remote sensing

MDPI

Article

Label-Efficient Fine-Tuning for Remote Sensing Imagery Segmentation with Diffusion Models

Yiyun Luo [1], Jinnian Wang [2,3,*], Jean Sequeira [4], Xiankun Yang [2,3], Dakang Wang [2,3], Jiabin Liu [2,5], Grekou Yao [1] and Sébastien Mavromatis [1]

[1] G-Mod, LIS, CNRS, Aix-Marseille University, 13009 Marseille, France; yiyun.luo@etu.univ-amu.fr (Y.L.)
[2] Institute of Aerospace Remote Sensing Innovations, Guangzhou University, Guangzhou 510006, China
[3] School of Geography and Remote Sensing, Guangzhou University, Guangzhou 510006, China
[4] 2ik Company, 13360 Marseille, France
[5] Faculty of Geography, Yunnan Normal University, Kunming 650500, China
* Correspondence: jnwang@gzhu.edu.cn

Abstract

High-resolution remote sensing imagery plays an essential role in urban management and environmental monitoring, providing detailed insights for applications ranging from land cover mapping to disaster response. Semantic segmentation methods are among the most effective techniques for comprehensive land cover mapping, and they commonly employ ImageNet-based pre-training semantics. However, traditional fine-tuning processes exhibit poor transferability across different downstream tasks and require large amounts of labeled data. To address these challenges, we introduce Denoising Diffusion Probabilistic Models (DDPMs) as a generative pre-training approach for semantic features extraction in remote sensing imagery. We pre-trained a DDPM on extensive unlabeled imagery, obtaining features at multiple noise levels and resolutions. In order to integrate and optimize these features efficiently, we designed a multi-layer perceptron module with residual connections. It performs channel-wise optimization to suppress feature redundancy and refine representations. Additionally, we froze the feature extractor during fine-tuning. This strategy significantly reduces computational consumption and facilitates fast transfer and deployment across various interpretation tasks on homogeneous imagery. Our comprehensive evaluation on the sparsely labeled dataset MiniFrance-S and the fully labeled Gaofen Image Dataset achieved mean intersection over union scores of 42.7% and 66.5%, respectively, outperforming previous works. This demonstrates that our approach effectively reduces reliance on labeled imagery and increases transferability to downstream remote sensing tasks.

Keywords: remote sensing imagery segmentation; fine-tuning; denoising diffusion probabilistic model; self-supervised learning

check for updates

Academic Editors: Li Fang, Jian Yang and Yu Feng

Received: 20 June 2025
Revised: 15 July 2025
Accepted: 18 July 2025
Published: 24 July 2025

Citation: Luo, Y.; Wang, J.; Sequeira, J.; Yang, X.; Wang, D.; Liu, J.; Yao, G.; Mavromatis, S. Label-Efficient Fine-Tuning for Remote Sensing Imagery Segmentation with Diffusion Models. *Remote Sens.* **2025**, *17*, 2579. https://doi.org/10.3390/rs17152579

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Remote sensing imagery (RSI) segmentation is critical for remote sensing interpretation, as it enables thematic mapping by delineating pixel properties [1]. Diverse classification criteria serve specific applications: Urban authorities target spatial elements like buildings and roads, while environmental agencies analyze forests, grasslands, and water bodies to track environmental trends. Additionally, hazards such as floods, forest fires, and earthquakes can be detected to assess societal impacts. Efficient and accurate imagery

interpretation is important for monitoring human activities in specific regions, tracking natural disasters, and understanding global-scale ecological transformations [2].

Initial remote sensing segmentation methods relied primarily on spectral reflectance features to differentiate between objects. Machine learning later enhanced interpretation accuracy by incorporating texture and contextual features through manual feature selection and classifier design [3]. However, scaling these methods across large, multi-sensor datasets has remained a significant challenge [4]. Advances in satellite and unmanned aerial vehicle (UAV) technology have since enabled high-quality, large-scale remote sensing data with high spatial-temporal resolution, becoming essential for land surface monitoring [5–7]. Data abundance has facilitated data-driven deep learning approaches, which leverage large labeled datasets to achieve superior accuracy and robustness over traditional methods, enabling precise end-to-end segmentation aligned with training data distributions [8,9].

As shown in Figure 1a, the methodology for remote sensing imagery segmentation typically involves a feature encoder and a decoder with a segmentor [10]. Feature encoders are pre-trained with the classifier to initialize weights with knowledge. The classifier is then replaced by a segmentor to enable pixel-level learning [11]. Due to the significant distribution gap between remote sensing datasets and the commonly used pre-training dataset, ImageNet [12], specialized pre-training datasets like BigEarthNet [13] and MillionAID [14] have been developed. Nevertheless, the requirement for over a million patch-level annotations makes these datasets expensive to construct, significantly increasing the demand for manually labeled data from domain experts [15]. In this context, strategically utilizing unlabeled imagery to enhance model generalization, refine feature representations, and reduce reliance on labeled data becomes crucial for effective segmentation [16]. In response to this imperative, self-supervised learning (SSL) methods have become an alternative to image classification pre-training [17].

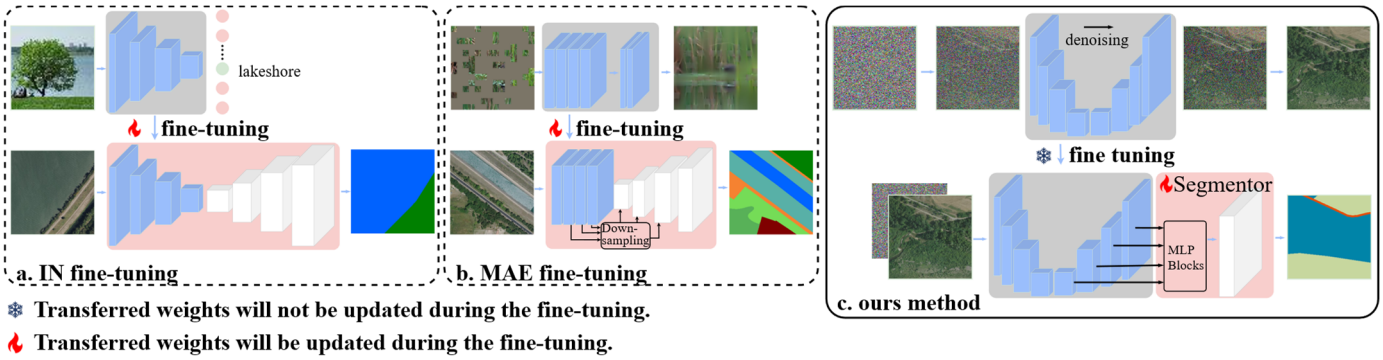

Figure 1. Fine-tuning pipeline comparison. Pipeline (**a**) is IN fine-tuning, pre-trained by image classification using the ImageNet dataset. Pipeline (**b**) is MAE fine-tuning, pre-trained by reconstructing images from mask tokens. Pipeline (**c**) is our method, pre-trained by DDPM.

SatViT [18] collected 1.3 million remote sensing images from various sensors and pre-trained a Vision Transformer (ViT), as illustrated in the pipeline in Figure 1b, using Masked Autoencoding (MAE) [19]. ScaleMAE [20] and the Masked Angle-Aware Autoencoder (MA3E) [21] leverage geographic-specific information to bridge the semantic gap between RSI and natural images. Meanwhile, DDPMs [22] have proven highly effective in generating Earth observation imagery, outperforming previous methods like Generative Variational Autoencoders (VAEs) [23] in generation quality. Therefore, diffusion-based approaches have been rapidly adopted for a range of downstream tasks in both natural and remote sensing imagery interpretation. The Lightweight Diffusion Model (LWTDM) [24] balances inference speed and perceptual quality in remote sensing imagery super-resolution tasks through a simplified diffusion model. Similarly, DiffSeg [25] explores land use classification

by constraining the diffusion process to generate segmentation masks. DDPM for Change Detection (DDPM-CD) [26], on the other hand, demonstrates outstanding performance by employing the diffusion model as a feature extractor for change detection.

The excellent works above demonstrate the potential of diffusion models for feature extraction in remote sensing applications. Meanwhile, the following challenges have arisen: (1) leveraging unsupervised features to reduce the requirement for labeled samples during fine-tuning and (2) accelerating fine-tuning in downstream tasks while maintaining high accuracy. To tackle these challenges, we designed a novel paradigm for remote sensing imagery segmentation. As illustrated in Figure 1c, we utilize unlabeled remote sensing imageries for generative pre-training and then train the segmentor using multi-scale features from the diffusion model decoder. Unlike preceding pre-training methods, such as classification pre-training, which aims at predicting patch-level objects, the denoised images output by the diffusion model retain the resolution of the input image. This allows the diffusion model decoder to extract the semantics of the objects after pre-training.

Based on this, we have made the following contributions.

