remote sensing

Article

DiffuPrompter: Pixel-Level Automatic Annotation for High-Resolution Remote Sensing Images with Foundation Models

Huadong Li [1,2], **Ying Wei** [1,*], **Han Peng** [2] **and Wei Zhang** [2]

[1] College of Information Science and Engineering, Northeastern University, Shenyang 110819, China; 2110333@stu.neu.edu.cn

[2] Peng Cheng Laboratory, Shenzhen 518055, China; pengh@pcl.ac.cn (H.P.); zhangw05@pcl.ac.cn (W.Z.)

[*] Correspondence: weiying@ise.neu.edu.cn

Abstract: Instance segmentation is pivotal in remote sensing image (RSI) analysis, aiding in many downstream tasks. However, annotating images with pixel-wise annotations is time-consuming and laborious. Despite some progress in automatic annotation, the performance of existing methods still needs improvement due to the high precision requirements for pixel-level annotation and the complexity of RSIs. With the support of large-scale data, some foundational models have made significant progress in semantic understanding and generalization capabilities. In this paper, we delve deep into the potential of the foundational models in automatic annotation and propose a training-free automatic annotation method called DiffuPrompter, achieving pixel-level automatic annotation of RSIs. Extensive experimental results indicate that the proposed method can provide reliable pseudo-labels, significantly reducing the annotation costs of the segmentation task. Additionally, the cross-domain validation experiments confirm the powerful effectiveness of large-scale pseudo-data in improving model generalization performance.

Keywords: automatic labeling; instance segmentation; remote sensing; prompt generation; training-free

Citation: Li, H.; Wei, Y.; Peng, H.; Zhang, W. DiffuPrompter: Pixel-Level Automatic Annotation for High-Resolution Remote Sensing Images with Foundation Models. *Remote Sens.* **2024**, *16*, 2004. https://doi.org/10.3390/rs16112004

Academic Editors: Ahmed Shaharyar Khwaja and Filippo Biondi

Received: 23 April 2024
Revised: 24 May 2024
Accepted: 29 May 2024
Published: 2 June 2024

1. Introduction

With the development of deep learning, the interpretation of RSIs has also made significant progress [1]. Instance segmentation is a crucial part of remote sensing interpretation. However, it is also a data-intensive task that requires a significant amount of pixel-level annotations, which are labor-intensive and expensive, limiting the development of the task. To reduce the annotation costs, some scholars have introduced the automatic image annotation task, AIA for short, which has become an integral component of computer vision [2]. Although the scarcity of pixel-level annotated datasets in the remote sensing domain may be more severe, previous AIA methods have mainly focused on natural images. The perspective effect in natural images often results in distinct foreground and background elements. Therefore, some methods use the attention maps of classification networks to locate the foreground region and consider the image category as the foreground class to generate a pseudo-mask for objects [3–5]. In contrast, the top-down perspective and complex image content in RSIs diminish the effectiveness of these AIA methods.

Recently, the potential of big data has been further explored, leading to the emergence of many fundamental models trained on large-scale datasets, such as the Stable Diffusion Model (SDM) [6] and the Segment Anything Model (SAM) [7], which show significant contributions to downstream tasks [8]. Although fundamental models have become plentiful, none are tailored for RSI, which limits their application in RSI tasks. The main goal of this paper is not to create a fundamental model tailored for RSIs but to investigate the applicability of existing fundamental models to pixel-level AIA in RSIs.

With the development of text-guided generation models such as SDM, a pixel-level AIA technical route based on synthetic images has emerged in natural images [9,10]. These methods utilize generative models to synthesize data and generate masks for objects in synthetic images via vision–text alignment knowledge during generation [11]. Generative-based methods account for two assumptions: (1) the synthetic images are realistic enough to avoid domain shift issues between training and testing sets, and (2) the vision–text alignment knowledge can guide the generation of sufficiently accurate object masks. While these assumptions hold in natural images, they do not necessarily apply in RSIs. Generative models trained on natural images cannot synthesize RSIs realistically enough, and the complex and diverse scenes in remote sensing images seriously interfere with vision–text alignment, making it difficult to segment objects accurately. Some scholars incorporate the SAM model into instance segmentation methods for more accurate results [12]. After training on over one billion masks, the SAM model has demonstrated an outstanding ability to segment anything. However, SAM is a class-agnostic segmentation method and requires prior positional cues, such as points and bounding boxes, to segment target objects. These limitations prevent SAM from being directly applied to the instance segmentation task.

As shown in Figure 1, we present new insight into automatically obtaining mask annotations for authentic images using pre-trained foundational models to reduce the annotation cost of the instance segmentation task. Based on the insight, we propose a training-free prompt generation method called DiffuPrompter, which transforms SAM from a category-agnostic into a category-aware segmentation method to label RSIs automatically. Specifically, the proposed DiffuPrompter model leverages the text-concept grounding capabilities of a pre-trained diffusion model to provide coarse localization results for target objects. These localization results are then used as visual segmentation prompts for the SAM model, enabling precise segmentation of the target objects. We tested various automatic annotation methods on remote sensing datasets, and the experimental results validated the superiority of DiffuPrompter, i.e., it achieved 27.3% and 15.4% AP on the NWPU and iSAID datasets, respectively. Furthermore, the cross-domain study demonstrates the positive impact of pseudo-labels on improving model generalization performance, providing a valuable reference for future research.

The main contributions of this paper can be summarized as follows:

1. We present the novel insight that it is possible to automatically obtain the mask annotation of authentic images using off-the-shelf foundational models.
2. We propose a training-free prompt generation method, DiffuPrompter, that transforms SAM from a class-agnostic segmenter to a class-aware segmenter to label RSIs automatically.
3. We tested several automatic annotation methods on remote sensing datasets, and the extensive results validated the superiority of the proposed DiffuPrompter while proving the positive impact of pseudo-labels on enhancing model generalization performance. The results may provide a reference for future work.

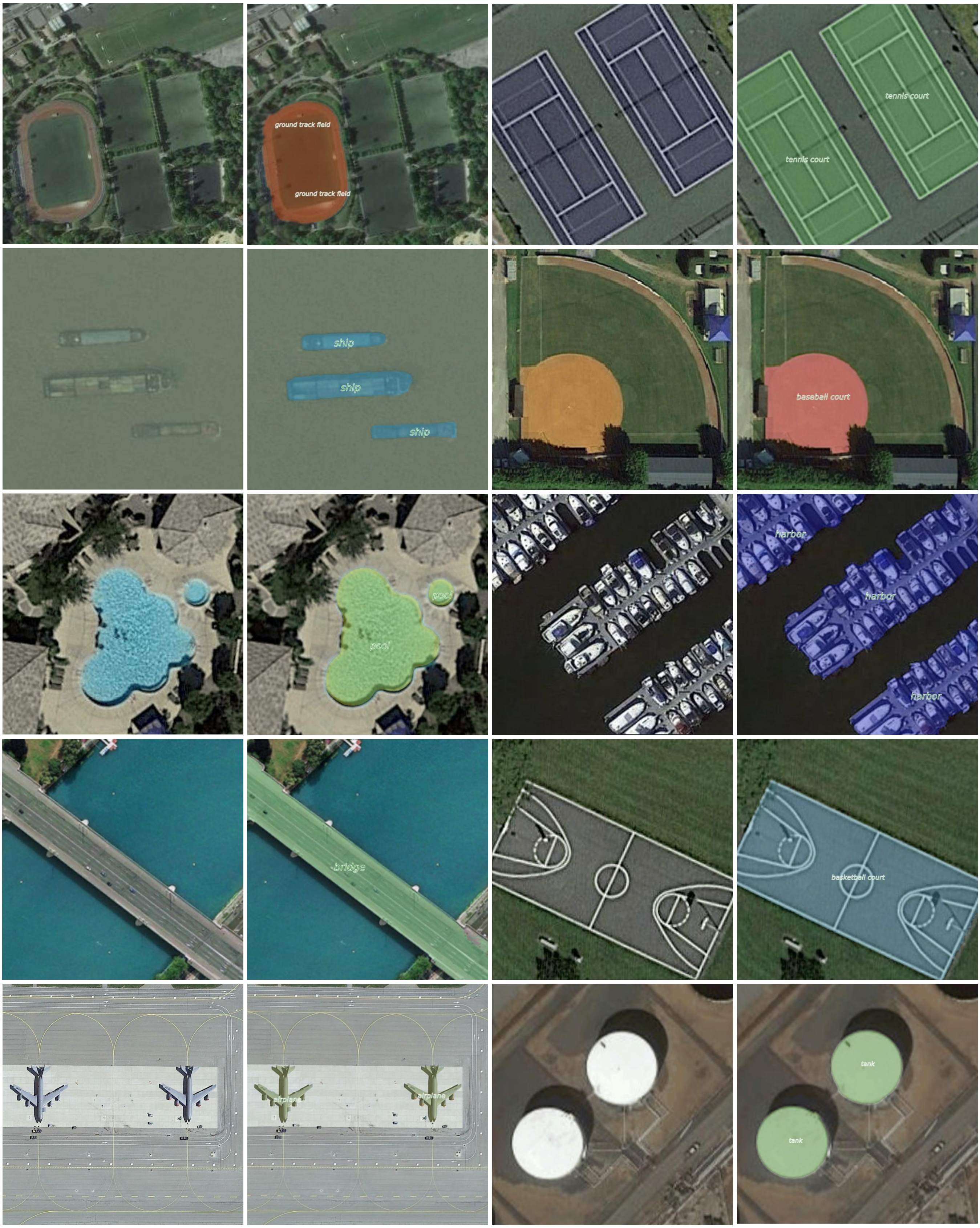


Figure 1. DiffuPrompter classification images, with pixel-level annotations labeled by DiffuPrompter.

