun. Scattering-keypoint-guided network for oriented ship detection in high-resolution and large-scale sar images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:11162–11178, 2021.
- Hao Chen and Zhenwei Shi. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sensing*, 12(10), 2020. ISSN 2072-4292. doi:10.3390/rs12101662. URL <https://www.mdpi.com/2072-4292/12/10/1662>.
- Hao Chen, Zipeng Qi, and Zhenwei Shi. Remote sensing image change detection with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. 场景理解的统一感知解析。In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018年9月。Yi Wang, Nassim Ait Ali Brah am, Zhitong Xiong, Chenying Liu, Conrad M Albrecht, and Xiao Xiang Zhu. Ssl4eo-s12: 用于地球观测的大型多模态、多时态数据集, 用于无监督学习[软件和数据集]。*IEEE Geoscience and Remote Sensing Magazine*, 11(3):98–106, 2023。Di Wang, Jing Zhang, Bo Du, Minqiang Xu, Lin Liu, Dacheng Tao, and Liangpei Zhang. Samrs: 使用分割一切模型扩展遥感分割数据集的设计。*Advances in Neural Information Processing Systems*, 2024b。En ze Xie, Wenhui Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: 基于变压器的简单高效语义分割设计。*Advances in neural information processing systems*, 34:12077–12090, 2021b。Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: 基于变压器的语义分割。In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 第7262–7272页, 2021。Xiao Liu, Fei Jin, Shuxiang Wang, Jie Rui, Xibing Zuo, Xiaobing Yang, and Chuanxiang Cheng. 基于全模态或缺失模态的多模态在线知识蒸馏框架用于土地利用/覆盖分类。*IEEE Transactions on Geoscience and Remote Sensing*, 2024b 。Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 带有空洞可分离卷积的编码器-解码器用于语义图像分割。In *Proceedings of the European conference on computer vision (ECCV)*, 第8 01–818页, 2018。Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. 基于上下文编码的语义分割。In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 第7151–7160页, 2018。Hengshu ang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: 场景解析的点式空间注意力网络。In *Proceedings of the European conference on computer vision (ECCV)*, 第267–283页, 2018。

Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, 和 Wenyu Liu. Ccnet: 交叉注意力机制在语义分割中的应用. 在 *Proceedings of the IEEE/CVF international conference on computer vision*, 页码 603–6 12, 2019.

Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, 和 Hanqing Lu. 场景分割中的双注意网络. 在 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 第 3146–3154 页, 2019。

曹越, 徐佳瑞, 林Stephen, 魏方云, 和 胡涵. Gcnet: 非局部网络遇见压缩-激励网络及其扩展. 在 *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 页码 0–0, 2019. 王志瑞, 康玉卓, 曾轩, 王越磊, 张婷, 和 孙贤. Sar-aircraft-1.0: 高分辨率sar飞机检测与识别数据集。*Journal of Radars*, 12(4):906–922, 2023. 郑格, 刘松涛, 王锋, 李泽明, 和 孙剑. Yolox: 超越2021年的yolo系列. *arXiv preprint arXiv:2107.08430*, 2021. 任少卿, 何凯明, 金瑞斯, 和 孙剑. Faster r-cnn: 朝实时目标检测迈进的区域建议网络。*IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. 蔡钊伟和Nuno Vasconcelo s. Cascade r-cnn: 深入高质量目标检测. 在 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 页码 6154–6162, 2018. 杨泽, 刘绍辉, 胡涵, 王立威, 和 林Stephen. Reppoints: 点集表示法用于目标检测. 在 *Proceedings of the IEEE/CVF international conference on computer vision*, 页码 9657–9666, 2019. 学富坤, 学富嘉美, 王志瑞, 和 孙贤. 高分辨率和大规模sar图像中的方向船舶检测散射关键点引导网络。*IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:11162–11178, 2021. 陈浩和石振伟. 基于时空注意力的方法和新数据集用于遥感图像变化检测.*Remote Sensing*, 12(10), 2020. ISSN 2072 -4292. doi:10.3390/rs12101662. URL <https://www.mdpi.com/2072-4292/12/10/1662>. 陈浩, 倪子peng, 和 石振伟. 使用变压器的遥感图像变化检测。*IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021.

Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. Oriented r-cnn for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3520–3529, 2021c.

Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. A id: 一种用于空中场景分类性能评估的标准数据集. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7): 3965–3981, 2017. Gong Cheng, Junwei Han, and Xiaoqiang Lu. 远程 sensing 图像场景分类：基准与最新进展. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Jun wei Han. 带方向的 R-CNN 用于目标检测. 在 *Proceedings of the IEEE/CVF international conference on computer vision*, 页码 3520–3529, 2021c.

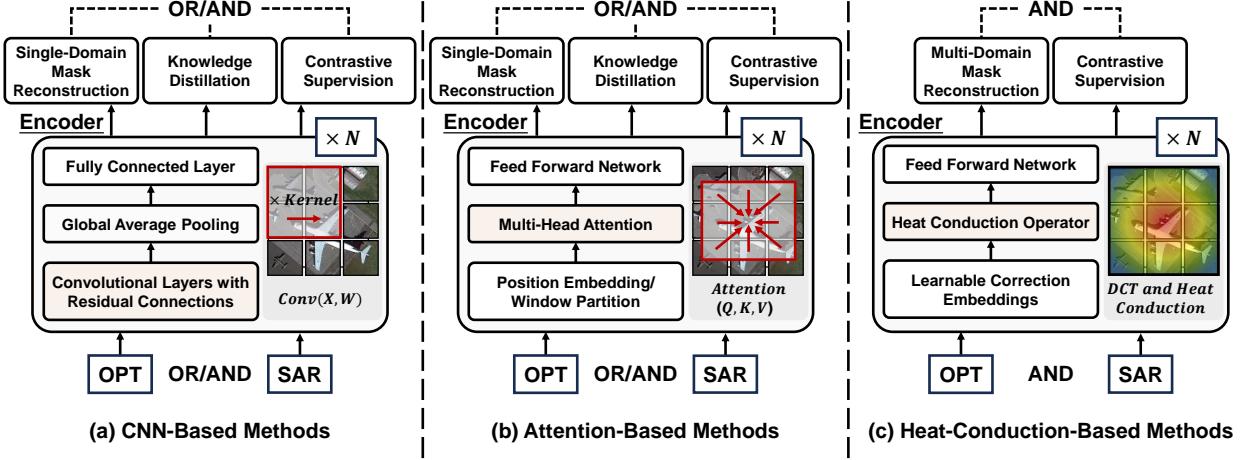


Figure 6: Comparison of the self-supervised training scheme for the heat-conduction-based RSFM with other methods. (a) CNN-Based methods Li et al. [2021], Manas et al. [2021], Mall et al. [2023], Ayush et al. [2021]. (b) Attention-Based methods Cong et al. [2022], Wang et al. [2022a], Tao et al. [2023], Reed et al. [2023], Mendieta et al. [2023], Bastani et al. [2023]. (c) Heat-Conduction-Based method (ours). In our visual encoder, the heat conduction operator is employed to replace the residual blocks in CNN-based networks, and the attention layers in attention-based networks. For optical (OPT) and SAR inputs, the dual constraints of multi-domain mask reconstruction and distance metrics for multi-modal feature representations provide self-supervised signals during the pre-training process. This approach transforms the visual semantic propagation into a process of thermal diffusion within a thermal space, guided by the scene and object characteristics, dynamically extracting global information across the entire image.