- Label-efficient fine-tuning: We propose a pre-training strategy on homogeneous unlabeled datasets to enhance the DDPM encoder. This strategy helps perform label-efficient fine-tuning, based on which we developed two self-supervised learning datasets based on homogeneous images from MiniFrance and the Gaofen Image Dataset (GID).
- Multi-scale features analysis from the unsupervised decoder: We activate and visualize intermediate-layer features, comparing the features extracted by the diffusion model's encoder and decoder. The curated features from the DDPM decoder are then used for remote sensing imagery segmentation.
- Scheduled noisy imagery input for dense prediction: We implement a noise-based augmentation strategy to enhance the generalization of diffusion models. Our experiments show that injecting a controlled proportion of Gaussian noise can effectively improve segmentation accuracy.
- Multi-layer perceptron (MLP) segmentor for frozen backbones: We design a simple MLP segmentor that effectively fuses and optimizes multi-scale features obtained from the frozen diffusion decoder, facilitating an easy and effective transfer to downstream segmentation tasks.

The rest of this paper is organized as follows. Section 2 presents selected works related to our methodology. Section 3 describes our proposed model for label-efficient fine-tuning. Section 4 reports the details, results, and analyses of our experiments. At the end, Section 5 provides our conclusion to this paper and perspectives on our future work.

2. Related Works

2.1. Pre-Training with Diffusion Models

2.1.1. Generative Pre-Training

General image features have been extracted by pre-training [27]. This accelerates the training during fine-tuning for specific downstream tasks, fosters the model convergence, and improves predictive accuracy [28,29]. ImageNet [12], as the largest dataset for image classification, provides an extensive representation of image dataset distribution, making it a popular choice for pre-training to initialize model weights. Nevertheless, as more advanced data augmentation techniques emerge and the cost of labeling grows, researchers are reevaluating the practicality of image classification pre-training [13,30]. Meanwhile, self-supervised learning, with its low dependency on labeled data and robust features, has emerged as a new option for pre-training [31].

Generative adversarial networks estimate the distributions of real-world data and implicitly calculate the similarity between them and the truth [32,33]. However, the pre-training these networks perform usually only serves them and is difficult to apply to broader real-world tasks. Auto-Encoders (AEs) are pioneering self-supervised methods in pre-training models [34–36]. AEs generate images based on latent variables obtained from the encoder. The objective function of AEs is to optimize the likelihood of generating output x by decoding from the latent variables. In $p(x) = \int p(z)p(x \mid z;\theta)dz$, z is the latent variable, θ is a parameterized vector, and the function $f(z;\theta)$ between z and θ is expressed by the conditional probability $p(x \mid z;\theta)$. This allows the total probability formula to specify the dependence of x on z. However, the latent variables z obtained from AEs lack the ability to generalize in large and complex scenarios. To address this, some methods add noise into the encoder to obtain stronger generalization capabilities. Recent methods [37–39] introduce an approximate posterior $p(x \mid z;\varphi)$ of the latent variable z to simulate the intractable truth posterior distribution $p(x \mid z;\theta)$. They achieve this through a reparametric trick that converts the encoder, which generates a latent representation z_i into the optimization of a variational lower bound $\mathcal{L}(\theta, \varphi; x_i)$ on the samples, thereby stabilizing the training of the model. Self-supervised pre-training models [40,41] significantly improve the efficiency and generalization of the models on visual tasks. However, the feature distributions modeled by AEs are still too complex to stabilize the quality of each generated sample while training the deep hierarchical model.

2.1.2. Diffusion Models

Denoising diffusion models [8,24–26,41–43] decompose the process of parameterization into T steps of Gaussian noise addition. The influence of noise on the input data at different steps is determined by the hyperparameters, variance schedule $\beta_1 \ldots \beta_T$. Meanwhile, the process of decoding becomes sampling from a series of latent variables with the same dimensionality. It is formed as $p_\theta(x_0) = \int p_\theta(x_{0:T})dx_{0:T}$, where $x_1,\ldots, x_T$ are variables with the same dimensionality and distribution as x_0. The joint function $p_\theta(x_{0:T})$ is defined as a Markov chain of Gaussian transitions, starting from x_T, where $p(x_T) = \mathcal{N}(x_T; 0, I)$. As shown in Figure 2, the approximate posterior distribution $q(x_{1:T}|x_0)$ of the diffusion model is fixed to a forward Markov chain. The latent variable is obtained by adding Gaussian noise with progressively increasing variance β. The posterior distribution of x_t given x_{t-1} is denoted as:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I\right), \tag{1}$$

where $\beta_1, ..., \beta_T$ is a noise schedule that increases between 0 and 1, introducing varying degrees of Gaussian noise for the asynchrony. Typically, either a linear or a cosine schedule is applied. Therefore, forward samples x_t with $x_t \sim q(x_t \mid x_{t-1})$ can be formulated as follows:

$$x_t = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\epsilon_t, \; with \; \epsilon_t \sim N(\epsilon; 0, I). \tag{2}$$

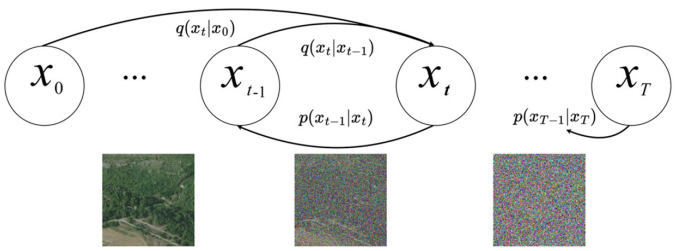

Figure 2. The visualized variational diffusion processes x_0 is the truth sample, and x_t is the pure noise conforming to a Gaussian distribution.

Defining the parameters $\alpha_t = 1 - \beta_t$ gives $\overline{\alpha_t} = \prod_{s=1}^{t} \alpha_s$. In the forward process, it becomes feasible to directly sample x_t at an arbitrary forward step based on the original data x_0. We get the posterior distribution of x_t given x_0:

$$q(x_t \mid x_0) = N\left(x_t\,;\ \sqrt{\overline{\alpha}_t}x_0, (1-\overline{\alpha}_t)I\right). \tag{3}$$

Thus, forward samples x_t with $x_t \sim q(x_t \mid x_0)$ can be formulated as:

$$x_t = \sqrt{\overline{\alpha_t}}x_{t-1} + \sqrt{1-\overline{\alpha_t}}\epsilon_0, \tag{4}$$

where ϵ_0 in Equation (4) and ϵ in Equation (2) are independent and identically distributed and follow standard normal distributions.

The estimation of Gaussian distribution $p_\theta(x_{t-1} \mid \mathrm{x_t}) = \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, t), \sigma^2 I\right)$ at the reverse process by a parameterized Markov chain, where the variance σ^2 is a constant and the mean $\mu_\theta(x_t, t)$ of the distribution, is estimated by the model f_θ. This model is optimized by calculating a variational bound via the negative log-likelihood. It can be simply derived as:

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathrm{x_0},\epsilon}[\parallel \mathrm{f}(\theta; \mathrm{x_t}, \mathrm{t}) - \epsilon \parallel_2^2], \text{ with } \epsilon = \mathcal{N}(0, I), \tag{5}$$

where the optimization objective $\mathcal{L}(\theta)$ is subject to the mathematical expectation $\mathbb{E}_{x_0,\epsilon}$, where x_0 and ϵ are considered known. Here, ϵ represents the noise conforming to the Gaussian distribution for the t step in the Markov chain. The gradient descent optimization of $\mathcal{L}(\theta)$ actually optimizes the prediction process of the noise ϵ.

With a total number of T sampling steps, the process from x_t sampling to x_{t-1} $(T \geq t > 0)$ can be formulated as:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(x_t, t)\right) + \sigma_t z, \tag{6}$$

where ϵ_θ is the function approximator, with progressive denoising starting from random Gaussian noise x_t, and $z \sim N\ (0,\ I)$.

2.2. Intermediate Features for Semantic Segmentation

High-quality multi-scale intermediate features are key to obtain excellent dense prediction results [44]. The concept of semantic segmentation using convolutional neural networks (CNNs) was popularized by Fully Convolutional Networks (FCNs) [11]. FCNs replace fully connected layers with convolutional layers, enabling the model to output segmentation masks. Early approaches prove that intermediate features could be effectively combined to get dense prediction [45], thus laying the groundwork for many later methods [46]. However, upsampling directly from the bottom features to obtain the prediction mask is unsatisfactory in predicting the details of the image. Subsequently, some methods fuse and refine multi-resolution features into the upsampling process as a way to retain fine-grained details while understanding the object semantics [47].

Unsupervised learning has been widely used to extract intermediate features for various dense prediction tasks [42,48]. Contrastive learning has been gradually adopted from tile- and image-level feature extraction for pixel-level feature extraction owing to stable training and faster convergence. Due to the limitations of the data augmentation methods used in contrastive learning, the extracted features from pre-trained models transfer poorly to distributions with large gaps [49]. Generative pre-training approaches aim for feature learning by recovering or generating from unlabeled data [50]. MAE [19] demonstrates a powerful feature learning capability as a pre-task by reconstructing masked patches. If it uses the Bidirectional Encoder Image Transformers (BEiT) [51] method to encode tokenized images, multiscale intermediate features from different patches can be extracted

for dense prediction. MaskDM [43] introduces the idea of patch masks for improving the training efficiency of diffusion models. DiffMAE [52] formulates the diffusion model as a condition for BEiT autoencoder inputs, thereby enabling transfer to downstream tasks and high-quality image reconstruction. The Masked Diffusion Model (MDM) [53] substitutes the masking mechanism with additive Gaussian noise in diffusion modeling and extracts features from the decoder with a fine-tuned segmentor. MA3E [21] introduces the explicit angle variation to learn rotation-invariant representations.