2. Theory and Methods

DiffuPrompter utilizes the pre-trained SDM to explore generating semantically explicit prompts for SAM, enabling it to generate masks for specified remote sensing objects automatically. Section 2.1 introduces the working principles of SDM and SAM. In Section 2.2.1, we introduce how to realize the grounding of textual concepts into input images, and we discuss noise suppression in Section 2.2.2. Section 2.2.3 introduces how to prompt SAM to segment specific objects.

2.1. Preliminary Knowledge

2.1.1. Overview of SDM

SDM [6] is derived from a perceptual compression model consisting of an autoencoder, a U-Net, and a decoder. Specifically, when we input an image $x \in \mathbb{R}^{H \times W \times 3}$, the encoder $\mathcal{E}$ encodes x into a latent representation $z = \mathcal{E}(x) \in \mathbb{R}^{h \times w \times c}$, and the decoder $\mathcal{D}$ reconstructs the image from z, i.e., $\tilde{x} = \mathcal{D}(z) = \mathcal{D}(\mathcal{E}(x))$. The encoder downsamples the image according to the sampling factor $f = H/h = W/w$, where f has different values in different layers of the encoder and U-Net, namely $f = 2^m$, where $m \in \mathbb{N}$.

The training process of SDM consists of a forward diffusion and a backward denoising stage. In the forward diffusion stage, SDM adds noise to z for T steps until z is completely replaced by noise $\mathcal{N}(0, 1)$. In the backward denoising stage, SDM learns to gradually remove the noise by U-Net based on the textual condition to recover z. Finally, z is decoded into an image by $\mathcal{D}(z)$. SDM achieves semantic mapping between visual and textual inputs through the cross-attention mechanism in U-Net [13]. To pre-process the condition text prompt $\mathcal{P}_{text}$, SDM introduces a domain-specific encoder τ_θ that projects $\mathcal{P}_{text}$ to an intermediate representation, which is then mapped to the intermediate layers of the U-Net via cross-attention layers as follows:

$$\begin{aligned} \text{Attention}(Q, K, V) &= \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V, \\ Q &= W_Q^{(i)} \cdot \varphi_i(z_t), \\ K &= W_K^{(i)} \cdot \tau_\theta(\mathcal{P}_{text}), \\ V &= W_V^{(i)} \cdot \tau_\theta(\mathcal{P}_{text}). \end{aligned} \tag{1}$$

Here, $\varphi_i(z_t) \in \mathbb{R}^{N \times d_\epsilon^i}$ refers the intermediate embedding in the U-Net implementation, and $W_V^{(i)} \in \mathbb{R}^{d \times d_\mathcal{E}^i}$ and $W_Q^{(i)} \in \mathbb{R}^{d \times d_\tau} \& W_K^{(i)} \in \mathbb{R}^{d \times d_\tau}$ are learnable projection matrices [13,14].

2.1.2. Overview of SAM

SAM is an interactive segmentation approach predicated on provided prompts such as instance points and bounding boxes. The mask-generation process can be expressed as follows:

$$\begin{aligned} z_{\text{img}} &= \mathcal{E}_{i\text{-}enc}(x), \\ z_{inter} &= \mathcal{E}_{p\text{-}enc}(\mathcal{P}_{inter}), \\ z_{mask} &= \mathcal{E}_{p\text{-}enc}(\mathcal{P}_{mask}), \\ z_{out} &= \text{Cat}\Big(T_{mc\text{-}filter}, T_{\text{IoU}}, z_{inter}\Big), \\ \mathcal{M} &= \mathcal{E}_{m\text{-}dec}\big(z_{img} + z_{mask}, z_{\text{out}}\big), \end{aligned} \tag{2}$$

where $z_{img} \in \mathbb{R}^{h \times w \times c}$ represents the latent representation of the input image; $\mathcal{P}_{inter}$ denotes the interactive prompts, including points and bounding boxes; $\mathcal{P}_{mask} \in \mathbb{R}^{k \times c}$ signifies the mask prompt tokens, which are from the previous prediction iteration; $\mathcal{E}_{p\text{-}enc}$ encodes prompts into features $z_{inter} \in \mathbb{R}^{h \times w \times c}$ and $z_{mask} \in \mathbb{R}^{h \times w \times c}$; $T_{mc\text{-}filter} \in \mathbb{R}^{4 \times c}$ and $T_{IoU} \in \mathbb{R}^{1 \times c}$ are the pre-inserted learnable tokens representing four different mask filters and their corresponding IoU predictions; and $\mathcal{M}$ denotes the predicted masks. The primary objective of DiffuPrompter is to provide $\mathcal{P}_{inter}$ (points and bounding boxes) for SAM to segment the target object.

2.2. Proposed Method

2.2.1. Textual Concept Grounding

Upon further exploration of the SDM training process, we discovered that, during the training of SDM, the U-Net restores the latent presentations of the input image step by step based on the text description. Equation (1) illustrates that textual concepts $V = W_V^{(i)} \cdot \tau_\theta(\mathcal{P}_{text})$ are directed into latent presentations through cross-attention of the cross-modal spatial transformer module in U-Net.

Based on this observation, we constructed a text semantic grounding pipeline centered on the cross-attention map. This pipeline utilizes classification datasets as its data source and grounds image categories into the images. As illustrated in Figure 2, given an image from a classification dataset, the textual description is achieved by using the 'Photo of a category' template to process the corresponding class name. Then, the cross-attention layer grounds each text semantic in the template into the visual space by cross-attention maps as follows:

$$\mathcal{A} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right), \tag{3}$$

where $\mathcal{A} \in \mathbb{R}^{H \times W}$ denotes the re-shaped attention map. For the j-th text token, e.g., airplane in Figure 3a, the corresponding cross-attention $\mathcal{A}_j \in \mathbb{R}^{H \times W}$ shows the visual location in $\varphi(z_t)$ of the j-th token.