A. Comparison details among RSFMs with different backbone networks

Previous research on RSFMs primarily utilized existing visual encoders to extract deep features, integrating various self-supervised learning strategies with decoder structures, and pre-training on large-scale RS datasets. Visual encoders, as the core components of these models, are generally divided into two categories in recent research: 1) CNN-based methods Li et al. [2021], Manas et al. [2021], Mall et al. [2023], Ayush et al. [2021], as shown in fig. 6 (a). These models typically adopt the ResNet18/50 framework He et al. [2016], with the residual module serving as the key learning structure. These approaches extract rich information from RS data through pixel masking reconstruction, expert geographical knowledge supervision or contrastive learning signals. 2) Attention-based methods, such as Cong et al. [2022], Wang et al. [2022a], Tao et al. [2023], Reed et al. [2023], Mendieta et al. [2023] and Bastani et al. [2023], as illustrated in fig. 6 (b). These models primarily utilize the ViT Dosovitskiy et al. [2020] and Swin Transformers Liu et al. [2021] as visual encoders, where the fundamental modules rely on attention mechanisms Vaswani et al. [2017] and feed forward networks (FFNs) to model global dependencies. Pre-training is typically conducted through masked reconstruction, knowledge distillation or contrastive signals to enhance the robustness of the model representations.

In summary, current RSFMs typically employ CNN-based or attention-based methods as visual encoders, innovating in learning and training strategies to enhance model performance. As shown in fig. 6 (c), RS-vHeat employs a heat-conduction-based visual encoder, with the heat conduction operator serving as the core computational module. During self-supervised learning, it applies frequency-domain and spatial-domain masking reconstruction constraints, along with an additional contrastive loss, which differentiates it significantly from existing RSFMs.

B. Preliminary of heat conduction

Inspired by the physical principle of heat conduction, vHeat Wang et al. [2024a] considers a region as a two-dimensional region $D \in \mathbb{R}^2$. Then, for each point (x, y) in the region, its temperature is $u(x, y, t)$ at time t , and the initial condition is $t = 0$. The heat conduction propagation on this region can be expressed:

$$\frac{\partial u}{\partial t} = k \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) \quad (9)$$

where k represents the thermal diffusivity. We denote the Fourier Transform and its inverse using the symbols \mathcal{F} and \mathcal{F}^{-1} , respectively. After taking the Fourier Transform on both sides of the equals sign in eq. (9), we formulate the

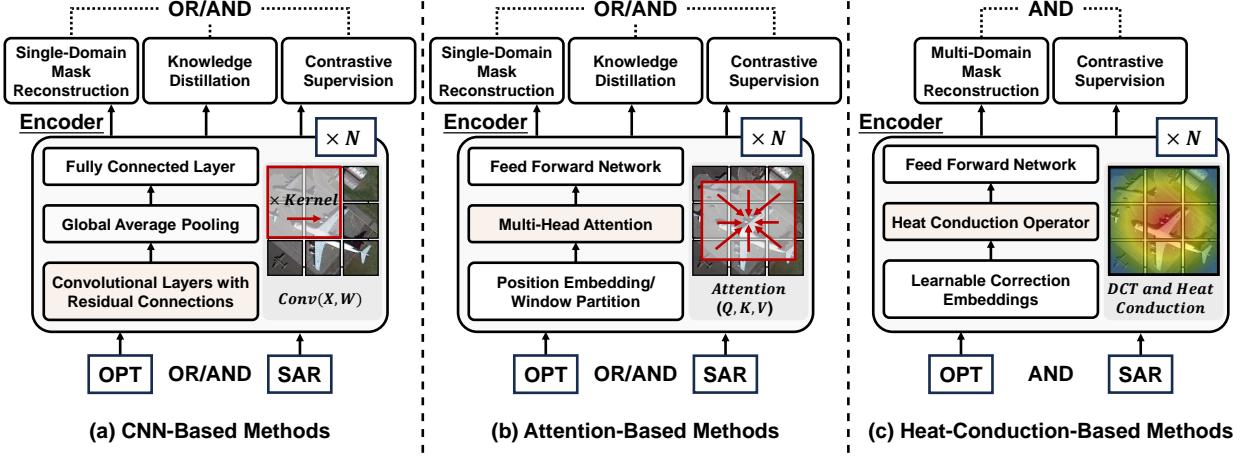


图 6: 基于热传导的 RSFM 与其它方法的自监督训练方案比较。(a) 基于 CNN 的方法: Li 等人 [2021], Manas 等人 [2021], Mall 等人 [2023], Ayush 等人 [2021]。(b) 基于注意力的方法: Cong 等人 [2022], Wang 等人 [2022a], Tao 等人 [2023], Reed 等人 [2023], Mendieta 等人 [2023], Bastani 等人 [2023]。(c) 基于热传导的方法 (我们自己的方法)。在我们的视觉编码器中, 使用热传导算子来替换 CNN 基础网络中的残差块以及注意力网络中的注意力层。对于光学 (OPT) 和 SAR 输入, 多域掩码重构的双重约束和多模态特征表示的距离度量在预训练过程中提供了自监督信号。这种方法将视觉语义传播转化为在热空间内的热扩散过程, 由场景和对象特性引导, 动态提取整幅图像中的全局信息。