2.3. Diffusion Models in Remote Sensing

Diffusion models have created significant ripples in the realm of image generation, yet their vast potential for application in downstream tasks remains largely unexplored. Remote sensing image perception works based on diffusion models can be broadly categorized into two groups. The first is conditional generation with downstream objectives [54], which trains a conditional diffusion model end-to-end during a single training phase, enabling direct prediction without additional processing [8]. The second is unconditional pre-training followed by fine-tuning, where features extracted from the pre-trained model serve as prior knowledge for various tasks [17,42].

Additionally, diffusion models are widely applied in remote sensing for imagery generation, augmentation, and interpretation [25]. Satellite image generation mainly involves text-to-image and image-to-image processes to produce diverse map visualizations [55,56]. For image augmentation, diffusion models are utilized in super-resolution, cloud removal, and denoising tasks [57]. In remote sensing interpretation, they support dense prediction tasks such as land cover classification and change detection [58]. For instance, SegDiff [8] directly combines the results obtained from diffusion models into a final segmentation map. Multi-Class Segmentation (MCS) [59] refines predicted segmentation maps by comparing them with RGB images and calculating losses against the ground truth. SpectralDiff [60] improves classification performance on multispectral and hyperspectral data by incorporating channel attention, and the Background Suppression Diffusion Model (BSDM) [61] achieves high-performance hyperspectral anomaly detection by treating background information as noise.

Our study demonstrates that diffusion models trained on remote sensing data can extract generalizable features across diverse land cover classes and scenes. These self-supervised features seamlessly transfer to downstream tasks without fine-tuning the entire model. Experiments show that pre-training on diverse scenes enables strong segmentation performance with limited labels. We pre-train diffusion models and fine-tune a lightweight segmentor on remote sensing datasets. To further improve the accuracy in the remote sensing segmentation, we adopt a secondary pre-training strategy. In this strategy, the model is first trained on a large-scale unlabeled dataset, followed by domain-specific adaptation, significantly reducing the time required for model training. Although we did not obtain refined feature representations for specific categories in the generated images, the majority of land cover categories of interest were well generated. These categories provided excellent feature representations during the fine-tuning of the segmentor for multiple foreground categories, even after freezing the pre-training parameters.

3. Methodology

Diffusion models have been highly successful in image generation, demonstrating their ability to accurately model feature distributions for dense prediction [62]. This capability is relevant to tasks involving semantic segmentation, where dense estimation is essential. Therefore, we propose a fine-tuning approach based on diffusion models as an alternative to the conventional pre-training paradigm for remote sensing imagery

segmentation and design a novel segmentor that can efficiently utilize features obtained from diffusion models. As shown in Figure 3, we obtain a total of 12 layers of features from the pre-trained diffusion model. These muti-scale features are acquired at different noise levels, integrated, and optimized in the segmentor using channel-wise MLP blocks. Then, processed features at different resolutions are uniformly upsampled to the $\frac{1}{4}$ of the original image resolution to generate the final segmentation results.

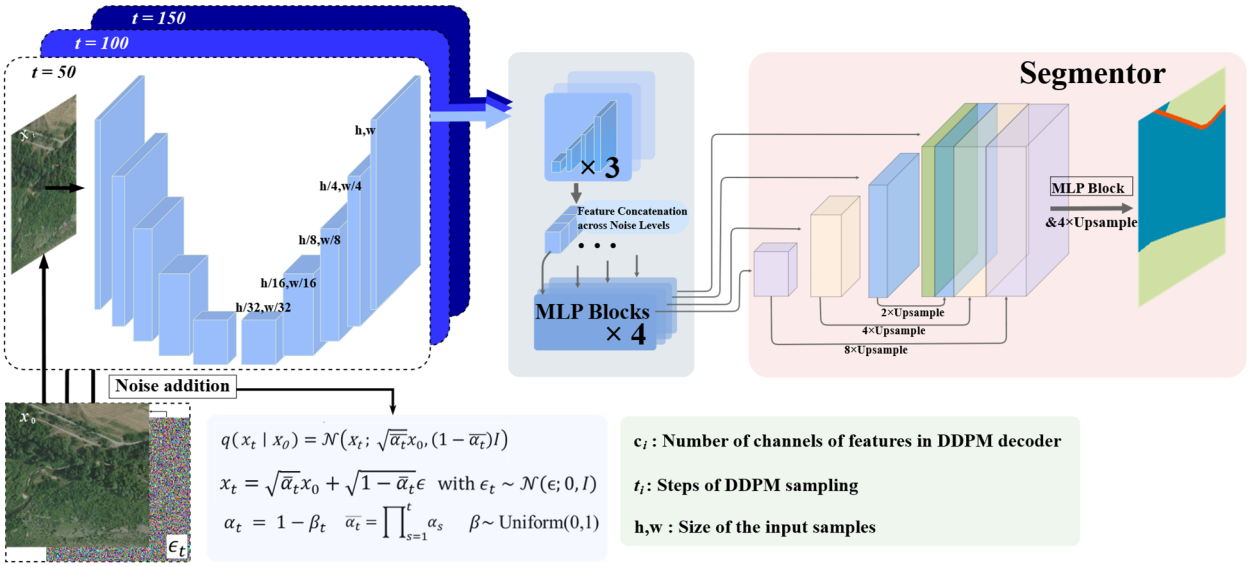

Figure 3. Framework of our model for semantic segmentation of remote sensing imagery using the diffusion model for MLP segmentor fine-tuning.

3.1. A Generative Pre-Training Paradigm for Semantic Segmentation

Our proposed method explores the feasibility of DDPM for dense pixel-level image generation of complex objects and for pre-training remote sensing image segmentation models. Typically, the process of remote sensing images segmentation consists of two main phases: pre-training and fine-tuning.

We consider the objective function for pre-training using the classification model with the formula:

$$\mathcal{L}_{cls}(\theta_p) = \mathbb{E}_{(X,Y)\sim P_p}\left[f(\theta_p; X, Y)\right], \tag{7}$$

where θ_p denotes the parameters of the neural network, (X, Y) are image-label pairs from the pre-training dataset, which follows the distribution P_p, and $f(\cdot)$ is the classification loss function. For clarity, we use $\mathbb{E}$, the expectation, to denote the target loss calculation of the pairs (X, Y).

The main goal of performing pre-training is to optimize θ_p with robust feature extraction ability. The training objective, fine-tuned on the semantic segmentation dataset based on θ_p, is defined by:

$$\mathcal{L}_{seg}\left(\theta_f\right) = \mathbb{E}_{(X',Y')\sim P_f}\left[f\left(\theta_f; X', Y'\right)\right], \tag{8}$$

where θ_f is initialized from θ_p, and (X', Y') denotes samples from the segmentation data distribution P_f. In traditional settings, this step involves updating the entire network, including both the encoder and decoder.

In contrast, our proposed approach freezes the feature extractor obtained from unsupervised generative pre-training, and only the segmentation head is fine-tuned. The fine-tuning objective becomes:

$$\mathcal{L}_{seg}\left(\theta_f\right) = \mathbb{E}_{E(X'),Y'}\left[f\left(\theta_f; E(X'), Y'\right)\right], \tag{9}$$

where $E(X')$ denotes the features extracted by the frozen encoder. In contrast to conventional generative learning, DDPMs generate images by predicting noise in a stepwise manner. Notably, similar image features are represented differently across varying noise levels, with the noise level being controlled by the diffusion step t. Therefore, the objective function of the method proposed in this paper for DDPM pre-training is defined by:

$$\mathcal{L}_{gen}(X'_t) = \mathbb{E}_{X_{0\sim P_f}}\left[f\left(\theta_g; E(X'_t), t\right)\right], \tag{10}$$

where the input of fine-tuning $E(X'_t)$ is affected simultaneously by the variables x and t, and when $t = 0$, there is no noise added to the image X'_0.

3.2. Fine-Tuning Strategy

Pre-training is a kind of pretext task, indicating that the training process is not aligned with the user's genuine goals. However, it aims to understand the universal features of the target dataset [63]. Unsupervised learning and self-supervised learning simplify the process of data acquisition by eliminating the necessity of using manual labels in the feature extraction process, and they are more capable of learning deep representations than supervised learning [64]. As shown in the upper part of Figure 4, in remote sensing imagery interpretation, extracting features from large-scale archived imagery using a self-supervised model is advantageous. These features are transferred to the downstream task with frozen parameters to achieve label-efficient fine-tuning, as depicted in the lower part of Figure 4.