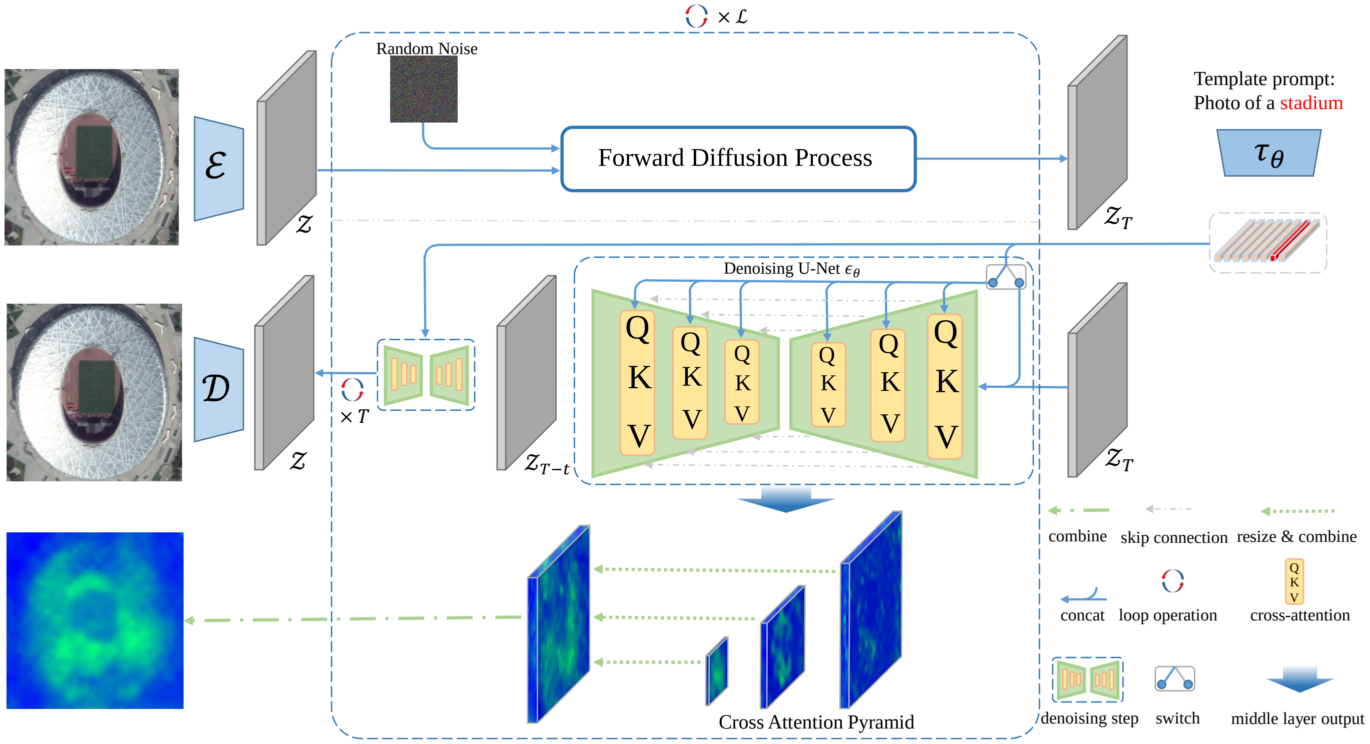


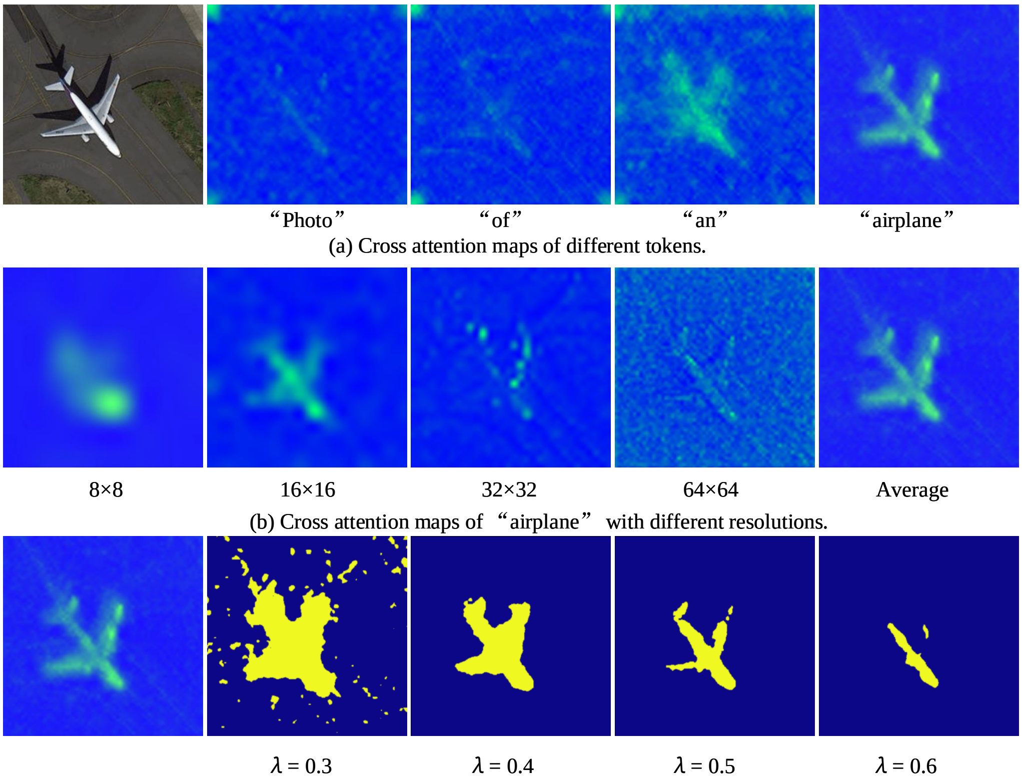
Figure 2. Pipeline for our method with the prompt 'Photo of a stadium'. DiffuPrompter mainly includes three steps: (1) Organize the object name into the template and use it as a text prompt. (2) Object mask proposal generation. (3) The denoising strategy is applied to refine the proposals.

We propose integrating grounding results at different resolutions to enhance the accuracy and robustness. The cross-attention pyramid $\mathcal{A}_j^s$ is obtained by applying Equation (3) to different layers in U-Net, where s denotes the attention map from the s-th layer of U-Net. We extract four resolutions in this paper, i.e., 8×8, 16×16, 32×32, and 64×64, as shown in

Figure 3b. Then, we aggregate multi-scale grounding results in the cross-attention pyramid by calculating the average map as follows:

$$\hat{\mathcal{A}}_j = \text{Norm}\left(\frac{1}{S}\sum_{s\in S}\frac{\mathcal{A}_j^s}{\max(\mathcal{A}_j^s)}\right), \tag{4}$$

where S represents the total number of layers (i.e., four for U-Net). Finally, the attention maps are transformed into probability maps through a normalization layer to facilitate subsequent binarization processing.



(a) Cross attention maps of different tokens.

(b) Cross attention maps of “airplane” with different resolutions.

(c) Binarization mask with different thresholds λ .

Figure 3. Cross-attention maps of SDM. Text prompt: ‘Photo of an airplane’.

2.2.2. Denoise by Noise

Figure 3a indicates that the highlighted regions in the cross-attention map correlate with the regions where that input token is presented. However, the maps are significantly noisy. Figure 4 compares the cross-attention maps between natural and remote-sensing images. Figure 4a shows a precise attention map for the ‘horse’ region. On the contrary, the attention map in Figure 4b for an RSI only roughly indicates the ‘playground’ region, showing a very weak correlation with regions a human would pick out as meaningful. Therefore, it is necessary to denoise the cross-attention maps in SDM before using them to localize remote sensing objects. However, the noise points tend to be localized and demonstrate high randomness in their distribution. Hence, achieving precise removal of these noise points poses a significant challenge.

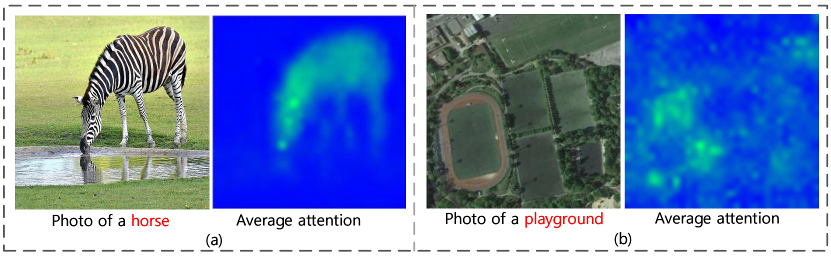


Figure 4. The cross-attention maps of natural and remote-sensing images.

Inspired by [15], we propose a Loop-Sampling Averaging Denoising (LSAD) strategy to suppress noise interference. In LSAD, we model the observed cross-attention map as a combination of a noise-free attention map and additive noise, as follows:

$$\mathcal{A}(x,y) = A(x,y) + \mathcal{N}(x,y), \tag{5}$$

where $\mathcal{A}(x,y)$ represents the value of the captured cross-attention map at coordinates (x,y), $A(x,y)$ signifies the value of the noise-free map, and $\mathcal{N}(x,y)$ denotes the value of the noise. The denoising process is the procedure of approximating $A(x,y)$ from the known $\mathcal{A}(x,y)$. For multiple cross-attention maps of the same input image and token, the $A(x,y)$ will remain constant, and $\mathcal{N}(x,y)$ is random. Thus, the mean of $\mathcal{L}$ maps of the same image can be represented as follows:

$$\begin{aligned}\bar{\mathcal{A}}(x,y) &= \frac{1}{\mathcal{L}}\sum_{i=1}^{\mathcal{L}}[A(x,y) + \mathcal{N}_i(x,y)] \\ &= A(x,y) + \frac{1}{\mathcal{L}}\sum_{i=1}^{\mathcal{L}}\mathcal{N}_i(x,y),\end{aligned} \tag{6}$$

where $\bar{\mathcal{A}}(x,y)$ denotes the mean value of the maps at coordinates (x,y), and $\mathcal{L}$ is the total number of maps considered. As the noise is random and unrelated, the expectation of the mean value approximates zero, i.e., $\sum_{i=1}^{k}\mathcal{N}_i(x,y) \approx 0$. Therefore, the expected mean and variance of the cross-attention maps can be expressed as follows:

$$\mathbb{E}_{\bar{\mathcal{A}}(x,y)} = A(x,y), \tag{7}$$

$$\sigma_{\bar{\mathcal{A}}(x,y)} = \frac{1}{\sqrt{\mathcal{L}}}\sigma_{\mathcal{N}(x,y)}, \tag{8}$$

where σ represents the standard deviation. Equation (7) shows that the expected mean value of multiple cross-attention maps is a map without noise. However, there will be some disturbances, and the standard deviation determines the noise's intensity. The essence of denoising is reducing the standard deviation. Equation (8) indicates that, by increasing the value of $\mathcal{L}$, i.e., increasing the number of averaged maps, the noise can be suppressed effectively.

The main challenge of applying Equation (8) to denoise cross-attention maps lies in introducing random noise onto $\mathcal{A}(x,y)$. As mentioned in Section 3, the SDM is trained to construct a clear image from Gaussian noise by removing Gaussian noise step by step. Noise contributes to generation diversity [16], implying attention maps are variable. Therefore, injecting Gaussian noise into the latent embeddings of the input image will result in multiple noisy cross-attention maps. Fortunately, the forward diffusion process in SDM is the noise addition process. Therefore, given an image, we perform an iterative forward diffusion process, preserving the cross-attention pyramid maps during each loop. Subsequently, we apply LSAD to them as follows:

$$\bar{\mathcal{A}}_j(x,y) = \frac{1}{\mathcal{L}}\sum_{l\in\mathcal{L}}\hat{\mathcal{A}}_j^l(x,y), \tag{9}$$

where l represents the l-th sampling iteration. Figure 5 illustrates the workflow of the LSAD algorithm.

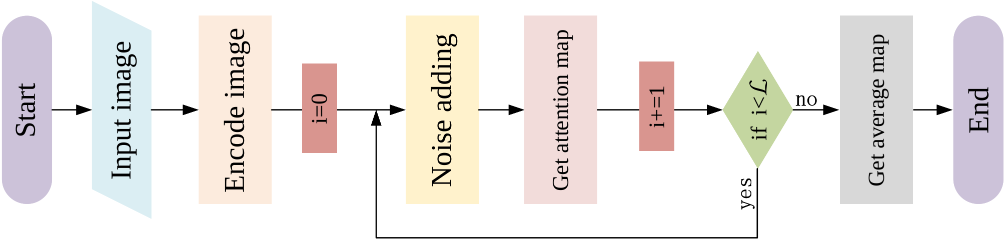


Figure 5. Flow chart of LSAD.