A. 不同骨干网络的RSFMs比较细节

Previous research on RSFMs 主要利用现有的视觉编码器提取深度特征, 结合各种自监督学习策略与解码器结构, 并在大规模RS数据集上进行预训练。视觉编码器作为这些模型的核心组件, 在最近的研究中通常被分为两类: 1) 基于CNN的方法, 如Li等人[2021]、Manas等人[2021]、Mall等人[2023]、Ayush等人[2021], 如图6(a)所示。这些模型通常采用He等人[2016]提出的ResNet18/50框架, 其中残差模块是关键的学习结构。这些方法通过像素掩码重建、专家地理知识监督或对比学习信号从RS数据中提取丰富的信息。2) 基于注意力的方法, 如Cong等人[2022]、Wang等人[2022a]、Tao等人[2023]、Reed等人[2023]、Mendieta等人[2023]和Bastani等人[2023], 如图6(b)所示。这些模型主要使用Dosovitskiy等人[2020]提出的ViT和Liu等人[2021]提出的Swin Transformers作为视觉编码器, 其中基本模块依赖于注意机制Vaswani等人[2017]和前馈网络(FFNs)来建模全局依赖关系。预训练通常通过掩码重建、知识蒸馏或对比学习信号来增强模型表示的鲁棒性。

总结来说, 当前的RSFMs通常采用基于CNN或注意力机制的方法作为视觉编码器, 并在学习和训练策略上进行创新以提升模型性能。如图6(c)所示, RS-vHeat采用基于热传导的视觉编码器, 热传导算子作为核心计算模块。在自监督学习过程中, 它应用了频域和空域的掩码重建约束, 并附加了一个对比损失, 这与现有的RSFMs有显著区别。

B. 热传导的初步

受热传导物理原理的启发, Wang et al. [2024a] 将一个区域视为二维区域 $D \in \mathbb{R}^2$ 。然后, 对于该区域中的每个点 (x, y) , 其在时间 t 时的温度为 $u(x, y, t)$, 初始条件为 $t = 0$ 。该区域中的热传导传播可以表示为:

$$\frac{\partial u}{\partial t} = k \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) \quad (9)$$

其中 k 表示热扩散率。我们分别用符号 \mathcal{F} 和 \mathcal{F}^{-1} 表示傅里叶变换及其逆变换。在等号两边对 eq. (9) 进行傅里叶变换后, 我们得到

calculation of physical heat equation as:

$$\mathcal{F}\left(\frac{\partial u}{\partial t}\right) = k\mathcal{F}\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) \quad (10)$$

We represent the result of the Fourier transform of $u(x, y, t)$ as follows:

$$\tilde{u}(\omega_x, \omega_y, t) := \mathcal{F}(u(x, y, t)) \quad (11)$$

The left and right of eq. (10) can be reformulated as

$$\mathcal{F}\left(\frac{\partial u}{\partial t}\right) = \frac{\partial \tilde{u}(\omega_x, \omega_y, t)}{\partial t} \quad (12)$$

$$\mathcal{F}\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) = -(\omega_x^2 + \omega_y^2)\tilde{u}(\omega_x, \omega_y, t) \quad (13)$$

Furthermore, the eq. (10) is expressed as an ordinary differential equation in the frequency domain:

$$\frac{d\tilde{u}(\omega_x, \omega_y, t)}{dt} = -k(\omega_x^2 + \omega_y^2)\tilde{u}(\omega_x, \omega_y, t) \quad (14)$$

To solve $\tilde{u}(\omega_x, \omega_y, t)$ in eq. (14), we use $\tilde{f}(\omega_x, \omega_y)$ to represent the Fourier Transform of $f(x, y)$, and we can get the following result under the initial condition of $\tilde{u}(\omega_x, \omega_y, t)|_{t=0}$:

$$\tilde{u}(\omega_x, \omega_y, t) = \tilde{f}(\omega_x, \omega_y)e^{-k(\omega_x^2 + \omega_y^2)t} \quad (15)$$

Finally, the values in the frequency domain are converted back to the space domain by inverse Fourier Transform, and we get the general solution of heat equation in the spatial domain expressed as follows:

$$u(x, y, t) = \mathcal{F}^{-1}(\tilde{f}(\omega_x, \omega_y)e^{-k(\omega_x^2 + \omega_y^2)t}) = \frac{1}{4\pi^2} \int_{\tilde{D}} \tilde{f}(\omega_x, \omega_y)e^{-k(\omega_x^2 + \omega_y^2)t} e^{i(\omega_x x + \omega_y y)} d\omega_x d\omega_y \quad (16)$$

C. Implementation details of the masking strategy

Given the multi-modal input (optical and SAR), denoted as $I(x, y, c) = \{I_o, I_s\}, I_o \in \mathbb{R}^{H \times W \times 3}, I_s \in \mathbb{R}^{H \times W \times 1}$, the process begins by applying the DCT along each image dimension $c = 1, \dots, C$, extracting 2D planes from the spatial domain $I(x, y)$ and converting them into its frequency representation $\tilde{I}(u, v)$. This transformation concentrates low-frequency information in the top-left corner of the frequency spectrum:

$$\tilde{I}(u, v) = \frac{2}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} I(x, y) \cos \frac{(2x+1)u\pi}{2M} \cos \frac{(2y+1)v\pi}{2N} \quad (17)$$

where M and N denote the width and height of the input image, respectively.

To address signals across different frequency ranges, we apply a sector mask to the transformed image. Centered at the top-left, this mask separates the image into distinct high-frequency $\tilde{I}^{high}(u, v)$ and low-frequency $\tilde{I}^{low}(u, v)$ regions:

$$\tilde{I}^{low}(u, v), \tilde{I}^{high}(u, v) = \tilde{M} \odot \tilde{I}(u, v) \quad (18)$$

The binary mask \tilde{M} , sized $(M \times N)$, is applied to each dimension c using the operator \odot . Each element of \tilde{M} takes a value of either 0 or 1.

After applying the mask, we perform the IDCT to convert the processed frequency representation back to its spatial representation along each dimension:

$$I^{low}(x, y) = \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} \frac{2}{\sqrt{MN}} \tilde{I}^{low}(u, v) \cos \frac{(2x+1)u\pi}{2M} \cos \frac{(2y+1)v\pi}{2N} \quad (19)$$

$$I^{high}(x, y) = \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} \frac{2}{\sqrt{MN}} \tilde{I}^{high}(u, v) \cos \frac{(2x+1)u\pi}{2M} \cos \frac{(2y+1)v\pi}{2N} \quad (20)$$

Where $I^{low}(x, y)$ and $I^{high}(x, y)$ denote the low- and high-frequency representation that are converted back to their spatial domain after applying the mask. The results are then concatenated to restore the original dimensionality.

