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 [31], [40], [39], [2], as shown in Fig. 6 (a). These models typically adopt the ResNet18/50 framework [25], 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 [15], [56], [53], [44], [41] and [3], as illustrated in Fig. 6 (b). These models primarily utilize the ViT [17] and Swin Transformers [35] as visual encoders, where the fundamental modules rely on attention mechanisms [55] 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 [62] 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:

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

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 calculation of physical heat equation as:

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

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

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

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

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

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

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

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

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 $\tilde{u}(\omega_x, \omega_y, t)|_{t=0}$:

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

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:

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

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 $c = 1, \ldots, C$, 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:

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

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

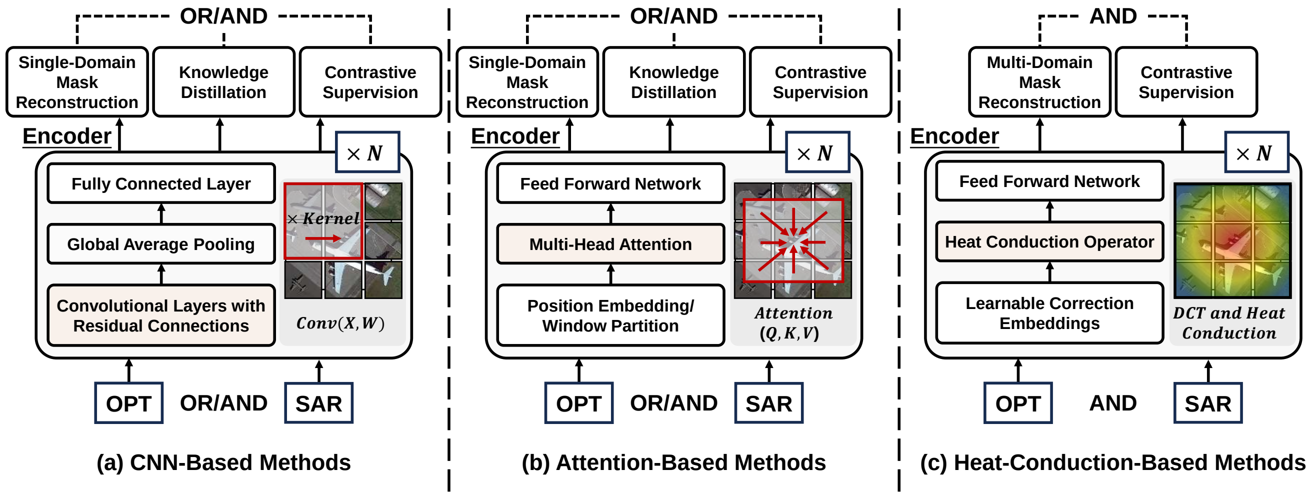

Figure 6. Comparison of the self-supervised training scheme for the heat-conduction-based RSFM with other methods. (a) CNN-Based methods [31], [40], [39], [2]. (b) Attention-Based methods [15], [56], [53], [44], [41], [3]. (c) Heat-Conduction-Based method (ours). In our visual encoder, the heat conduction 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.

high-frequency $\tilde{I}^{high}(u,v)$ and low-frequency $\tilde{I}^{low}(u,v)$ regions:

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

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:

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

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.

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.

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

We utilize RS-vHeat as the visual encoder and implemented UPerNet [65] 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 [47] 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 [63] 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.

3) The Air-PolSAR-Seg dataset [61] 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 [32] 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