Figure 6 visualizes the performance of LSAD on natural and remote sensing images. It can be observed that there is no noticeable noise in the cross-attention map of the natural image. LSAD does not show a significant enhancement effect on the cross-attention map. In contrast, there is much noise in the cross-attention map of the RSI, making it challenging to locate the target object accurately. After LSAD processing, the noise is effectively suppressed, making the highlighted areas more meaningful.

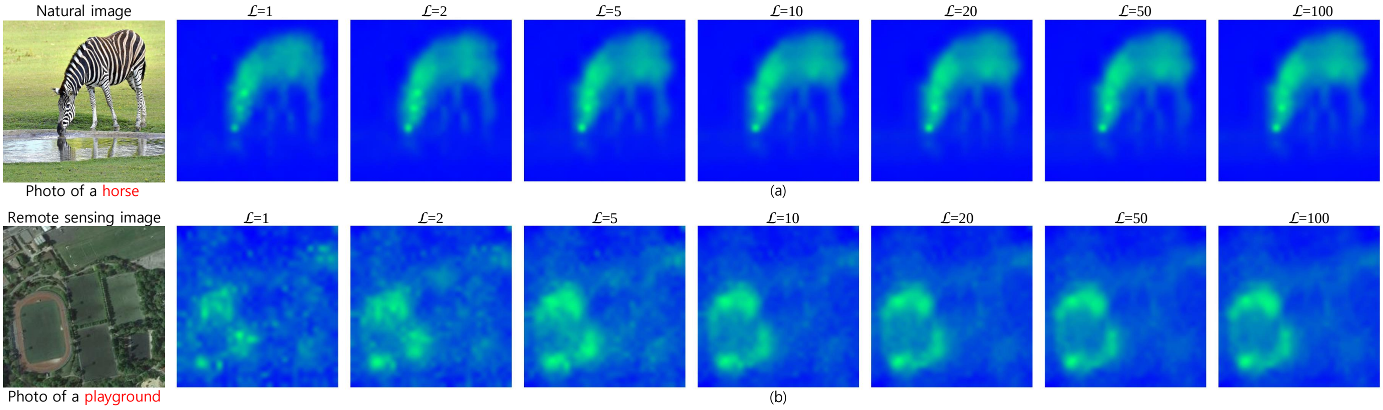


Figure 6. Visualization of denoising effects with different sampling times; t = 40 was applied to each sampling process. (**a**) visualizes the denoising effect on a natural image; (**b**) visualizes the denoising effect on a remote sensing image.

2.2.3. Prompt for SAM

Given a normalized average attention map $\bar{\mathcal{A}} \in \mathbb{R}^{H\times W}$ for the j-th text token to get the target object region (e.g., 'airplane'). As shown in Figure 3c, the solution to the binarization process is using a threshold value γ and refining with DenseCRF [17] as follows:

$$\mathcal{B} = \text{DenseCRF}\Big(\big[\gamma; \hat{\mathcal{A}}_j\big]_{\text{argmax}}\Big). \tag{10}$$

As shown in Figure 7, we take the minimum bounding box and centroid of the areas in $\mathcal{B}$ with a value of 1 as the box and point prompts for SAM. Then, SAM will output a mask list based on the prompts. We select the mask with the highest IoU, with the attention mask in $\mathcal{B}$ as the final segmentation result. If the selected mask contains multiple closed intervals, it is considered to have multiple objects, such as the boat, airplane, tennis court, and storage tank in Figure 7. If the selected mask contains a single closed interval, we consider that there is a single object, such as the playground. At this point, we have constructed a training-free, pixel-level AIA pipeline using SDM and SAM.

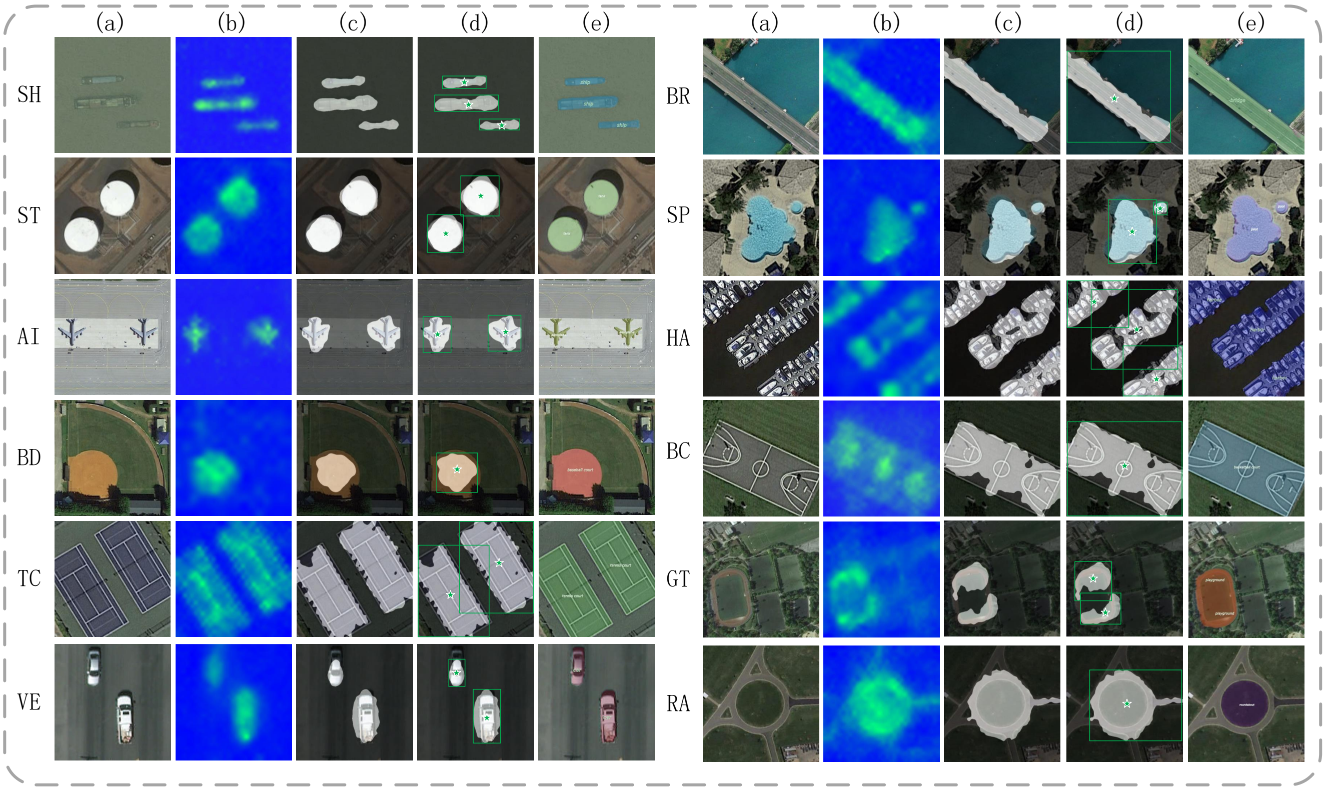


Figure 7. Visualization of the DiffuPrompter mask generation process: (**a**) original image, (**b**) cross-attention map, (**c**) binarized map, (**d**) box and point prompts, (**e**) segmentation result.

3. Results

3.1. Datasets

iSAID: iSAID [18] is a large-scale dataset for remote sensing instance segmentation inherited from DOTA [19]. The spatial resolutions of images range between 800 and 13,000. We split them into 512 × 512 patches during training and testing. It contains 15 classes of 655, 451 instances in 2806 images: ship, storage tank, baseball diamond, tennis court, basketball court, playground, bridge, large vehicle, small vehicle, helicopter, swimming pool, roundabout, soccer ball field, plane, and harbor.

NWPU VHR-10: NWPU VHR-10 [20] is another widely used dataset for object detection of RSIs. It has 800 high-resolution images, among which 650 are positive and 150 are negative, without any objects of interest. This dataset contains annotations of 10 object categories: airplane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, and vehicle.

Classification Dataset: We selected target images corresponding to the segmentation dataset from 11 classification datasets: UC Merced Land Use Dataset [21], WHU-RS19 [22,23], RSSCN7 [24], RS_C11 [25], NWPU-RESISC45 [26], AID [26], RSD46-WHU [27,28], PatternNet [29], OPTIMAL-31 [30], CLRS [31], and DLR Munich Vehicle [32]. Ultimately, we collected 9300 RSIs across 12 categories of 0.3∼3 m resolution: airplane, ship, storage tank, baseball diamond, swimming pool, tennis court, basketball court, roundabout, ground track field, harbor, bridge, and vehicle. The corresponding classes are selected when testing on the iSAID and NWPU datasets. The spatial resolution is uniformly set to 512 × 512.