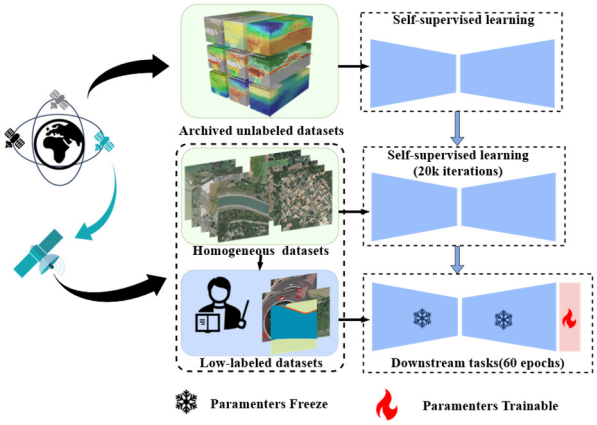

Figure 4. Illustration of the two-step unsupervised pre-training and fine-tuning process towards the downstream task.

Given the spatial similarity in the specific semantic features of remote sensing images, we employ a model pre-trained on very high-resolution satellite imagery sourced from Google Earth Engine [7], followed by training on our specific datasets. To leverage the powerful generalization capabilities of DDPM and achieve optimal semantic segmentation results, as shown in Figure 4, we built a three-step fine-tuning process for remote sensing tasks. First, perform unsupervised fine-tuning on the pre-trained diffusion model using the training set. Second, extract intermediate representations from the DDPM. The training iteration ratio between the large-scale dataset and the specific dataset is set at 1:1. Finally, train a high-performance segmentation model with a simple segmentor using these intermediate representations. In comparison to traditional pre-training methods, generative pre-training does not need to update all the parameters, thereby improving training speed and data utilization efficiency.

3.3. Feature Selection

Intermediate representations are typically derived from the encoder component of a classification model that has been pre-trained on ImageNet [65]. Due to the considerable computational expense associated with recognizing high-resolution features, generating pixel-level predictions from the decoder presents a greater challenge compared to obtaining patch-level predictions from the encoder during pre-training on extensive datasets [66]. Meanwhile, pre-training in classification tasks focuses excessively on the global features of the image. This diminishes the ability to extract the contextual features that are essential in segmentation tasks. In our study, the intermediate representations of DDPM are carefully analyzed. Although the training objective in generative learning is not related to semantic categories, some studies have shown that in the latent feature space of unconditional generative models, it is possible to distinguish foreground and background pixels as well as specific semantic features in the samples, even without supervisory information about labels [67]. In order to test the feature representativeness learned in different datasets and different pre-training methods, we use T-SNE to visualize the features expressed by different classes of pixels. As shown in Figure 5, the features pre-trained without labels in the encoder and decoder parts of the DDPM can discriminate pixels with similar semantics without fine-tuning. Among all T-SNE visualization results, the DDPM decoder achieves the best feature separation, indicating that DDPM pre-training is able to distinguish different pixel classes to a certain extent before being transferred to the downstream task. In light of this, we extracted semantic features from different layers of the DDPM decoder and adapted them to different downstream tasks.

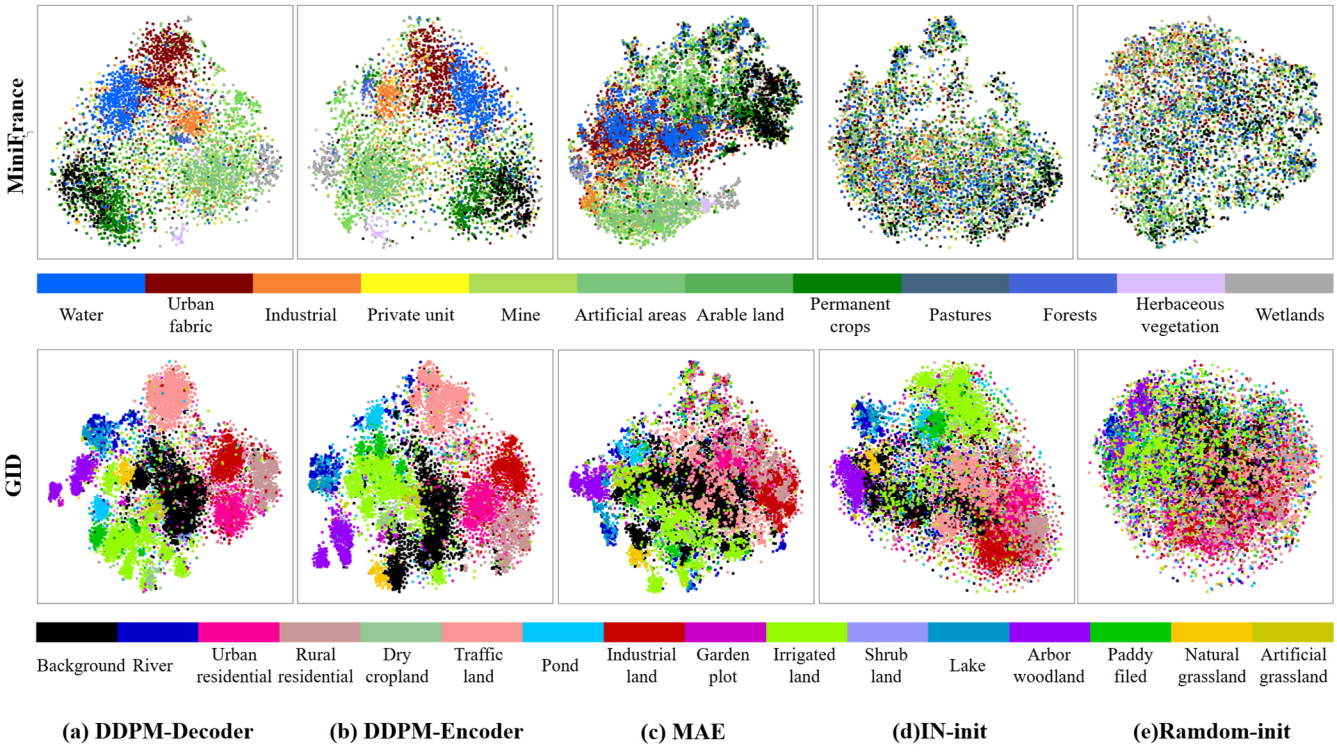

Figure 5. TSNE visualization of intermediate features for different pre-training methods.

Additionally, as DDPM operates as a progressive denoising model, $f_\theta(x_t, t)$ is a prediction of the noise in step t. In our feature extraction of an original image sample, the presence of a certain level of noise can effectively enhance the extraction of features at different scales because the interclass similarity of features between different species affects the recognition of semantic features, thereby affecting the segmentation accuracy [68]. As

shown in Figure 6, we extracted the feature maps with {4, 8, 16, 32} times down-sampling in the encoder and decoder, respectively, and clustered them by k-means in different noise ratios. We found that the semantic objects of the target category remain coherent under unsupervised learning.

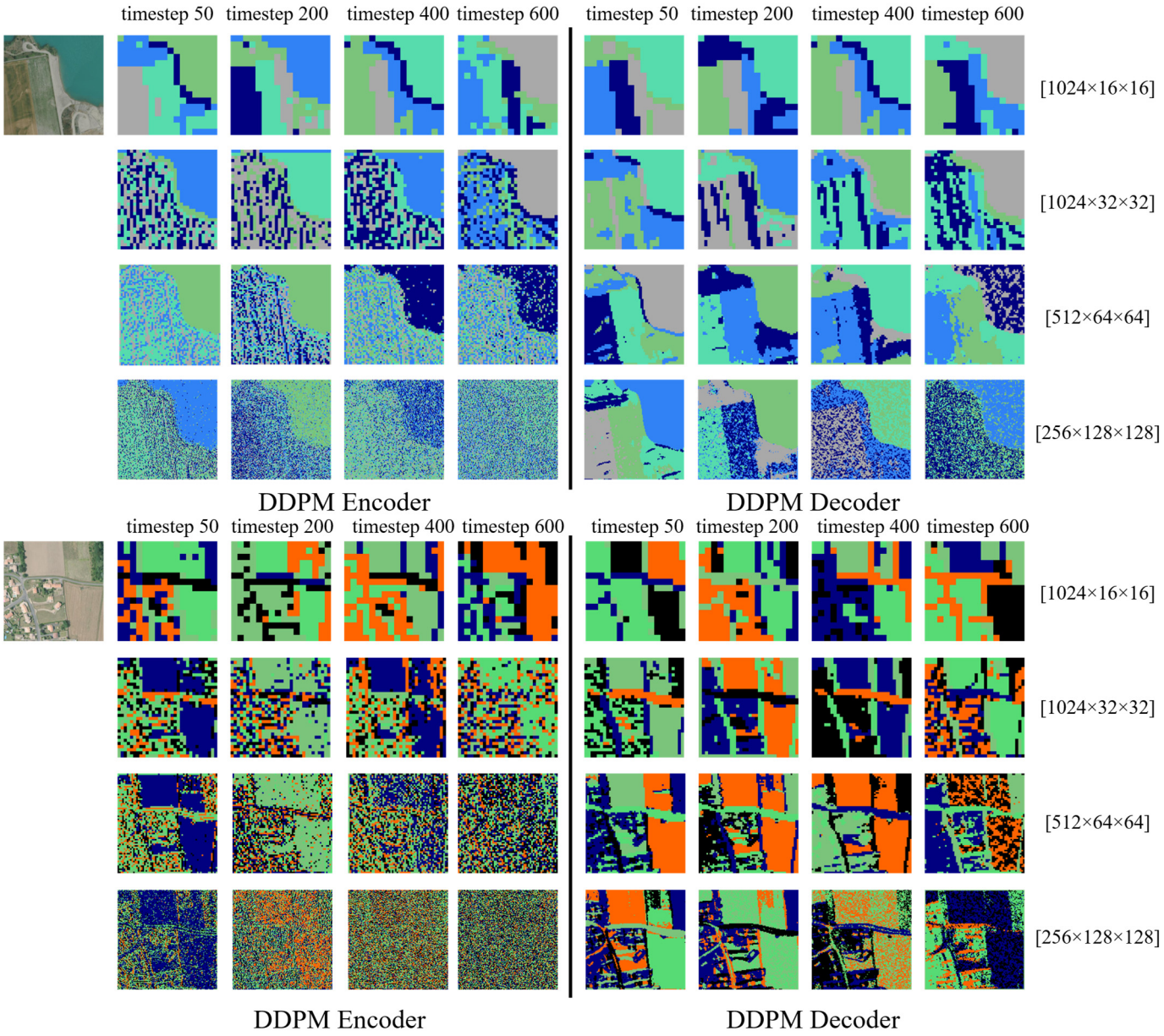

Figure 6. Visualization of k-means results for the intermediate features in samples of the MiniFrance test set, with sample resolution of 256 × 256.