计算物理热方程为：

$$\mathcal{F}\left(\frac{\partial u}{\partial t}\right) = k\mathcal{F}\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) \quad (10)$$

我们表示 $u(x, y, t)$ 的傅里叶变换结果如下：

$$\tilde{u}(\omega_x, \omega_y, t) := \mathcal{F}(u(x, y, t)) \quad (11)$$

等号 (10) 的左右可以重新公式化为

$$\mathcal{F}\left(\frac{\partial u}{\partial t}\right) = \frac{\partial \tilde{u}(\omega_x, \omega_y, t)}{\partial t} \quad (12)$$

$$\mathcal{F}\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) = -(\omega_x^2 + \omega_y^2)\tilde{u}(\omega_x, \omega_y, t) \quad (13)$$

此外，eq. (10) 在频域中表示为一个常微分方程：

$$\frac{d\tilde{u}(\omega_x, \omega_y, t)}{dt} = -k(\omega_x^2 + \omega_y^2)\tilde{u}(\omega_x, \omega_y, t) \quad (14)$$

为了解决 eq. (14) 中的 $\tilde{u}(\omega_x, \omega_y, t)$ ，我们使用 $\tilde{f}(\omega_x, \omega_y)$ 来表示 $f(x, y)$ 的傅里叶变换，并在初始条件为 $\tilde{u}(\omega_x, \omega_y, t)|_{t=0}$ 的情况下可以得到以下结果：

$$\tilde{u}(\omega_x, \omega_y, t) = \tilde{f}(\omega_x, \omega_y)e^{-k(\omega_x^2 + \omega_y^2)t} \quad (15)$$

最后，通过逆傅里叶变换将频域中的值转换回空间域，并得到如下的热方程在空间域中的通用解：

$$u(x, y, t) = \mathcal{F}^{-1}(\tilde{f}(\omega_x, \omega_y)e^{-k(\omega_x^2 + \omega_y^2)t}) = \frac{1}{4\pi^2} \int_{\bar{D}} \tilde{f}(\omega_x, \omega_y)e^{-k(\omega_x^2 + \omega_y^2)t} e^{i(\omega_x x + \omega_y y)} d\omega_x d\omega_y \quad (16)$$

C. 遮罩策略的实现细节

给定多模态输入（光学和SAR），表示为 $I(x, y, c) = \{I_o, I_s\}$, $I_o \in \mathbb{R}^{H \times W \times 3}$, $I_s \in \mathbb{R}^{H \times W \times 1}$ ，过程从沿每个图像维度 c = 应用DCT开始，从中提取空间域的2D平面 $I(x, y)$ 并将其转换为其频率表示 $\tilde{I}(u, v)$ 。这种变换将低频信息集中在频率谱的左上角：

$$\tilde{I}(u, v) = \frac{2}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} I(x, y) \cos \frac{(2x+1)u\pi}{2M} \cos \frac{(2y+1)v\pi}{2N} \quad (17)$$

其中 M 和 N 分别表示输入图像的宽度和高度。

为了应对不同频率范围的信号，我们在变换后的图像上应用一个扇区掩码。该掩码以左上角为中心，将图像分离成独立的高频 $\tilde{I}^{high}(u, v)$ 和低频 $\tilde{I}^{low}(u, v)$ 区域：

$$\tilde{I}^{low}(u, v), \tilde{I}^{high}(u, v) = \tilde{M} \odot \tilde{I}(u, v) \quad (18)$$

二进制掩码 \tilde{M} ，大小为 $(M \times N)$ ，使用操作符 \odot 应用于每个维度 c 。 \tilde{M} 中的每个元素的值为 0 或 1。

应用掩码后，我们执行IDCT，将其处理后的频率表示转换回每个维度的空间表示：

$$I^{low}(x, y) = \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} \frac{2}{\sqrt{MN}} \tilde{I}^{low}(u, v) \cos \frac{(2x+1)u\pi}{2M} \cos \frac{(2y+1)v\pi}{2N} \quad (19)$$

$$I^{high}(x, y) = \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} \frac{2}{\sqrt{MN}} \tilde{I}^{high}(u, v) \cos \frac{(2x+1)u\pi}{2M} \cos \frac{(2y+1)v\pi}{2N} \quad (20)$$

Where $I^{low}(x, y)$ 和 $I^{high}(x, y)$ 表示低频和高频表示，经过掩码处理后转换回空域。然后将结果连接起来以恢复原始维度。

D. Configuration and visualization results of fownstream task datasets

RS-vHeat is trained on 10 datasets across 4 downstream tasks. In this section, we provide detailed information about the datasets and experimental configurations.