3.2. Evaluation Metrics

We adopted the commonly used mean average precision (mAP) metric to evaluate the performance of the proposed method. When the mask of an instance exists an intersection-over-union (IoU) with a mask of ground truth above a threshold and its predicted category matches the label, the prediction is considered to be true positive. In this study, we employ $\mathbf{AP}$, $\mathbf{AP_{50}}$, $\mathbf{AP_{75}}$, $\mathbf{AP}_S$, $\mathbf{AP}_M$, and $\mathbf{AP}_L$ for evaluation. **AP** refers to metrics averaged across all 10 IoU thresholds (0.50:0.05:0.95) and all categories. A larger **AP** value denotes more accurate predicted instance masks and superior instance segmentation performance. $\mathbf{AP_{50}}$ represents the calculation under the IoU threshold of 0.50, while $\mathbf{AP_{75}}$ embodies a stricter metric corresponding to the calculation under the IoU threshold 0.75. Therefore, if we havethe same $\mathbf{AP_{75}}$ and $\mathbf{AP_{50}}$ values, qhere the $\mathbf{AP_{75}}$ indicates more accurate instance masks. $\mathbf{AP}_L$ is set for large targets (area > 96^2); $\mathbf{AP}_M$ is set for medium targets (32^2 < area < 96^2); and $\mathbf{AP}_S$ is set for small targets (area < 32^2).

3.3. Implementation Details

In this paper, we do not train any parameters in the SDM and SAM. Due to the lack of information about vehicle sizes in the remote sensing classification datasets, we merged ‘small-vehicle’ and ‘large-vehicle’ as one category, i.e., ‘vehicle’, during testing on iSAID. Additionally, since the “helicopter” and “soccer ball field” categories do not exist in the classification dataset, we did not generate pseudo-labels for them. Finally, we collected 9300 RSIs. The total pseudo-label categories and their abbreviations annotated in this paper are: AI-airplane, SH-ship, ST-storage tank, BD-baseball diamond, TC-tennis court, BC-basketball court, RA-roundabout, PL-playground (ground track field), SP-swimming pool, HA-harbor, BR-bridge, and VE-vehicle. Mask R-CNN [33], Cascade R-CNN [34], and Mask2Former [34] were used as the baseline to evaluate our method. Eight Tesla V100 GPUs were used to generate pseudo-labels, which took approximately 96 h.

3.4. Qualitative Experiments

Visualizations in Figure 7 depict the intermediate results of DiffuPrompter in generating pseudo-labels. It can be observed that the proposed method can accurately segment any number of target objects, significantly increasing the number of positive samples in the pseudo-labels. In the Stable Diffusion model, time steps control the noise intensity, affecting the results of DiffuPrompter. Figure 8 visualizes the cross-attention maps at different time steps and shows that the cross-attention map at time step 40 is the clearest, maintaining the structure of the target object. The reason may be that, at t = 40, the artificially introduced noise intensity is close to the inherent noise intensity in the cross-attention map. Therefore, LSAD can effectively suppress noise without losing the contours of the target objects due to excessive noise. Thus, we use cross-attention in step 40 to conduct experiments throughout the paper.

3.5. Ablation Study

We also performed an extensive ablation analysis to better understand the effectiveness of each proposed module in our DiffuPrompter.

3.5.1. Comparison with Attention Map under Different Thresholds

Figure 3c illustrates the impact of different thresholds on the binary image. It is evident that the λ value significantly impacts the prompt quality for SAM. Table 1 qualitatively compares the segmentation performance under different binary thresholds. The term “cross-attention” means using the binarized attention map as pseudo-labels to train the segmentation model. The experimental results indicate that setting the threshold to 0.4 provides the best guidance for the model to segment target objects in both methods. Therefore, the value of λ in subsequent experiments is set to 0.4. Additionally, the performance of DiffuPrompter far exceeds that of the attention map across all thresholds. This can be attributed to the superior segmentation ability of SAM, which provides accurate pseudo-masks for the segmentation model.

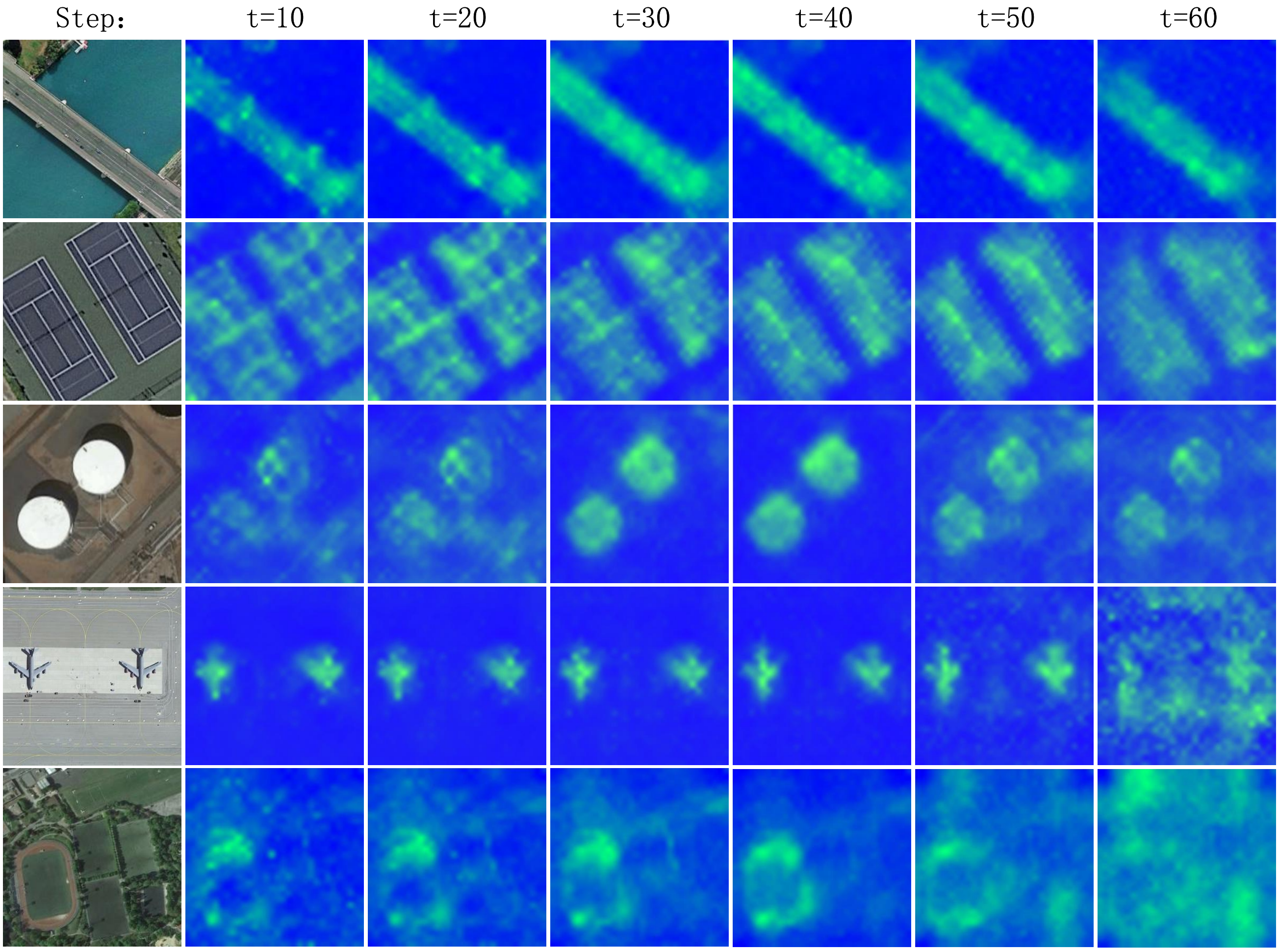


Figure 8. Visualization of cross-attention maps with different time steps. The $\mathcal{L}$ in LSAD is set to 50.

Table 1. The performance of Mask R-CNN trained by pseudo-labels generated by DiffuPrompter vs. cross-attention with different λ thresholds.

Mask	λ	NWPU			iSAID		
		AP	AP $_{50}$	AP$_{75}$	AP	AP$_{50}$	AP$_{75}$
Cross-Attention	0.3	10.3	17.7	13.2	3.1	7.4	5.5
	0.4	17.2	25.1	13.5	9.2	15.2	10.6
	0.5	15.6	22.3	19.7	5.3	9.1	7.4
DiffuPrompter	0.3	22.5	37.9	30.1	5.7	15.6	10.9
	0.4	27.3	50.2	36.1	15.4	31.2	19.4
	0.5	25.4	47.9	33.0	13.1	27.5	16.3

3.5.2. Sampling Times

Table 2 provides the related ablation study for sampling times in LSAD. The results reflect that, with the increase in sampling iterations, there is a significant improvement in segmentation performance, and the performance stabilized after reaching 50 sampling iterations. We ultimately adopted 50 iterations as the universal sampling count to balance performance and economy.

Table 2. The performance of Mask R-CNN trained on pure synthesis data under different $\mathcal{L}$ loop sampling times.