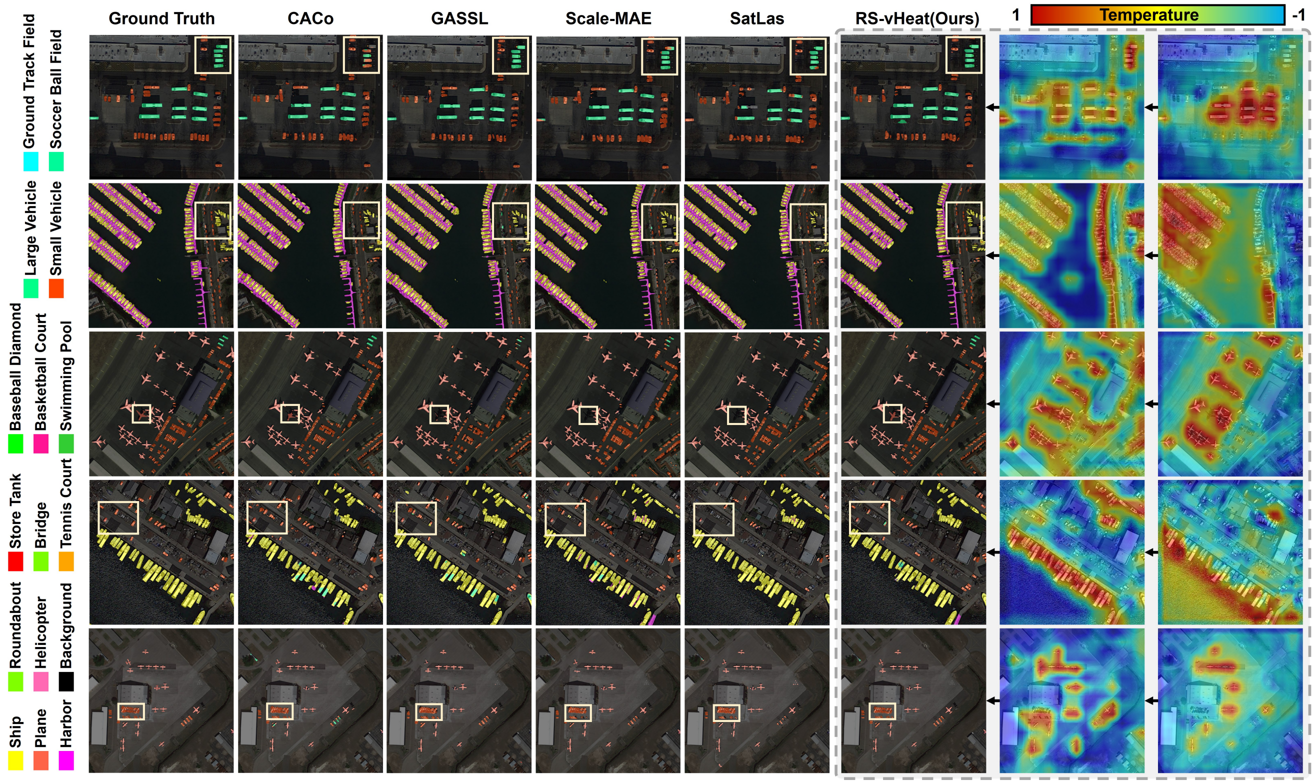

Figure 7. The qualitative results of RS-vHeat and several representative RSFMs on the iSAID dataset. Each column from left to right represents: ground truth, CACo (ResNet-18), GASSL (ResNet-50), Scale-MAE (ViT-L), Satlas (Swin-B), and the results from our model, RS-vHeat. The last two columns on the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

Table 10. Comparison of AP$_{50}$ for each category, mAP$_{50}$ and mAP$_{75}$ 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 [45]	TPAMI'2016	85.0	97.2	78.5	74.0	55.1	72.9	70.1	76.1	62.2
Cascade R-CNN [4]	CVPR'2018	87.4	97.5	74.0	78.0	54.5	68.3	69.1	75.7	58.9
RepPoints [73]	ICCV'2019	89.8	**97.9**	71.4	73.0	55.7	51.8	68.4	72.6	53.3
SKG-Net [19]	JSTARS'2021	79.3	78.2	66.4	65.0	65.1	69.6	71.4	70.7	46.4
SA-Net [78]	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**

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 [51] and SkySense [24], 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-PolSAR-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 [21] 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 [68] 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 [52] is an optical fine-grained dataset with