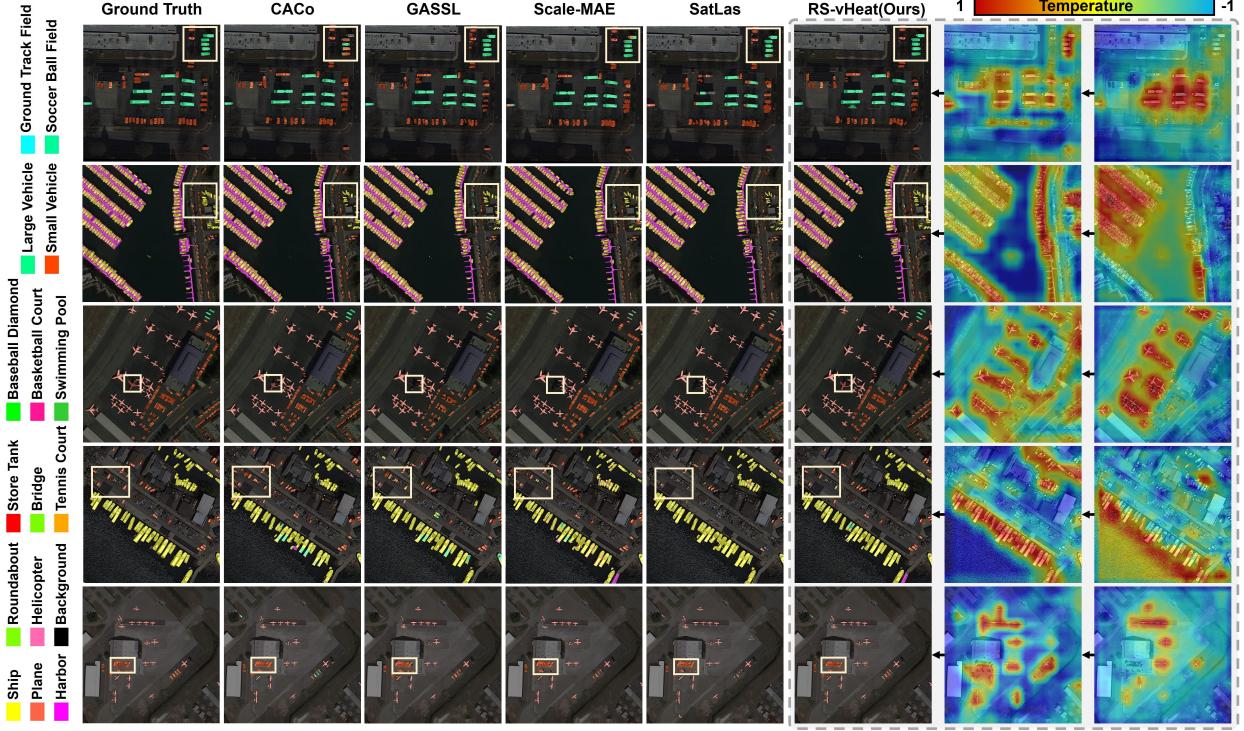


Figure 7: The qualitative results of RS-vHeat and several representative RSFMs on the iSAID dataset. Each column from left to right represents: ground truth, CACo (ResNet-18), GASSL (ResNet-50), Scale-MAE (ViT-L), Satlas (Swin-B), and the results from our model, RS-vHeat. The last two columns on the right visualize the output variations of RS-vHeat across the final two stages.

Table 10: Comparison of AP₅₀ for each category, mAP₅₀ and mAP₇₅ on SAR-AIRcraft-1.0 with other specialized models.

Method	Publication	A330	A320/A321	A220	ARJ21	Boeing737	Boeing787	Other	mAP ₅₀ ↑	mAP ₇₅ ↑
Faster R-CNN Ren et al. [2016]	TPAMI'2016	85.0	97.2	78.5	74.0	55.1	72.9	70.1	76.1	62.2
Cascade R-CNN Cai and Vasconcelos [2018]	CVPR'2018	87.4	97.5	74.0	78.0	54.5	68.3	69.1	75.7	58.9
RepPoints Yang et al. [2019]	ICCV'2019	89.8	97.9	71.4	73.0	55.7	51.8	68.4	72.6	53.3
SKG-Net Fu et al. [2021]	JSTARS'2021	79.3	78.2	66.4	65.0	65.1	69.6	71.4	70.7	46.4
SA-Net Zhirui et al. [2023]	RADARS'2023	88.6	94.3	80.3	78.6	59.7	70.8	71.3	77.7	62.8
RS-vHeat (Ours)	-	98.4	97.9	81.1	89.3	82.0	79.8	81.1	87.1	67.4

D.1. Single- and multi-modal semantic segmentation

We utilize RS-vHeat as the visual encoder and implemented UPerNet Xiao et al. [2018] with cross-entropy loss for the output head. Additionally, we employ the AdamW optimizer with a learning rate of 6e-5 and conduct a warm-up of 1500 iterations.

Dataset. We evaluated our model on three single-modal datasets and one multi-modal dataset:

- 1) The Potsdam dataset Sherrah [2016] comprises 38 images. This dataset is annotated with six classes, each having a resolution of 6000 × 6000 pixels. The input resolution is set to 512 pixels.
- 2) The iSAID dataset Waqas Zamir et al. [2019] comprises 2,806 images with varying resolutions, primarily focusing on urban environments. The dataset includes annotations for 15 different categories and we utilize an image size of 896 pixels as the input for the model.

D. 下游任务数据集的配置和可视化结果

RS-vHeat 在 4 个下游任务的 10 个数据集中进行训练。在本节中，我们提供了关于数据集和实验配置的详细信息。

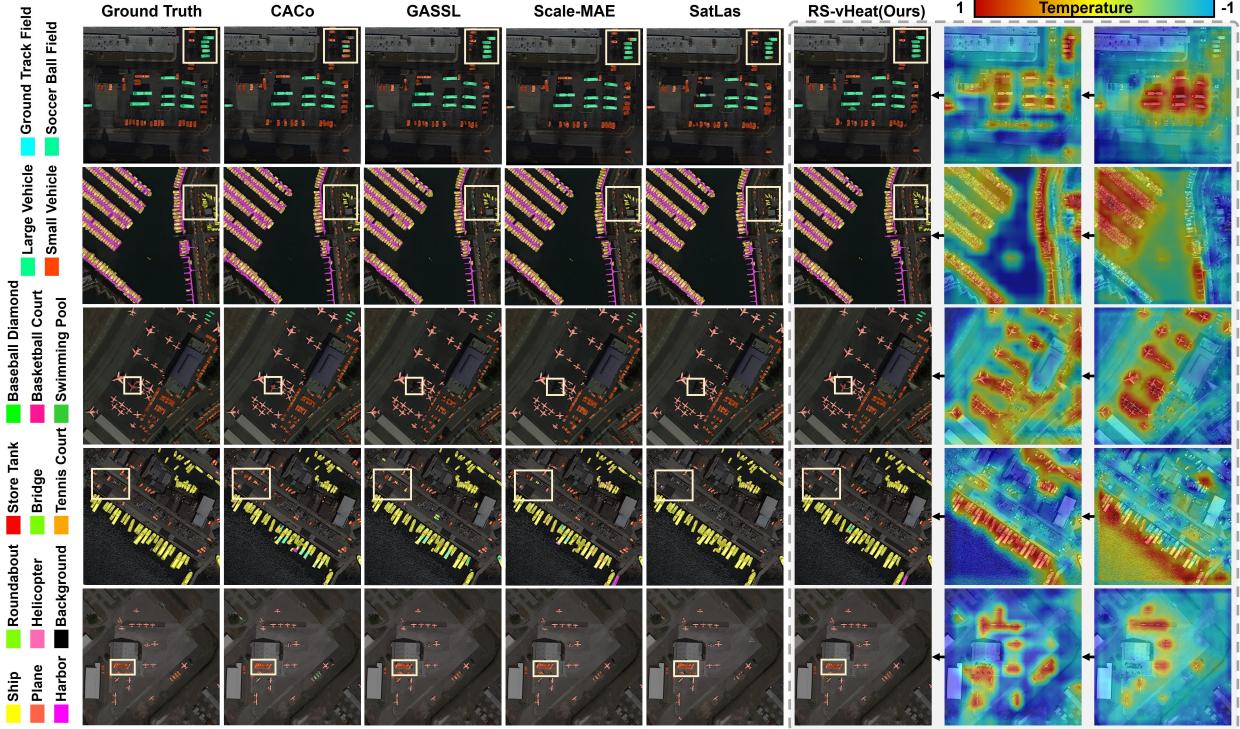


Figure 7: RS-vHeat 和几个代表性的 RSFM 在 iSAID 数据集上的定性结果。从左到右每一列分别表示：ground truth, CACo (ResNet-18), GASSL (ResNet-50), Scale-MAE (ViT-L), Satlas (Swin-B), 以及我们模型 RS-vHeat 的结果。最右边的两列可视化了 RS-vHeat 在最后两个阶段的输出变化。