Times	**NWPU**			**iSAID**		
	AP	**AP$_{50}$**	**AP$_{75}$**	**AP**	**AP$_{50}$**	**AP$_{75}$**
1	4.3	6.7	4.2	2.4	3.7	2.1
5	7.6	10.3	8.6	4.5	7.3	6.5
10	10.3	12.5	10.5	5.4	9.6	7.8
20	20.5	48.6	33.3	12.5	27.8	16.5
50	27.3	50.2	36.1	15.4	31.2	19.4
100	27.2	50.4	36.2	15.6	31.1	19.5

3.6. Segmentation Performance Comparison

NWPU: Table 3 presents instance segmentation results on the NWPU. The baseline segmentation methods trained on the data labeled by DiffuPrompter can reach approximately half of the performance achieved with pure real data, e.g., 27.3% vs. 58.3% for mask *AP* of Mask R-CNN. Additionally, further fine-tuning on 600 (25% off) real data can achieve performance comparable to training on pure real data, e.g., 55.6% mask *AP* after fine-tuning vs. 58.3% mask *AP* training on pure real data with Mask R-CNN; 59.1% mask *AP* after fine-tuning vs. 59.8% mask *AP* training on pure real data with Cascade R-CNN; and 60.9% mask *AP* after fine-tuning vs. 61.3% mask *AP* training on pure real data with Mask2Former.

iSAID: Table 4 presents the results on iSAID. iSAID is more challenging than NWPU, as it includes more objects and complex backgrounds. Even in the absence of pseudo-labels for helicopter and football field categories, DiffuPrompter and iSAID still can present a competitive result, i.e., 34.8% vs. 35.1% mask *AP* of Mask R-CNN; 35.1% vs. 35.6% mask *AP* of Cascade Mask R-CNN; 36.8% vs. 37.1% mask *AP* of Mask2Former, when trained on 9300 pseudo-data and 2200 real data (saved 21.4% in manner effort).

Table 3. The performance of Mask R-CNN and Cascade R-CNN on the NWPU. 'P' and 'R' refer to 'Pseudo' and 'Real'.

Training Set	**Method**	**Size**	**AP**	**AP$_{50}$**	**AP$_{75}$**	**AP$_S$**	**AP$_M$**	**AP$_L$**
Training with Pure Real Label								
NWPU	Mask R-CNN	R: 0.8 k (all)	58.3	90.2	60.7	40.9	56.6	61.1
	Cascade R-CNN	R: 0.8 k (all)	59.8	91.9	66.6	45.3	60.0	67.3
	Mask2Former	R: 0.8 k (all)	61.3	92.5	68.6	46.3	62.7	69.5
	Mask R-CNN	R: 0.6 k	50.2	83.1	55.4	31.2	48.3	57.5
	Cascade R-CNN	R: 0.6 k	55.4	86.1	62.4	40.3	49.7	62.1
	Mask2Former	R: 0.6 k	56.2	87.1	64.1	42.5	53.8	64.2
Training with Pure Pseudo-Label								
DiffuPrompter	Mask R-CNN	P: 9.3 k	27.3	50.2	36.1	17.2	25.1	35.4
	Cascade R-CNN	P: 9.3 k	29.9	52.3	38.1	19.4	28.7	35.2
	Mask2Former	R: 9.3 k	30.3	53.4	40.1	21.3	30.5	37.2
Training with Pseudo and Real Label								
DiffuPrompter & NWPU	Mask R-CNN	P: 9.3 k R: 0.4 k	50.6	85.3	54.2	32.4	53.3	58.6
	Cascade R-CNN	P: 9.3 k R: 0.4 k	52.1	86.5	59.1	33.4	55.3	64.0
	Mask2Former	P: 9.3 k R: 0.4 k	54.6	88.7	64.2	40.9	53.6	62.1
	Mask R-CNN	P: 9.3 k R: 0.6 k	55.6	89.3	60.2	37.4	56.3	61.6
	Cascade R-CNN	P: 9.3 k R: 0.6 k	59.1	90.5	65.1	45.3	58.9	67.0
	Mask2Former	P: 9.3 k R: 0.6 k	60.9	92.6	66.9	46.0	61.8	68.2

Table 4. The performance of Mask R-CNN and Cascade R-CNN on the iSAID. Mask *AP* is for 15 classes. 'P' and 'R' refer to 'Pseudo' and 'Real'.

Training Set	Method	Size	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Training with Pure Real Label								
iSAID	Mask R-CNN	R: 2.8 k	34.8	57.4	37.0	20.5	43.2	50.3
	Cascade R-CNN	R: 2.8 k	35.6	57.8	38.0	20.8	44.3	52.7
	Mask2Former	R: 2.8 k	37.1	59.4	40.2	20.5	45.1	55.6
	Mask R-CNN	R: 2.2 k	29.5	52.3	31.4	12.5	37.7	46.9
	Cascade R-CNN	R: 2.2 k	31.2	51.4	31.6	13.5	39.7	47.2
	Mask2Former	R: 2.2 k	33.8	53.1	33.6	14.1	39.9	47.8
Training with Pure Pseudo-Label								
DiffuPrompter	Mask R-CNN	P: 9.3 k	15.4	31.2	19.4	9.3	20.6	33.7
	Cascade R-CNN	P: 9.3 k	16.1	32.5	21.5	11.1	22.7	36.2
	Mask2Former	R: 9.3 k	17.9	36.2	24.3	14.5	22.5	36.3
Training with Pseudo and Real Label								
DiffuPrompter & iSAID	Mask R-CNN	P: 9.3 k R: 1.4 k	31.6	54.3	32.4	15.3	41.7	48.4
	Cascade R-CNN	P: 9.3 k R: 1.4 k	32.7	55.1	34.3	15.8	43.3	51.1
	Mask2Former	P: 9.3 k R: 1.4 k	33.4	56.4	35.3	16.3	45.7	54.5
	Mask R-CNN	P: 9.3 k R: 2.2 k	35.1	57.3	36.4	17.3	42.7	50.4
	Cascade R-CNN	P: 9.3 k R: 2.2 k	35.2	57.1	38.3	17.8	43.9	52.5
	Mask2Former	P: 9.3 k R: 2.2 k	36.8	58.7	38.9	17.9	45.2	55.3

3.7. Domain Generalization

Table 5 delivers the results of cross-domain validation, which can evaluate the generalization performance. We tested the performance on overlapping categories under the two datasets. The results indicate that DiffuPrompter plays a prominent role in domain generalization, e.g., 22.5% **AP** with DiffuPrompter and NWPU vs. 17.9% **AP** with NWPU on the iSAID test set, and 50.3% **AP** with DiffuPrompter and iSAID vs. 47.2% **AP** with iSAID on the NWPU test set.

Table 5. Performance of domain generalization between different datasets. Mask R-CNN with ResNet50 is used as the baseline.

Training Set	Data Size	Test Set	AP	AP_{50}	AP_{75}
NWPU	R: 0.8 K (all)	iSAID	17.9	29.0	20.1
iSAID	R: 2.8 K (all)	NWPU	47.2	78.5	51.0
DiffuPrompter & NWPU	9.3K + R: 0.6K	iSAID	22.5	40.3	23.0
DiffuPrompter & iSAID	9.3K + R: 2.2K	NWPU	45.4	72.6	48.1
DiffuPrompter & NWPU	9.3K + R: 0.8 K (all)	iSAID	25.2	45.1	26.0
DiffuPrompter & iSAID	9.3K + R: 2.8 K (all)	NWPU	50.3	82.7	53.5

3.8. Comparison with the State of the Art

Figure 9 visualizes the output results of two advanced AIA methods and our DiffuPrompter. CAM originates from classification models and focuses more on discriminative regions while neglecting details. Consequently, the masks it generates often fail to cover the object entirely. DiffuMask, a recently proposed advanced algorithm, also leverages diffusion models to generate pseudo-masks for objects. It optimizes attention maps in U-Net through noise learning and uses them as pseudo-labels. However, DiffuMask is designed to generate pseudo-labels for synthetic images, which limits its annotation ability for authentic images. From Figure 9, it is evident that the optimized attention map is still easily disturbed by the complex background in RSIs. Moreover, the additional training process significantly increases its annotation cost compared to DiffuPrompter. In contrast, the proposed DiffuPrompter can accurately locate and label masks for target objects without training.

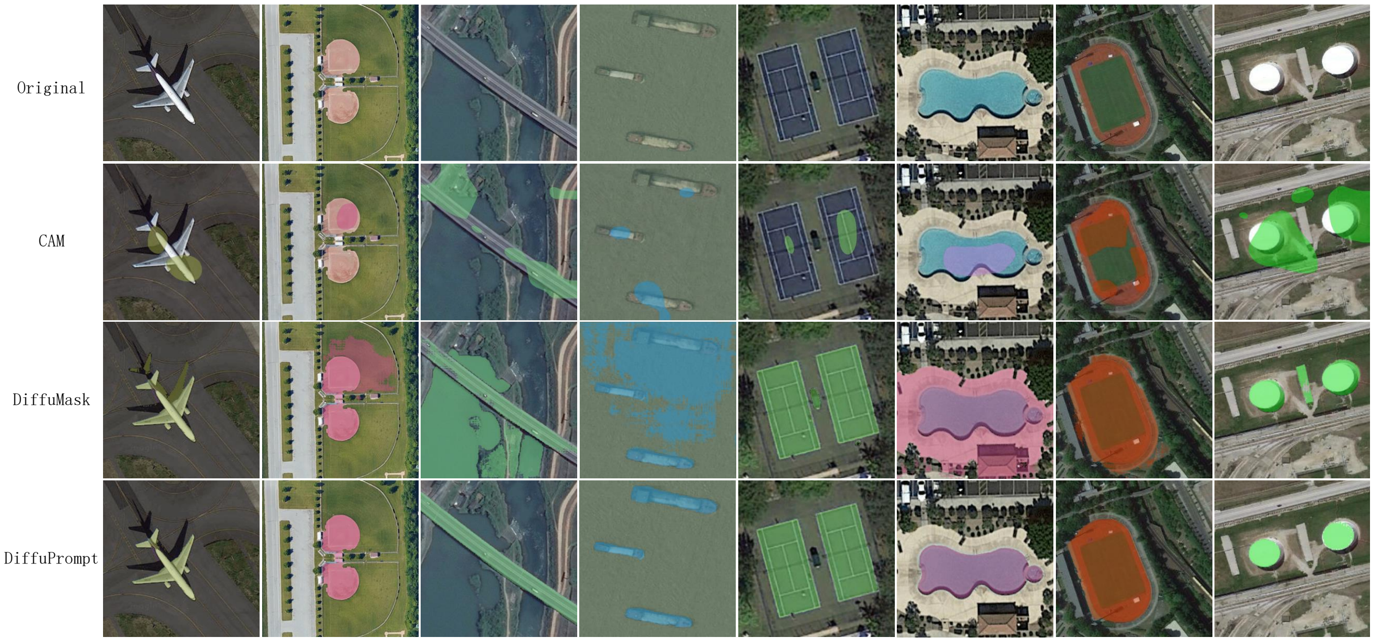


Figure 9. Some examples of pseudo-labels generated by different methods.