Consequently, we extracted three sets of feature maps with different noise levels using the decoder in the UNet model instead of choosing noise-free images for training. Intermediate representations from 12 different layers were extracted for each sample at different noise steps {50, 100, 150} for subsequent fine-tuning.

3.4. MLP-Based Segmentor Architecture

The analysis of the intermediate feature representations presented above indicates that the latent features extracted by the unconditional generative model already exhibit excellent semantic expressiveness and capture relevant geometric structures. To obtain more robust representations under varying noise conditions, features at the same spa-

tial resolution but extracted from different noise levels are concatenated. Compared to pixel-wise addition, feature concatenation retains a richer set of representations, enhancing the segmentor's ability to express the features across noise variations. However, concatenation also introduces redundancy, increasing both memory consumption and computational complexity.

Building on this, the segmentor is designed based on a lightweight MLP architecture, aiming to optimize feature selection for specific semantic classes while maintaining computational efficiency. The MLP segmentor refines the multi-scale feature maps extracted from the diffusion model decoder and produces dense, pixel-wise segmentation predictions for the target dataset.

As illustrated in Figure 7, each block within the segmentor consists of three linear layers, two batch normalization (BN) layers, and two non-linear activation functions. Furthermore, a residual connection is added before each BN layer to stabilize training and facilitate feature reuse. Unlike convolutional structures that operate over local spatial neighborhoods, the MLP segmentor performs transformations exclusively along the channel dimension, processing each spatial position independently. Specifically, given concatenated features at a resolution level, the segmentor maps $\mathbf{F}_{\text{input}} \in \mathbb{R}^{3c_i} \rightarrow \mathbf{F}_{\text{output}} \in \mathbb{R}^{c_i}$, where c_i denotes the number of output channels at the i-th resolution.

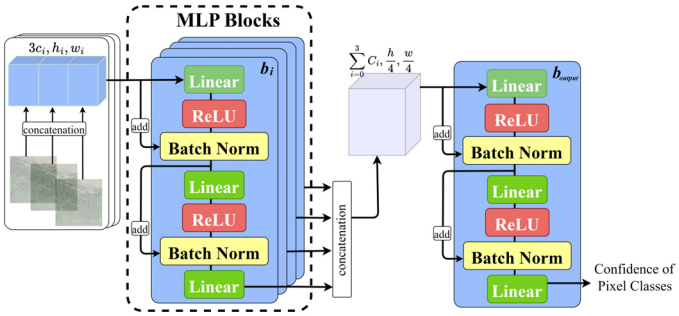

Figure 7. Architecture of the MLP-based segmentor used for label-efficient fine-tuning.

The computation inside each MLP block can be formulated as follows:

$$\begin{aligned} \mathbf{z}_1 &= \sigma\big(\mathrm{BN}_1\big(\mathrm{Linear}_1\big(\mathbf{F}_{\text{input}}\big) + \mathbf{F}_{\text{input}}\big)\big), \\ \mathbf{z}_2 &= \sigma(\mathrm{BN}_2(\mathrm{Linear}_2(\mathbf{z}_1) + \mathbf{z}_1)), \\ \mathbf{F}_{\text{output}} &= \mathrm{Linear}_3(\mathbf{z}_2), \end{aligned} \tag{11}$$

where $\sigma(\cdot)$ denotes the non-linear activation function ReLU. Here, each residual addition improves feature propagation and mitigates gradient vanishing issues, while BN layers normalize feature distributions to accelerate convergence. Following feature fusion across scales, the final dense prediction map is produced via an additional MLP block. In this final stage, the optimized features are first upsampled to $\frac{1}{4}$ of the input resolution and then concatenated. The final MLP projects these combined features into a per-pixel category probability map:

$$\mathbf{P} = \mathrm{Softmax}(\mathrm{MLP}(\mathrm{Concat}(\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3))), \tag{12}$$

where $\mathbf{P} \in \mathbb{R}^{H/4 \times W/4 \times C}$ denotes the pixel-wise class probability map with C categories.

The MLP-based segmentor effectively bridges the gap between generative pre-training and dense prediction, enabling scalable and efficient remote sensing imagery segmentation.

4. Experimental Results

4.1. Datasets

MiniFrance [69] is a large-scale, semi-supervised dataset comprising 14 categories and over two thousand very high-resolution Earth observation images collected from various regions across France. The key characteristic of this dataset is its diverse scenarios and land cover category tailored to different environments. As illustrated in Figure 8, although the urban structures in the three displayed images differ, they all belong to the urban fabric category.



Figure 8. Three different urban fabric structures in the MiniFrance dataset.

From the MiniFrance dataset, we extracted 488 aerial imageries with a resolution of 10,000 × 10,000 pixels, of which 100 were randomly selected as the unlabeled dataset for pre-training. The labeled imageries primarily cover two main types of scenes from four cities: Nantes and Saint-Nazaire, northwestern Atlantic coast cities dominated by plains and hills, and Marseille and Martigues, southern mountainous Mediterranean coastal cities. We randomly divide all tile maps into 256 × 256 resolution slices, allocating 40% to the training set, 10% to the validation set, and the remaining 50% to the test set. In addition, we divided the MiniFrance dataset into three different fine-tuned training sets, labeled MiniFrance-S, MiniFrance-M, and MiniFrance-L, comprising 2%, 40%, and 100% of the total training imagery, respectively.

GID [70] consists of 150 Gaofen-2 satellite imageries with pixel-level annotations. The dataset is organized into two tiers: a coarse land cover classification dataset with 5 categories and a fine land cover classification dataset with 15 categories, differentiated by their level of detail.

With this dataset, we take 140 images of 6800 × 7200 pixels with only coarse labels as homogeneous datasets for pre-training, and we fine-tune the segmentation task on 10 images with fine labels. GID training set images come from more than 30 different regions in China and are well characterized by their wide coverage, rich features, and high resolution. Actually, the GID data represent high spatial resolution satellite images, and models that perform well on the GID data can generalize well to different satellite imagery analysis tasks. Meanwhile, comparing satellite images and MiniFrance aerial images as two heterogeneous datasets can help explore the generalization ability of generative models on heterogeneous data.

4.2. Experimental Implementation

4.2.1. Pre-Training Setup

We used an A4000 graphics card for the two-step pre-training process of the DDPM model on the PyTorch 2.6.0 framework. We cropped the samples from the MiniFrance and GID datasets to ensure that the input size is 256 × 256 pixels. The parameters trained over 200k iterations on the large-scale remote sensing dataset were transferred, and training continued for an additional 20k iterations. The optimizer used was AdamW with the linear warm-up strategy. We started the learning rate at 1×10^{-6}, and after 10k iterations of warm-up, the learning rate was kept at 1×10^{-5}.

The four different pre-training paradigms compared in our experiments were random initialization, pre-training on ImageNet-1k classification tasks, pre-training with the unsupervised MAE method, and pre-training with diffusion models.

4.2.2. Hyperparameters

For segmentor fine-tuning, since the pre-training parts were frozen, obtaining decoder features offline and inputting them into the segmentor for prediction was considered. However, given the offline features required approximately about 60G of memory, the online generation method was applied. At this stage, the optimizer AdamW with an initial learning rate of 0.0002 and momentum decay of 0.0001 was applied, and we uniformly chose cosine annealing as the optimization strategy for the learning rate. Training was performed over 300 iterations. For comparison experiments, we applied a unified data augmentation method, including linear stretching in a random interval of 5–15%, $90 \times k$ (k = 1, 2, 3) random angle rotation, and image inversion.

4.2.3. Evaluation

We adopted intersection over union (IoU) to evaluate the performance of models. Within the semantic segmentation task, IoU measures the overlap between predictions and ground truth for each category. The mean IoU (mIoU) is defined as the ratio between the correctly predicted area (true positives) and the union of the predicted and ground truth areas:

$$mIoU = \frac{1}{C}\sum_{C=1}^{C}\left(\frac{TP_c}{TP_c + FP_c + FN_c}\right), \tag{13}$$

where TP_c, FP_c, and FN_c denote the number of true positives, false positives, and false negatives for class c, respectively.