表10: SAR-AIRcraft-1.0上每类别AP₅₀的比较, 以及其他专门模型的mAP₅₀和mAP₇₅

Method	Publication	A330	A320/A321	A220	ARJ21	Boeing737	Boeing787	Other	mAP ₅₀ ↑	mAP ₇₅ ↑
Faster R-CNN Ren et al. [2016]	TPAMI'2016	85.0	97.2	78.5	74.0	55.1	72.9	70.1	76.1	62.2
Cascade R-CNN Cai and Vasconcelos [2018]	CVPR'2018	87.4	97.5	74.0	54.5	68.3	69.1	75.7	58.9	
RepPoints Yang et al. [2019]	ICCV'2019	89.8	97.9	71.4	73.0	55.7	51.8	68.4	72.6	53.3
SKG-Net Fu et al. [2021]	JSTARS'2021	79.3	78.2	66.4	65.0	65.1	69.6	71.4	70.7	46.4
SA-Net Zhirui et al. [2023]	RADARS'2023	88.6	94.3	80.3	78.6	59.7	70.8	71.3	77.7	62.8
RS-vHeat (Ours)	-	98.4	97.9	81.1	89.3	82.0	79.8	81.1	87.1	67.4

D.1. 单模态和多模态语义分割

我们使用 RS-vHeat 作为视觉编码器，并使用 Xiao 等人 [2018] 实现的 UPerNet 与交叉熵损失函数作为输出头。此外，我们采用 AdamW 优化器，学习率为 6e-5，并进行 1500 次迭代的 warm-up。

数据集。我们评估了我们的模型在三个单模态数据集和一个跨模态数据集上：

1) 帕茨坦数据集 Sherrah [2016] 包含 38 张图片。该数据集用六类进行标注，每类的分辨率为 6000 × 6000 像素。输入分辨率为 512 像素。

2) iSAID数据集Waqas Zamir等[2019]包含2,806张具有不同分辨率的图像, 主要关注城市环境。该数据集包括 15 个不同类别的注释, 并且我们使用 896 像素的图像作为模型的输入。

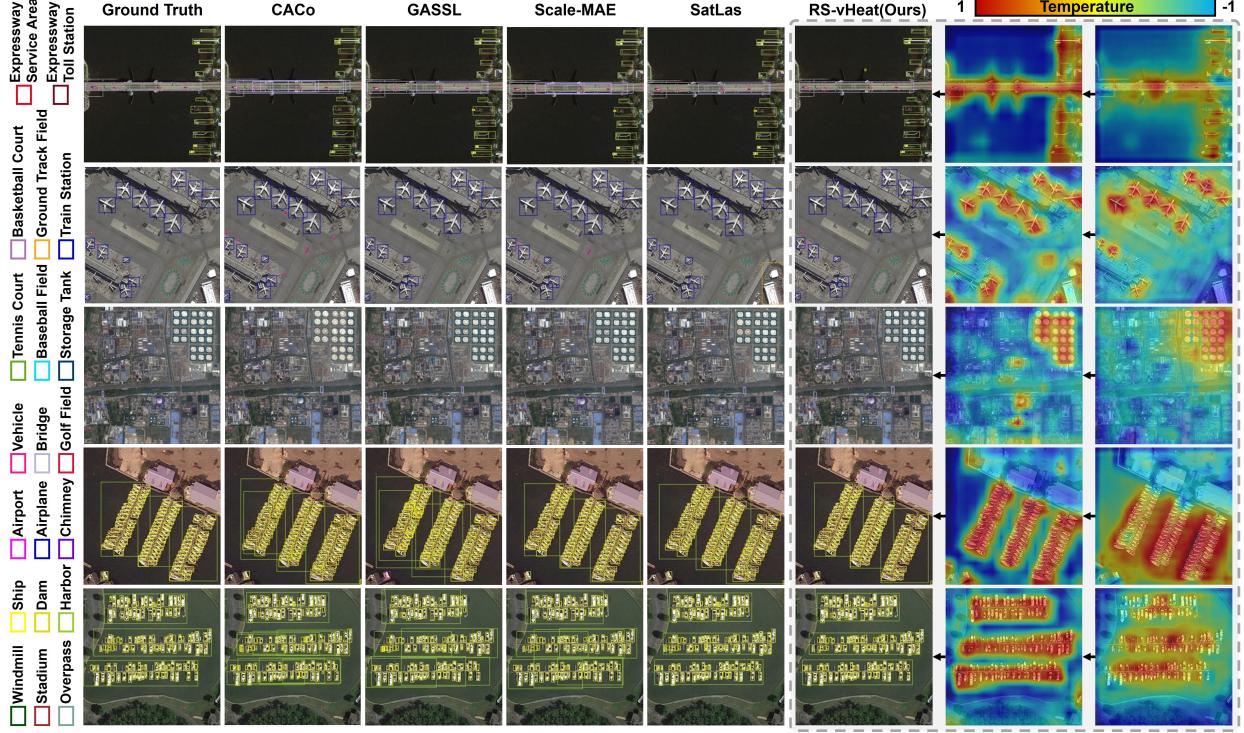


Figure 8: The qualitative results of RS-vHeat and several representative RSFMs on the DIOR dataset. Each column from left to right represents: ground truth, CACo (ResNet-18), GASSL (ResNet-50), Scale-MAE (ViT-L), Satlas (Swin-B), and the results from our model, RS-vHeat. The last two columns on the right visualize the output variations of RS-vHeat across the final two stages.

3) The Air-PoSAR-Seg dataset Wang et al. [2022b] focuses on polarimetric SAR images. It offers a region measuring 9082×9805 pixels and includes 2,000 image patches, each sized 512×512 . The dataset features pixel-wise annotations covering six categories. We adopt a size of 512 pixels for the image input.

4) The WHU-OPT-SAR dataset Li et al. [2022] is a multi-modal segmentation dataset with a resolution of 5 meters. It includes optical and SAR data from the same region, categorized into seven classes. Each image has a size of 5556×3704 pixels. We uniformly cropped the multi-modal images to a pixel size of 256 for model input.

Metric. Following the configurations of RingMo Sun et al. [2022b] and SkySense Guo et al. [2024], we evaluate the mean Intersection over Union (mIoU) on the iSAID dataset and test the mean F1 score (mF1) on the Potsdam dataset. For the AIR-PoSAR-Seg dataset, we use three metrics: mIoU, Overall Accuracy (OA) and Average Accuracy (AA). We assess OA and User's Accuracy on the WHU-OPT-SAR dataset following the setup outlined in the corresponding paper.