Table 6 quantitatively compares the performance of the proposed method with other state-of-the-art algorithms. The results in Table 6 show that the proposed method significantly outperforms the baselines in terms of accuracy. However, there is still room for improvement in annotation speed.

Table 6. Comparison of different AIA methods. The results are from segmentation methods trained on pseudo-labels constructed by these AIA methods. Seconds/im represents the time consumed by labeling each image with one V100 GPU.

Method		**Seconds/im**	**NWPU**			**iSAID**		
			AP	**AP$_{50}$**	**AP$_{75}$**	**AP**	**AP$_{50}$**	**AP$_{75}$**
CAM	Mask R-CNN	2.5	8.4	15.3	10.7	3.1	7.4	5.5
	Mask2Former		7.9	16.1	11.2	3.5	6.9	5.7
DiffuMask	Mask R-CNN	35.3	12.3	17.9	13.1	9.5	13.6	11.2
	Mask2Former		12.9	18.1	14.5	10.2	14.2	10.9
DiffuPrompter	Mask R-CNN	297.3	27.3	50.2	36.1	15.4	31.2	19.4
	Mask2Former		30.3	53.4	40.1	17.9	36.2	24.3

Table 7 compares DiffuPrompter with some advanced weakly supervised methods. Here, we denote the supervision type as: $\mathcal{I}$ (image-level label), $\mathcal{B}$ (box-level label). DiffuPrompter$_M$ and DiffuPrompter$_C$ represent Mask R-CNN and Cascade Mask R-CNN trained with pseudo-labels generated by DiffuPrompter, respectively. The results show that DiffuPrompter$_M$ and DiffuPrompter$_C$ significantly outperform methods based on image-level supervision, e.g., 29.9% **AP** with DiffuPrompter$_C$ vs. 13.3% **AP** with BESTIE on NWPU, and achieve performance comparable to methods based on object-level supervision, e.g., 29.9% **AP** with DiffuPrompter$_C$ vs. 29.8% **AP** with MGWI-Net on NWPU. In conclusion, DiffuPrompter can perform better with less manual labor than existing advanced methods.

Table 7. Performance comparison with some weakly supervised methods on the NWPU and iSAID datasets. 'Sup' refers to the supervision type.

Method	Sup	NWPU			iSAID		
		AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
CAM [35]	$\mathcal{I}$	8.4	15.3	10.7	3.1	7.4	5.5
SEC [36]	$\mathcal{I}$	21.4	30.0	23.1	4.2	12.7	10.3
AffinityNet [37]	$\mathcal{I}$	22.5	37.9	30.1	5.7	15.6	10.9
BESTIE [38]	$\mathcal{I}$	13.3	25.9	13.8	7.3	9.4	8.5
BoxSup [39]	$\mathcal{B}$	27.5	53.1	37.9	10.3	20.2	14.1
MGWI-Net [40]	$\mathcal{B}$	29.8	**62.9**	25.8	-	-	-
DiffuPrompter$_M$	$\mathcal{I}$	27.3	50.2	36.1	11.4	22.6	14.4
DiffuPrompter$_C$	$\mathcal{I}$	**29.9**	52.3	**38.1**	**13.1**	**25.5**	**16.3**

Table 8 compares the class-wise performance of DiffuPrompter$_M$ against several advanced supervised methods trained by ground truth. The results indicate that most categories achieve nearly half the performance of supervised methods. However, the performance of "bridge" and "vehicle" is significantly lower than that of supervised methods. We hypothesize that this is due to the significant scale difference between the pseudo-labels and the segmentation dataset for these two categories. The pseudo-labels do not adequately guide the segmentation of these objects in the target dataset. Therefore, addressing the cross-dataset issues between pseudo-labels and the target dataset is worth further investigation.

Table 8. Comparison of class-wise results with advanced supervision methods on NWPU dataset.

Method	AP	AI	SH	ST	BD	TC	BC	PL	HB	BR	VE
Mask R-CNN	58.3	28.4	52.8	69.6	81.4	59.6	69.6	84.3	60.7	25.8	50.6
PANet [18]	64.8	50.6	53.5	78.4	83.5	73.0	78.1	87.2	58.6	33.8	51.6
PointRend [41]	65.4	54.5	53.2	75.7	84.3	72.4	74.4	90.1	58.8	35.9	54.7
BlendMask [42]	65.7	48.1	51.1	79.8	84.0	72.4	76.7	91.5	58.9	39.6	54.6
CATNet [43]	73.3	51.9	64.4	87.1	89.4	75.8	79.7	95.0	65.0	53.2	72.0
DiffuPrompter$_M$	27.3	17.9	27.5	35.3	37.1	30.4	31.1	40.1	27.9	9.9	15.8

4. Discussion

To obtain more accurate prompts, we fine-tuned SDM on a combination of some RSI caption datasets (RSICD [44], UCM-captions [45], and Sydney-captions [45]). However, the cross-attention maps of the fine-tuned SDM, as shown in Figure 10, cannot provide precise guidance for SAM. The reason may be that the image caption datasets for RSIs are too small in scale and insufficient to convey accurate text–visual correspondence information to the SDM model. Therefore, the SDM used in this paper has yet to be fine-tuned. It is worth exploring how to fine-tune the SDM model on remote sensing datasets to obtain a more familiar understanding of remote sensing objects.

Another factor affecting performance is domain discrepancy. The significant differences between the classification and object detection datasets result in the knowledge provided by the pseudo-labels not being well applied to the dataset. Therefore, addressing the cross-domain issue between pseudo-labels and target data is also a valuable research direction.

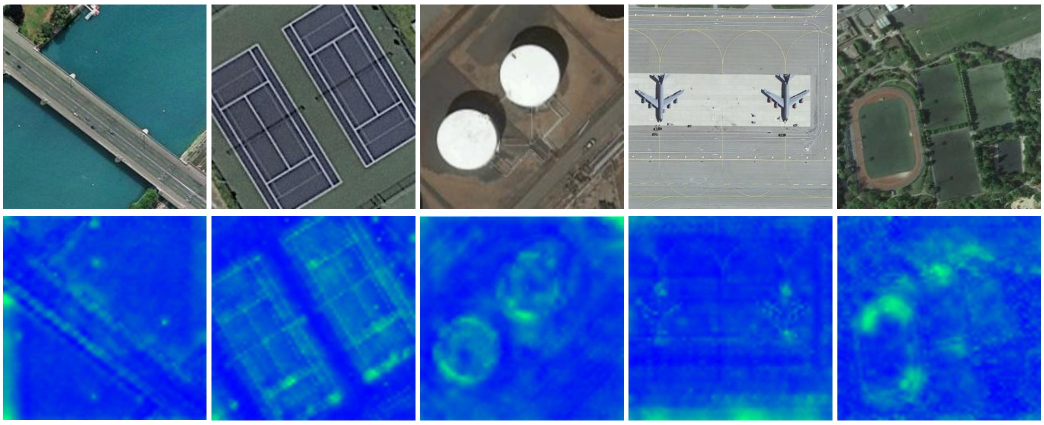


Figure 10. Cross-attention map with fine-tuned SDM.

In this paper, we determined the hyperparameter values based on the overall performance of the segmentation algorithm. However, the optimal hyperparameters vary slightly across different categories. Table 9 presents the category-level results under different parameters, indicating that designing adaptive hyperparameters could further improve the quality of the pseudo-labels. Therefore, developing an adaptive pseudo-labeling algorithm tailored to each category is a promising direction for future research.

Table 9. Class-wise results of Mask R-CNN on the NWPU VHR-10 test set with different values of λ and **t**.

t	λ	**AP**	**AI**	**SH**	**ST**	**BD**	**TC**	**BC**	**PL**	**HA**	**BR**	**VE**
30	0.3	12.3	8.9	11.5	8.6	22.5	11.3	21.6	15.9	9.6	9.1	4.1
	0.4	20.5	16.7	19.7	15.7	18.4	22.6	28.5	26.4	27.8	11.9	17.3
	0.5	18.5	16.0	20.7	19.7	19.4	25.0	12.9	19.2	23.9	10.6	17.5
40	0.3	22.5	27.7	30.0	21.1	30.0	19.7	28.0	22.1	21.8	10.5	14.1
	0.4	27.3	17.9	27.5	35.3	37.1	30.4	31.1	40.1	27.9	9.9	15.8
	0.5	25.4	35.0	35.0	20.1	35.0	18.9	18.9	35.0	30.8	12.7	12.6
50	0.3	19.5	23.7	14.2	25.0	21.1	24.0	21.1	11.3	23.0	25.0	6.4
	0.4	23.7	30.6	19.1	26.3	32.0	28.8	29.4	33.0	16.7	8.7	12.4
	0.5	20.1	32.4	30.5	20.9	24.4	15.9	21.0	24.6	16.8	8.5	6.2

5. Conclusions

This paper introduces a novel insight that demonstrates the possibility of automatically annotating object masks for RSIs by leveraging off-the-shelf foundational models. Building on the SDM and SAM models, we propose an AIA method, namely DiffuPrompter, capable of leveraging the text semantic grounding knowledge in SDM to generate semantically precise SAM prompts, enabling it to acquire instance masks autonomously. We comprehensively test the effectiveness of the proposed method on two general datasets. We first evaluate the efficacy of each component in DiffuPrompter through ablation studies. Then, the cross-domain validation experiments confirm the significant effectiveness of large-scale pseudo-data in improving model generalization performance. Finally, we compare our method with other state-of-the-art algorithms, and the results demonstrate the superiority of our proposed method over existing ones.