Given that there is no consistency in the number of parameters activated in different pre-training models, and the number of parameters is recognized as an important metric affecting model performance, the sizes of the models were controlled across different baselines, and the number of parameters activated in different experiments was reported. Most parameters are derived from the convolution procedure. For a single convolution, the parameter count is calculated as $Params = \left(K^2 \times C_{in} + 1\right) \times C_{out}$, where K is the size of the convolution kernel, and C_{in} and C_{out} are the numbers of channels of the input features and output features, respectively.

4.3. Ablation Study

4.3.1. Two-Step Pre-Training

Given the accessibility of archived free satellite imagery for researchers, we implemented a two-step pre-training process for diffusion models, aiming to further reduce reliance on labeled data and to mitigate the impact of distributional shifts caused by differences in spectral, spatial, radiometric, and directional characteristics during transfer learning. Out of the 200k pre-training iterations, 180k were performed using archived

satellite images, followed by 20k iterations of homogeneous data training. Here, we define homogeneous data as imagery produced from the same sensor. As shown in Figure 9, with only 2% of labeled data in the MiniFrance dataset, the one-step pre-training on archived data achieved a 40.0% mIoU in the segmentation, which is 5.7% higher than pre-training using ImageNet classification. If an additional 20k iterations are trained with homogeneous data, the mIoU increases to 42.7%, aligning closely with the mIoU obtained using only homogeneous data.

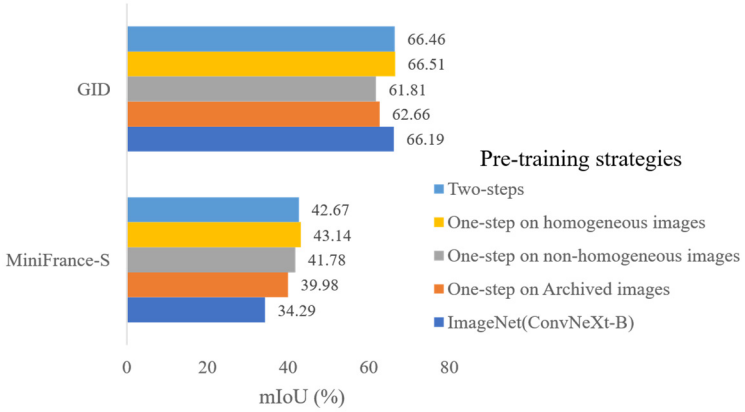

Figure 9. Comparison of pre-training strategies on remote sensing imagery segmentation.

Interestingly, even when using non-homogeneous data for learning the downstream task, the mIoU remains comparable to that of archived data alone, indicating that a small number of training iterations on non-homogeneous data do not adversely affect model generalization. The 180k iterations are sufficient to enable the diffusion model to learn generic object features in remote sensing imagery effectively. Additionally, as shown in Figure 10, we tested different percentages of training iterations for homogeneous data and found that dedicating 20k iterations to homogeneous data provides an optimal balance between datasets.

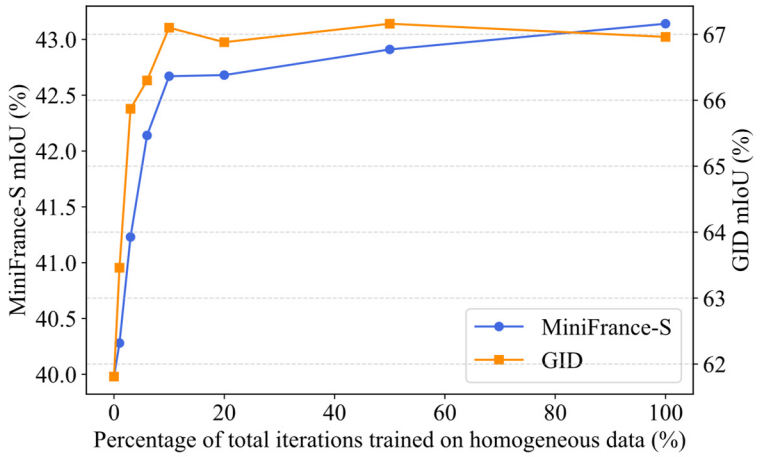

Figure 10. Effect of the percentage of training iterations on mIoU with homogeneous data in the MiniFrance-S and GID datasets.

4.3.2. Intermediate Features in Segmentor

Unlike classification tasks, segmentation tasks require dense prediction with a focus on contextual information and spatial features. Therefore, high-quality and high-resolution features are essential for improving segmentation accuracy. To fully leverage the high-resolution features extracted by diffusion models, we fused four different resolutions of features from the diffusion model decoder. As shown in Table 1, the best segmentation

results, with an mIoU of 42.7%, were achieved by integrating four solutions, $\frac{H}{16} \times \frac{W}{16}$, $\frac{H}{8} \times \frac{W}{8}$, $\frac{H}{4} \times \frac{W}{4}$, and $\frac{H}{2} \times \frac{W}{2}$, from the diffusion model decoders into the segmentor on the MiniFrance-S dataset. When using multi-resolution features output from the diffusion model encoders, the segmentor achieves an mIoU of 41%. Notably, as resolution increases, features in decoders become more representative until the $\frac{H}{2} \times \frac{W}{2}$ resolution, where too many irrelevant details are captured, reducing segmentation accuracy. The best single-scale result mIoU of 41% was obtained using the $\frac{H}{4} \times \frac{W}{4}$ resolution features from the decoder. Meanwhile, switching the segmentor from the MLP segmentor to the UPer head, which has a larger number of parameters, did not improve the mIoU when using the multi-resolution features from diffusion encoders. This indicates that the diffusion decoder is inherently effective at extracting semantic features and performing dense prediction. As a result, transfer learning after pre-training only requires a lightweight segmentor to guide a specific segmentation task excellently.

Table 1. Ablation study on the effect of intermediate feature selection, reporting mIoU on the MiniFrance-S dataset.

	DDPM Encoder		**DDPM Decoder**					**Ft Paras * (M)**
	1/16	**Multi-Resolution**	**1/16**	**1/8**	**1/4**	**1/2**	**Multi-Resolution**	
MLP segmentor	35.4	40.6	38.5	39.7	41.6	36.3	42.7	16.3
UPerNet [46]	-	42.4	-	-	-	-	42.5	57.3

* Ft Paras: The fine-tuned parameters, consistent with those shown in the subsequent tables.

4.3.3. Effect of Noise Scales

The input to the original diffusion model's reverse process is an image that has been given arbitrary Gaussian noise. The image context is recovered by predicting the distribution of the noise. As a result, the presence of noise helps diffusion models make better feature extractions in dense prediction tasks. Therefore, different scales of Gaussian noise are added to the input of our proposed DDPM-based segmentation model, which are used to improve noise-sensitive feature extraction performance in remote sensing imagery segmentation. We varied the noise scale by adjusting the timesteps t, allowing us to train the MLP segmentor with different combinations of samples at varying noise levels. As shown in Table 2, introducing light noise ($t \leq 100$) improves mIoU, with the best results achieved when timesteps t = 50, 100, and 150, resulting in a 4.8% mIoU improvement compared to no noise. Beyond t = 100, the benefits of added noise start to diminish, with mIoU dropping when the timestep is increased further.

Table 2. Effect of noise level on the MiniFrance-S dataset. The class abbreviations are UB—urban fabric, Ind.—industrial, commercial, public, military, PU—private and transport unit, AA—artificial non-agricultural vegetated areas, AL—arable land, PC—permanent crops, Pas.–pastures, For.—forest, HV—herbaceous vegetation associations, and WL—wetlands.

	UB	**Ind.**	**PU**	**Mine**	**AA**	**AL**	**PC**	**Pas.**	**For.**	**HV**	**WL**	**Water**	**mIoU**
t = [0]	59.4	39.3	21.2	13.5	17.9	34.1	43.3	46.4	40.4	25.0	56.5	57.3	37.9
t = [50]	62.7	42.1	21.3	13.9	19.2	37.1	44.4	47.7	42.2	28.0	60.2	61.5	40.0
t = [100]	62.4	43.0	21.8	14.2	18.8	36.3	45.1	48.4	42.9	31.6	61.0	60.5	40.5
t = [100, 50]	63.1	43.4	24.1	15.7	19.7	36.3	44.6	48.8	42.8	29.2	60.9	64.4	41.1
t = [200, 100, 50]	62.6	43.0	22.1	13.5	19.1	36.0	44.4	47.4	40.9	31.5	60.8	62.5	40.3
t = [150, 100, 50]	58.1	44.2	57.0	10.7	18.0	42.3	35.8	46.6	36.6	45.1	61.0	56.6	42.7

5. Discussion

To evaluate the effectiveness of diffusion model pre-training in remote sensing imagery segmentation, we compared our approach against three prevailing pre-training paradigms using varying sample sizes: random initialization, ImageNet-based pre-training, and three unsupervised MAE-based pre-training methods. MAE-based methods include the standard Masked Autoencoders, serving as the foundational masked pre-training framework, while ScaleMAE introduces multi-scale awareness to better capture geospatial structures. MA3E further extends this approach with a masked angle-aware encoder. This sequential progression not only reflects recent advances in the field but also provides a comprehensive benchmark for assessing the advantages of our proposed diffusion-based pre-training strategy.