Additional Results. The fig. 7 displays the process visualizations and prediction results for the iSAID dataset, which display that the heat-conduction-based backbone exhibits adaptive characteristics when capturing features across different layers.

D.2. Object Detection

We conduct coarse- and fine-grained experiments on optical and SAR datasets to demonstrate the robustness of RS-vHeat. In the horizontal bounding boxes (HBB) task, we employ SGD as the optimizer, with a base learning rate set to 0.01. A warm-up phase of 3 epochs is conducted. YOLOX Ge et al. [2021] is used as the output head, and experiments are conducted using cross-entropy loss and IoU loss. In the oriented bounding box (OBB) task, we adjust the base learning rate to 1e-4. The warm-up phase consists of 500 iterations. Oriented RCNN Xie et al. [2021c] is used as the output head, applying cross-entropy and Smooth \mathcal{L}_1 loss.

Dataset. Our model is tested on three challenging object detection datasets:

- 1) FAIR1M Sun et al. [2022a] is an optical fine-grained dataset with objects annotated using OBB, encompassing five major categories, further divided into 37 subcategories. The dataset contains over 40,000 images. Following the official

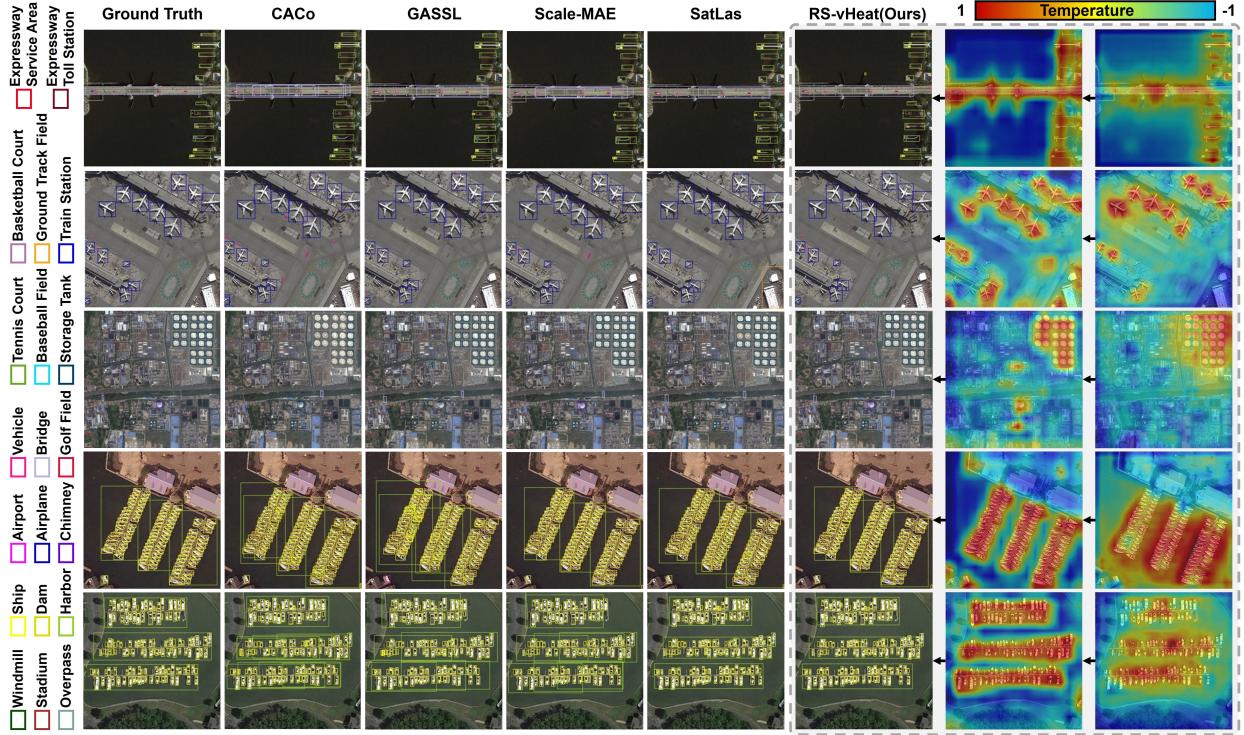


图 8：RS-vHeat 和几个代表性的 RSFM 在 DIOR 数据集上的定性结果。从左到右的每一列代表：真实值、CACo (ResNet-18)、GASSL (ResNet-50)、Scale-MAE (ViT-L)、Satlas (Swin-B) 以及我们模型 RS-vHeat 的结果。最右边的两列可视化了 RS-vHeat 在最终两个阶段的输出变化。

3) Air-PoSAR-Seg 数据集 [王等, 2022b] 专注于极化 SAR 图像。它提供了一个大小为 9082×9805 像素的区域，并包括 2000 个图像块，每个块的大小为 512×512 。该数据集包含六类像素级注释。我们采用 512 像素的大小作为图像输入。

4) WHU-OPT-SAR 数据集 [Li 等人, 2022] 是一个分辨率高达 5 米的多模态分割数据集。它包括同一区域的光学和 SAR 数据，并被分类为七个类别。每张图像的大小为 5556×3704 像素。我们均匀裁剪了多模态图像，使其像素大小为 256 以供模型输入。

度量。遵循 RingMo Sun 等人 [2022b] 和 SkySense Guo 等人 [2024] 的配置，我们在 iSAID 数据集上评估平均交并比 (mIoU)，在 Potsdam 数据集上测试平均 F1 分数 (mF1)。对于 AIR-PoSAR-Seg 数据集，我们使用三个指标：mIoU、总体准确率 (OA) 和平均准确率 (AA)。我们在 WHU-OPT-SAR 数据集上评估 OA 和用户准确率，遵循相应论文中概述的设置。

附加结果。图 7 显示了 iSAID 数据集的过程可视化和预测结果，这些结果表明基于热传导的骨干网络在跨不同层捕获特征时表现出适应性。

D.2. 物体检测

我们在光学和SAR数据集上进行粗粒度和细粒度实验，以展示RS-vHeat的鲁棒性。在水平边界框 (HBB) 任务中，我们使用SGD作为优化器，基学习率设置为0.01。进行3个epoch的预热阶段。YOLOX Ge et al. [2021] 用作输出头，并使用交叉熵损失和IoU损失进行实验。在定向边界框 (OBB) 任务中，我们将基学习率调整为 $1e-4$ 。预热阶段包括500个迭代。定向RCNN Xie et al. [2021c] 用作输出头，并应用交叉熵和Smooth \mathcal{L}_1 损失。

数据集。我们的模型在三个具有挑战性的目标检测数据集上进行了测试：

1) FAIR1M Sun 等人 [2022a] 是一个光学细粒度数据集，使用OBB 对对象进行标注，涵盖五大类，进一步细分为 37 个子类别。该数据集包含超过 40,000 张图像。遵循官方

split, we ultimately submitted the test results to the website to obtain accuracy measurements. We utilize an image size of 512 pixels as the input for the model.