Author Contributions: Conceptualization, H.L.; theory and methodology, H.L.; software (Python 3.9), H.L. and H.P.; visualization, H.L.; formal analysis, H.L.; writing—original draft preparation, H.L.; writing—review and editing, H.L.; supervision, Y.W.; project administration, Y.W. and W.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Open Project Program Foundation of the Key Laboratory of Opto-Electronics Information Processing, Chinese Academy of Sciences (OEIP-O-202002), the National Nature Science Foundation of China (grant No. 61871106 and No. 61370152), and the Key R & D projects of Liaoning Province, China (grant No. 2020JH2/10100029).

Data Availability Statement: The original contributions presented in the study are included in the article, and further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Chen, K.; Zou, Z.; Shi, Z. Building extraction from remote sensing images with sparse token transformers. *Remote Sens.* **2021**, *13*, 4441. [CrossRef]
2. Cheng, Q.; Zhang, Q.; Fu, P.; Tu, C.; Li, S. A survey and analysis on automatic image annotation. *Pattern Recognit.* **2018**, *79*, 242–259. [CrossRef]
3. Wu, T.; Huang, J.; Gao, G.; Wei, X.; Wei, X.; Luo, X.; Liu, C.H. Embedded discriminative attention mechanism for weakly supervised semantic segmentation. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 16765–16774. [CrossRef]
4. Xu, L.; Ouyang, W.; Bennamoun, M.; Boussaid, F.; Sohel, F.; Xu, D. Leveraging auxiliary tasks with affinity learning for weakly supervised semantic segmentation. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 6984–6993. [CrossRef]
5. Ru, L.; Zhan, Y.; Yu, B.; Du, B. Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 16846–16855. [CrossRef]
6. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 10684–10695. [CrossRef]
7. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment anything. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 4015–4026. [CrossRef]
8. Chen, J.; Chen, H.; Chen, K.; Zhang, Y.; Zou, Z.; Shi, Z. Diffusion models for imperceptible and transferable adversarial attack. *arXiv* **2023**, arXiv:2305.08192.
9. Zhang, Y.; Ling, H.; Gao, J.; Yin, K.; Lafleche, J.F.; Barriuso, A.; Torralba, A.; Fidler, S. Datasetgan: Efficient labeled data factory with minimal human effort. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 10145–10155.
10. Li, D.; Ling, H.; Kim, S.W.; Kreis, K.; Fidler, S.; Torralba, A. BigDatasetGAN: Synthesizing ImageNet with Pixel-wise Annotations. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 21330–21340. [CrossRef]
11. Wu, W.; Zhao, Y.; Shou, M.Z.; Zhou, H.; Shen, C. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 1206–1217. [CrossRef]
12. Chen, K.; Liu, C.; Chen, H.; Zhang, H.; Li, W.; Zou, Z.; Shi, Z. RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4701117. [CrossRef]
13. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Processing Syst.* **2017**, *30*. [CrossRef] [PubMed]
14. Jaegle, A.; Gimeno, F.; Brock, A.; Vinyals, O.; Zisserman, A.; Carreira, J. Perceiver: General perception with iterative attention. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 4651–4664. [CrossRef]
15. Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F.; Wattenberg, M. Smoothgrad: Removing noise by adding noise. *arXiv* **2017**, arXiv:1706.03825.
16. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851. [CrossRef]
17. Krähenbühl, P.; Koltun, V. Efficient inference in fully connected crfs with gaussian edge potentials. *Adv. Neural Inf. Process. Syst.* **2011**, *24*, 109–117. Available online: https://dl.acm.org/doi/10.5555/2986459.2986472 (accessed on 22 April 2024).
18. Waqas Zamir, S.; Arora, A.; Gupta, A.; Khan, S.; Sun, G.; Shahbaz Khan, F.; Zhu, F.; Shao, L.; Xia, G.S.; Bai, X. 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, Long Beach, CA, USA, 16–17 June 2019; pp. 28–37. Available online: https://api.semanticscholar.org/CorpusID:170079084 (accessed on 22 April 2024).

19. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983. [CrossRef]
20. Cheng, G.; Han, J.; Zhou, P.; Guo, L. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS J. Photogramm. Remote Sens.* **2014**, *98*, 119–132. [CrossRef]
21. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 270–279. [CrossRef]
22. Xia, G.S.; Yang, W.; Delon, J.; Gousseau, Y.; Sun, H.; Maître, H. Structural high-resolution satellite image indexing. In Proceedings of the ISPRS TC VII Symposium-100 Years ISPRS, Vienna, Austria, 5–7 July 2010; Volume 38, pp. 298–303. Available online: https://api.semanticscholar.org/CorpusID:18018842 (accessed on 22 April 2024).
23. Dai, D.; Yang, W. Satellite Image Classification via Two-Layer Sparse Coding with Biased Image Representation. *IEEE Geosci. Remote Sens. Lett.* **2011**, *8*, 173–176. [CrossRef]
24. Zou, Q.; Ni, L.; Zhang, T.; Wang, Q. Deep learning based feature selection for remote sensing scene classification. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 2321–2325. [CrossRef]
25. Zhao, L.; Tang, P.; Huo, L. Feature significance-based multibag-of-visual-words model for remote sensing image scene classification. *J. Appl. Remote Sens.* **2016**, *10*, 035004. [CrossRef]
26. Cheng, G.; Han, J.; Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
27. Long, Y.; Gong, Y.; Xiao, Z.; Liu, Q. Accurate object localization in remote sensing images based on convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2486–2498. [CrossRef]
28. Xiao, Z.; Long, Y.; Li, D.; Wei, C.; Tang, G.; Liu, J. High-resolution remote sensing image retrieval based on CNNs from a dimensional perspective. *Remote Sens.* **2017**, *9*, 725. [CrossRef]
29. Zhou, W.; Newsam, S.; Li, C.; Shao, Z. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS J. Photogramm. Remote* **2018**, *145*, 197–209. [CrossRef]
30. Wang, Q.; Liu, S.; Chanussot, J.; Li, X. Scene classification with recurrent attention of VHR remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 1155–1167. [CrossRef]
31. Li, H.; Jiang, H.; Gu, X.; Peng, J.; Li, W.; Hong, L.; Tao, C. CLRS: Continual learning benchmark for remote sensing image scene classification. *Sensors* **2020**, *20*, 1226. [CrossRef]
32. Liu, K.; Mattyus, G. Fast Multiclass Vehicle Detection on Aerial Images. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 1938–1942. [CrossRef]
33. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2961–2969. [CrossRef]
34. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-attention mask transformer for universal image segmentation. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 1290–1299. [CrossRef]
35. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929. [CrossRef]
36. Kolesnikov, A.; Lampert, C.H. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part IV 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 695–711.
37. Ahn, J.; Kwak, S. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4981–4990. [CrossRef]
38. Kim, B.; Yoo, Y.; Rhee, C.E.; Kim, J. Beyond semantic to instance segmentation: Weakly-supervised instance segmentation via semantic knowledge transfer and self-refinement. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 4278–4287. [CrossRef]
39. Dai, J.; He, K.; Sun, J. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In Proceedings of the 015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1635–1643. [CrossRef]
40. Chen, M.; Zhang, Y.; Chen, E.; Hu, Y.; Xie, Y.; Pan, Z. Meta-Knowledge Guided Weakly Supervised Instance Segmentation for Optical and SAR Image Interpretation. *Remote Sens.* **2023**, *15*, 2357. [CrossRef]
41. Kirillov, A.; Wu, Y.; He, K.; Girshick, R. Pointrend: Image segmentation as rendering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9799–9808.
42. Chen, H.; Sun, K.; Tian, Z.; Shen, C.; Huang, Y.; Yan, Y. Blendmask: Top-down meets bottom-up for instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8573–8581.

43. Liu, Y.; Li, H.; Hu, C.; Luo, S.; Luo, Y.; Chen, C.W. Learning to aggregate multi-scale context for instance segmentation in remote sensing images. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, 1–15 (Early Access). [CrossRef] [PubMed]
44. Lu, X.; Wang, B.; Zheng, X.; Li, X. Exploring Models and Data for Remote Sensing Image Caption Generation. *IEEE Trans. Geosci. Remote Sens.* **2017**, *56*, 2183–2195. [CrossRef]
45. Qu, B.; Li, X.; Tao, D.; Lu, X. Deep semantic understanding of high resolution remote sensing image. In Proceedings of the 2016 International Conference on Computer, Information and Telecommunication Systems (Cits), Kunming, China, 6–8 July 2016; pp. 1–5. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.