Table 3 shows the segmentation performance and parameter counts across different label ratios in the MiniFrance dataset. Our proposed model, coupled with a simple MLP segmentor, achieved the highest performance when activating all parameters and training on 100% labeled data, with an mIoU score of 48.6%.

Table 3. Comparison of mIoU in sub-datasets with different labeling ratios on the MiniFrance dataset.

		MiniFrance-S	MiniFrance-M	MiniFrance-L	Ft Params (M)
Random-init (ConvNeXt-B [71])		12.9	26.2	30.5	122.0
IN-init (ConvNeXt-B)	Frozen	16.6	28.3	32.6	21.6
	Fine-tuning	34.2	37.7	44.5	122.0
MAE [19]	Frozen	34.1	38.8	42.7	23.2
	Fine-tuning	42.0	41.4	43.5	96.4
ScaleMAE [20]	Frozen	32.9	40.1	43.9	23.2
	Fine-tuning	42.6	44.2	45.0	96.1
MA3E [21]	Frozen	36.3	40.5	43.3	23.2
	Fine-tuning	42.2	43.5	46.8	96.4
ours	Frozen	42.7	46.3	48.1	16.3
	Fine-tuning	41.4	47.1	48.6	107.5

When we froze parameters during fine-tuning, the mIoU decreased by only 0.5% for our method, while MAE decreased by 1.8% and IN-init decreased by 11.9% on the MiniFrance-L dataset. On the MiniFrance-S dataset, the less labeled data scenario, freezing pre-training parameters led to a 1.3% increase in mIoU for our method, whereas the MAE- and ImageNet-based methods saw mIoU reductions of 7.9% and 17.6%, respectively.

When focusing on the impact of decreased label availability in the downstream task while freezing pre-training parameters, our proposed method exhibits a 1.5% decrease in mIoU when moving from MiniFrance-L to MiniFrance-M, and a 5.7% decrease when moving from MiniFrance-L to MiniFrance-S. Corresponding decreases are observed in MAE by 2.9% and 7.6%, and in IN-init by 4.3% and 16.0%, respectively. This highlights the role of unsupervised learning in enhancing model performance in scenarios with limited labels for downstream tasks. Moreover, compared to MAE-based methods that rely on encoder-based pre-training, decoder-based diffusion models demonstrate higher accuracy in dense prediction tasks such as semantic segmentation.

In conclusion, pre-training with a diffusion model followed by fine-tuning with a simple MLP segmentor delivers competitive segmentation performance, using only 2%

labeled imagery. This approach nearly matches the mIoU score of randomly initialized models trained on 100% labeled data, demonstrating that diffusion model pre-training effectively equips models with strong feature extraction capabilities even before fine-tuning.

Generalizability of unsupervised representations. Tables 4 and 5 present the performance comparison on semantic segmentation datasets MiniFrance and GID. We evaluated our method against previous state-of-the-art (SotA) approaches, including the SETR, SegFormer and ConvNeXt-B for natural image segmentation, PFNet and LSKNet for remote sensing imagery segmentation, and MAE and MCS, which are pre-trained by scalable self-supervised learner.

Table 4. Comparison of IoU across state-of-the-art methods on the MiniFrance-L dataset.

Method	UB	Ind.	PU	Mine	AA	AL	PC	Pas.	For.	HV	WL	Water	mIoU	Ft-Paras
SETR [72]	58.4	36.7	11.4	11.3	16.1	34.5	45.2	45.9	38.6	18.6	51.3	43.7	34.3	73.2 M
SegFormer-B4 [73]	64.7	45.6	21.3	28.0	13.7	42.1	55.6	52.9	46.6	0.0	67.0	69.1	42.2	64.1 M
PFNet [74]	67.9	49.8	24.5	31.7	19.1	40.5	54.8	54.6	47.1	0.0	69.4	74.7	44.5	33.0 M
MCS [59]	59.9	79.7	24.9	20.1	13.6	23.3	44.2	54.2	77.9	29.2	68.0	52.0	45.6	11.8 M
ConvNeXt-B [71]	62.9	52.2	33.2	45.2	25.0	39.1	57.1	45.7	48.8	8.2	75.1	72.3	47.1	122.0 M
LSKNet-T-FPN [75]	69.9	53.2	28.8	39.9	19.3	43.6	58.6	52.9	53.2	5.4	74.6	75.5	47.9	15.0 M
Ours (MiniFrance-S)	58.1	44.2	57.0	10.7	18.0	42.3	35.8	46.6	36.6	45.1	61.0	56.6	42.7	16.3 M
Ours (MiniFrance-L)	67.3	49.7	31.7	34.9	20.5	40.8	54.8	56.4	50.4	22.6	75.5	77.2	48.6	16.3 M

Table 5. Comparison of IoU across state-of-the-art methods on the GID validation dataset. The class abbreviations are IDL—industrial land, UR—urban residential, RR—rural residential, TL—traffic land, PF—paddy field, IGL—irrigated land, DC—dry cropland, GP—garden plot, AW—arbor woodland, SL—shrub land, NG—natural grassland, and AG—artificial grassland.

Method	IDL	UR	RR	TL	PF	IGL	DC	GP	AW	SL	NG	AG	River	Lake	Pond	mIoU	Ft-Paras
SETR-M	68.7	79.3	60.7	69.6	55.7	91.5	55.1	38.8	59.5	35.3	79.5	24.5	56.6	73.6	14.1	58.0	73.2 M
SegFormer-B4	69.2	79.7	69.2	71.4	57.8	90.5	59.2	40.2	59.0	37.2	82.0	26.4	60.6	82.5	13.4	60.2	64.1 M
PFNet	69.6	79.3	65.3	71.2	55.8	89.4	62.2	46.8	62.5	30.4	82.1	22.5	61.1	82.2	22.3	60.2	33.0 M
MCS	75.2	76.5	66.7	48.7	59.8	56.2	75.3	43.9	75.6	40.1	63.3	81.8	75.2	71.2	21.5	62.1	11.8 M
MA3E	59.6	66.1	53.7	54.5	70.9	80.0	66.0	34.0	76.5	21.2	59.2	44.2	90.4	82.5	73.8	62.6	96.4 M
ConvNeXt-B	70.3	91.4	67.1	57.92	70.8	64.5	76.8	77.8	58.8	81.9	81.1	28.3	63.3	74.7	40.5	66.2	122.0 M
LSKNet-T-FPN	72.7	80.5	69.6	59.4	70.3	72.5	80.7	34.7	73.6	28.1	92.1	73.4	71.7	66.8	51.7	66.5	14.4 M
ours	71.8	92.9	69.9	59.7	70.6	63.8	81.0	40.3	81.9	21.0	92.4	78.2	74.0	59.3	49.2	67.1	16.3 M

This demonstrates that features learned from diffusion model pre-training can be seamlessly transferred to downstream tasks compared to traditional methods, especially when using limited labeled data.

Despite using fewer trainable parameters, our proposed method achieved an mIoU of 48.6% on MiniFrance-L and 67.1% on GID, outperforming all SotA approaches above. We also replicate the MCS method, which employs a customized diffusion model for dense prediction. Although MCS uses slightly fewer model parameters than our method, it requires at least 25 denoising steps to reach optimal performance, significantly increasing resource consumption per sample compared to our intermediate feature transfer approach. In these comparison experiments, all methods updated their parameters during fine-tuning, which minimizes the advantage typically offered by features learned through diffusion model pre-training.

Finally, the qualitative results shown in Figures 11 and 12 provide further support for the advantages of our method. Our DDPM-based model has a superior ability to identify long-tailed classes, such as 'pond,' 'arbor,' and 'woodland.' This suggests that unsupervised pre-training helps to mitigate the underrepresentation of rare classes that are commonly overshadowed during supervised training.