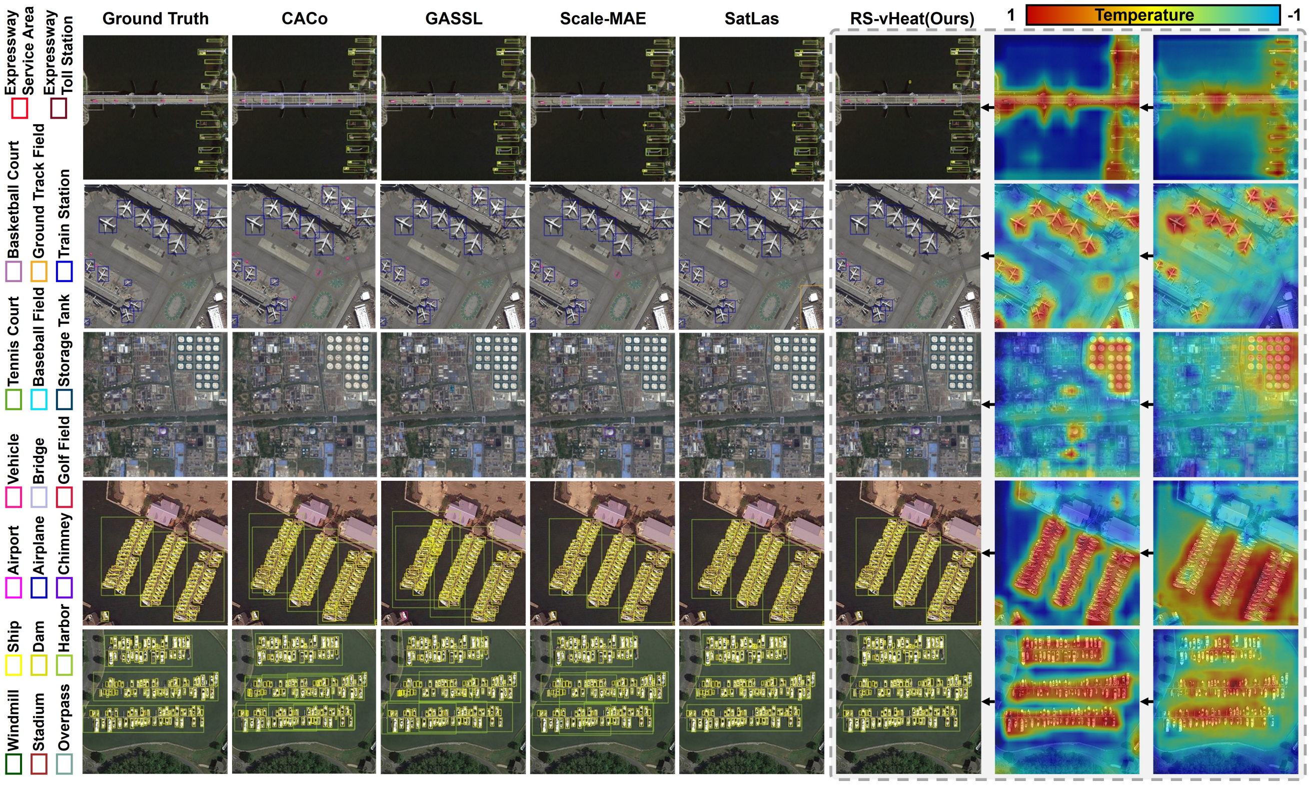

Figure 8. The qualitative results of RS-vHeat and several representative RSFMs on the DIOR dataset. Each column from left to right represents: ground truth, CACo (ResNet-18), GASSL (ResNet-50), Scale-MAE (ViT-L), Satlas (Swin-B), and the results from our model, RS-vHeat. The last two columns on the right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

objects annotated using OBB, encompassing five major categories, further divided into 37 subcategories. The dataset contains over 40,000 images. Following the official 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 [78] 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 [30] 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 AP_{50} for each category, mAP_{50} and mAP_{75}. mAP_{50} and mAP_{75} 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 Tab. 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 [7] 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 [6] 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) [64] 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 [13] 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 [24], using 20% and 50% of the AID dataset as training sets, and 10% and 20% of the NWPU-RESISC45 dataset as training sets.