2) SAR-AIRcraft-1.0 Zhirui et al. [2023] is a HBB fine-grained SAR aircraft object detection dataset designed for challenging scenarios, totaling 4,368 images. It encompasses seven fine-grained categories. We adopt a size of 640 pixels for the image input.

3) DIOR Li et al. [2020a] is an optical dataset that includes 20 categories. It comprises a total of 23,463 images and provides HBB annotations. We utilize an image size of 800 pixels as the input for the model.

Metric. On the FAIR1M and DIOR dataset, we evaluate the mAP (Mean Average Precision). For the SAR-AIRcraft-1.0 dataset, we evaluate the AP₅₀ for each category, mAP₅₀ and mAP₇₅. mAP₅₀ and mAP₇₅ represent the mAP at IoU thresholds of 0.5 and 0.75, respectively, with category-specific precision calculated at an IoU threshold of 0.5.

Additional Results. The visualization results of the DIOR dataset are shown in fig. 8. From the feature extraction process and results, RS-vHeat outperforms other RSFMs in terms of extracting dense RS objects. Additionally, we further refine the RS-vHeat extraction results for each class of the SAR-AIRcraft-1.0 dataset in table 10, highlighting its enhanced capability in recognizing various aircraft types in SAR scenarios compared to specialized object detection models.

D.3. Change Detection

We employ RS-vHeat as the visual encoder, accommodating images before and after transformation. AdamW optimizer is used with a base learning rate of 0.002 and we train for 200 epochs. The BIT architecture Chen et al. [2021] is utilized for subsequent image change analysis, with cross-entropy loss applied for the experiments.

Dataset. We use the LEVIR-CD dataset to train and test:

1) The LEVIR-CD dataset Chen and Shi [2020] consists of 637 image patch pairs obtained from Google Earth. Each patch has a size of 1024 × 1024 pixels. The dataset primary focus is on building-related changes, such as the emergence of new structures and the decline of existing ones. We utilize an image size of 256 pixels as the input.

Metric. We use F1-score to evaluate change detection performance. F1-score is the harmonic mean of precision and recall, providing a balanced measure of performance.

D.4. Image Classification

We extend our model by attaching a classification head designed to handle the classification task and employ cross-entropy loss for computation. We utilize AdamW as the optimizer with a learning rate of 5e-4, training for 300 epochs.

Dataset. We validate our model on two benchmark datasets as described below.

1) The Aerial Image Dataset (AID) Xia et al. [2017] consists of 30 categories, with each category containing approximately 220 to 420 images sized at 600 × 600 pixels, totaling 10,000 images.

2) The NWPU-RESISC45 dataset Cheng et al. [2017] is a RS image dataset comprising 45 categories, with a total of 31,500 images distributed across these categories. Each category consists of 700 images.

Metric. We use OA to evaluate classification performance. We follows standard practices in the field Guo et al. [2024], using 20% and 50% of the AID dataset as training sets, and 10% and 20% of the NWPU-RESISC45 dataset as training sets.

分割后，我们最终将测试结果提交到网站以获取准确度测量。我们使用512像素的图像作为模型的输入。

2) SAR-AIRcraft-1.0 翱瑞等 [2023] 是一个旨在应对具有挑战性场景的HBB细粒度SAR飞机对象检测数据集，总计包含4,368张图像。它涵盖了七个细粒度类别。我们采用640像素的图像输入。

3) DIOR Li等 [2020a] 是一个光学数据集，包括20个类别。它总共包含23,463张图像，并提供了HBB注释。我们使用800像素的图像作为模型的输入。

指标。在FAIR1M和DIOR数据集上，我们评估mAP（平均精确度）。对于SAR-AIRcraft-1.0数据集，我们为每个类别评估AP₅₀，mAP₅₀和mAP₇₅。mAP₅₀和mAP₇₅分别表示在IoU阈值为0.5和0.75时的mAP，且在IoU阈值为0.5时计算每个类别的精确度。

附加结果。DIOR 数据集的可视化结果如图 8 所示。从特征提取过程和结果来看，RS-vHeat 在提取密集 RS 对象方面优于其他 RSFMs。此外，我们还在表 10 中进一步细化了 SAR-AIRcraft-1.0 数据集每类 RS-vHeat 提取结果，突显了其在 SAR 场景中识别各种飞机类型方面增强的能力，优于专门的对象检测模型。

D.3. 变化检测

我们使用RS-vHeat作为视觉编码器，处理变换前后的图像。使用AdamW优化器，基础学习率为0.002，并训练200个周期。利用Chen等[2021]提出的BIT架构进行后续的图像变化分析，在实验中应用交叉熵损失。

数据集。我们使用LEVIR-CD数据集进行训练和测试：

1) LEVIR-CD数据集[陈和施, 2020]包含637个来自Google地球的图像块对。每个块的大小为1024 × 1024像素。该数据集的主要重点是建筑相关的变化，如新结构的出现和现有结构的衰退。我们使用256像素的图像作为输入。

度量。我们使用F1-score来评估变化检测性能。F1-score是精确率和召回率的调和平均值，提供了一个性能的平衡度量。

D.4. 图像分类

我们通过附加一个设计用于处理分类任务的分类头来扩展我们的模型，并使用交叉熵损失进行计算。我们使用 AdamW 作为优化器，学习率为 5e-4，训练 300 个epochs。

数据集。我们验证我们的模型在如上所述的两个基准数据集上 w.

1) 航空图像数据集 (AID) Xia等 [2017] 包含30个类别，每个类别包含约220到420张大小为600×600像素的图像，总共10,000张图像。

2) 北京理工大学-RESISC45 数据集 Cheng 等人 [2017] 是一个包含 45 个类别的 RS 图像数据集，总共包含 31,500 张图像，分布在这些类别中。每个类别包含 700 张图像。

指标。我们使用OA评估分类性能。我们遵循领域内的标准做法 Guo 等人 [2024]，使用AID数据集的20%和50%作为训练集，以及NWPU-RESISC45数据集的10%和20%作为训练集。