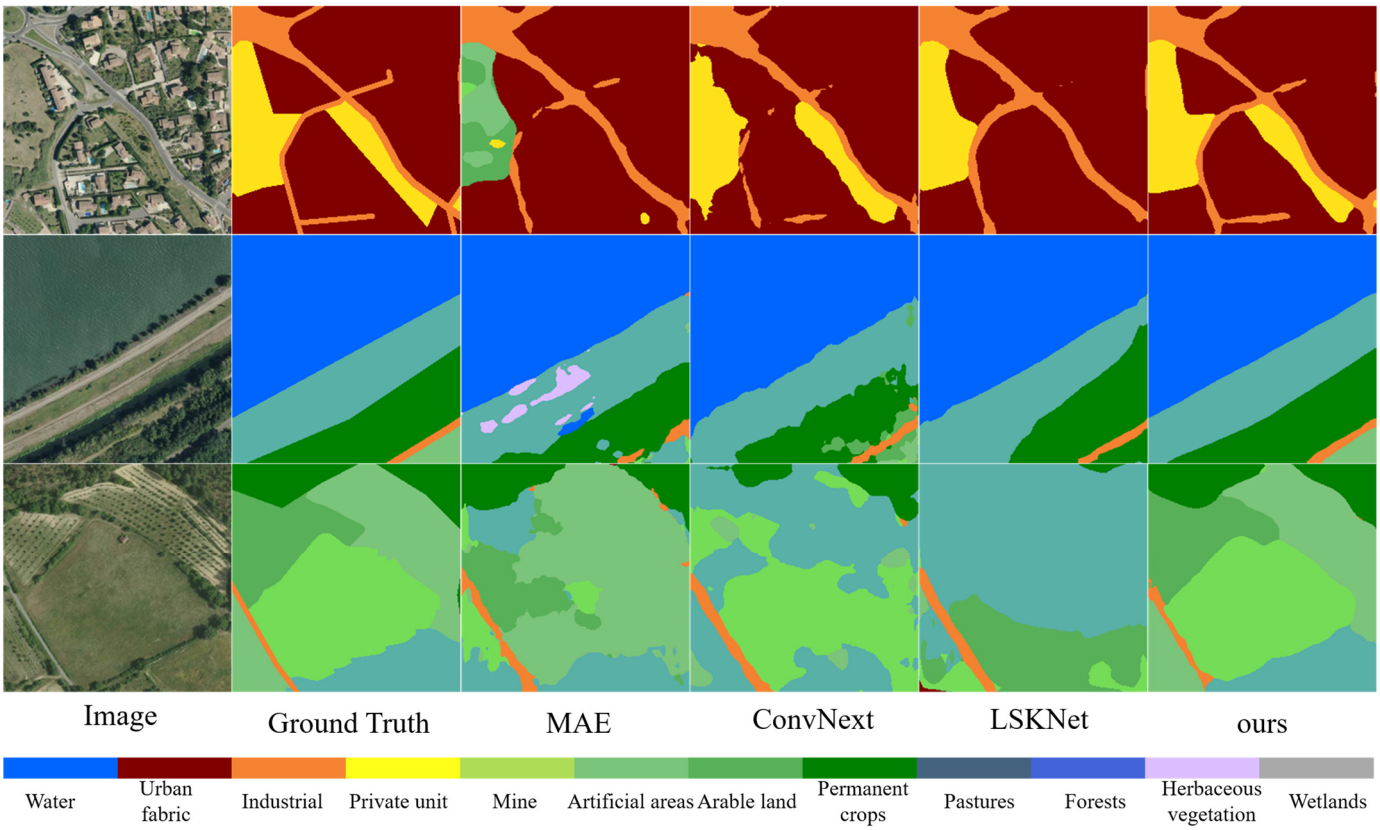

Figure 11. Visualization of results for the GID validation dataset.

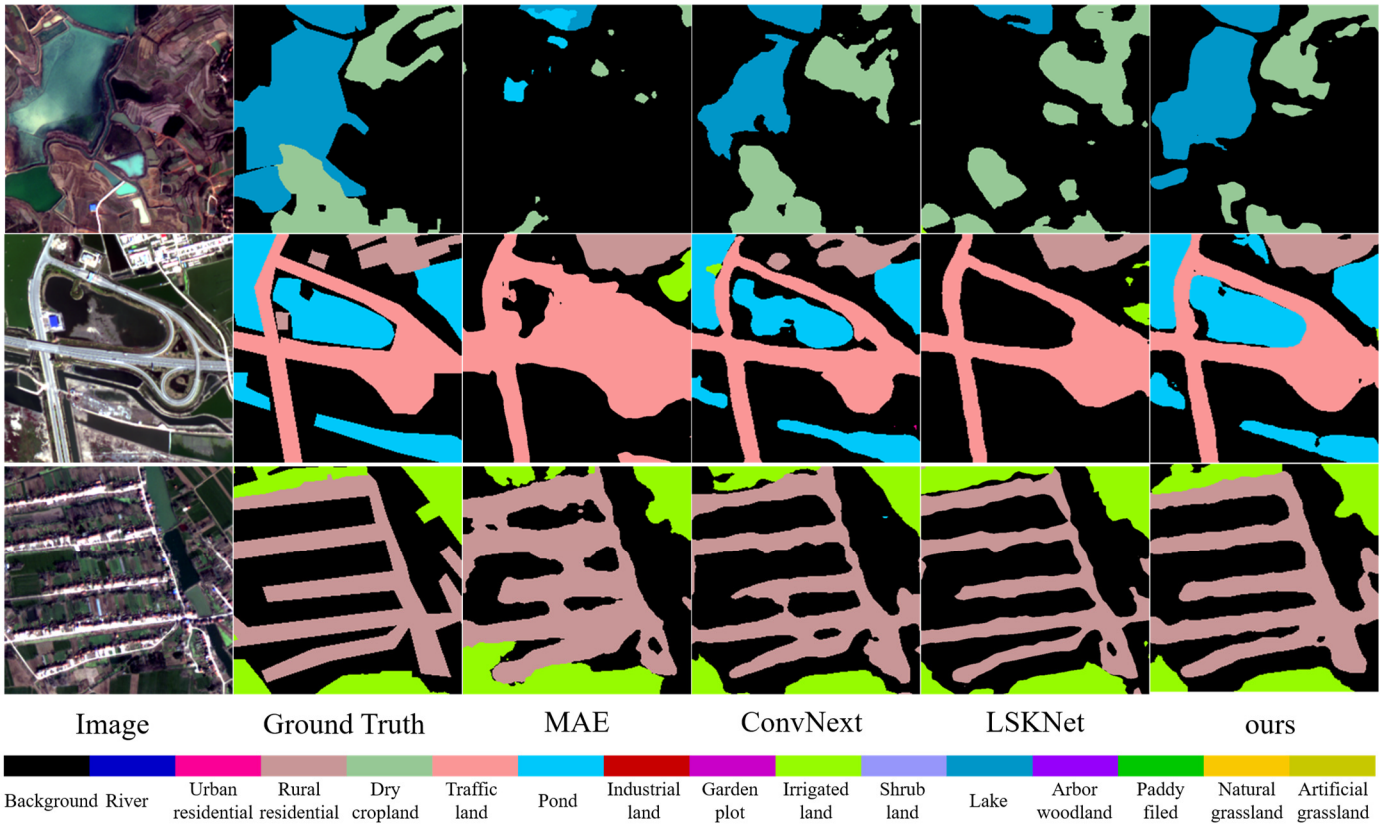

Figure 12. Visualization of results for the MiniFrance-L dataset.

6. Conclusions

We proposed a fine-tuning method based on the diffusion model for remote sensing imagery segmentation, aiming to reduce reliance on labeled data, improve data utilization efficiency, and lower application barriers. By leveraging large-scale archived remote sensing images for unsupervised pre-training, followed by domain adaptation using homogeneous data, we developed a highly generalizable DDPM for feature extraction. With multi-resolution features extracted from the DDPM decoder, we designed a lightweight MLP segmentor that requires minimal parameter fine-tuning while achieving SotA segmentation results. Our approach enables competitive performance even with only 2% labeled data, demonstrating the effectiveness of DDPM pre-training in preparing models for downstream tasks. Additionally, the learned features can be transferred to various applications with minimal adaptation, establishing our method as a robust unsupervised pre-training technique for remote sensing imagery segmentation.

It is worth noting that remote sensing imagery often contains complex and heterogeneous features: for example, varied object scales, mixed spectral responses, and noise induced by atmospheric or sensor conditions, which pose significant challenges for domain generalization and feature transfer. While our framework shows promise in mitigating some of these issues, further work is needed to enhance feature robustness and semantic alignment in cross-domain settings. In future research, we plan to explore the integration of additional language-based multimodal information to enhance the transferability of our model under varying conditions. Furthermore, we aim to further simplify the fine-tuning process, enabling more convenient adaptation to downstream tasks and reducing the complexity of practical deployment.

Author Contributions: Conceptualization, Y.L. and J.W.; methodology, S.M. and J.S.; validation, Y.L., J.L. and G.Y.; formal analysis, Y.L. and J.S.; data curation, Y.L.; writing—original draft preparation, Y.L.; writing—review and editing, X.Y. and D.W.; visualization, Y.L.; supervision, J.W. and S.M.; project administration, S.M. and J.W.; funding acquisition, J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Guangdong Provincial S & T Program, grant number 2024B1212080004. The financial support from the China Scholarship Council (CSC, Grant No. 202308440227) for the first author's living expenses is gratefully acknowledged.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: The Computer Science and Systems Laboratory (LIS) provided computational resources for the fine-tuning process. The Institute of Aerospace Remote Sensing Innovations (ARSI) provided computational resources for the pre-training process.

Conflicts of Interest: Author Jean Sequeira was employed by the company 2ik Company. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Zhang, X.; Zhou, Y.; Luo, J. Deep learning for processing and analysis of remote sensing big data: A technical review. *Big Earth Data* **2022**, *6*, 527–560. [CrossRef]
2. Sun, X.; Tian, Y.; Lu, W.; Wang, P.; Niu, R.; Yu, H.; Fu, K. From single- to multi-modal remote sensing imagery interpretation: A survey and taxonomy. *Sci. China Inf. Sci.* **2023**, *66*, 140301. [CrossRef]
3. Sobrino, J.A.; Raissouni, N. Toward remote sensing methods for land cover dynamic monitoring: Application to Morocco. *Int. J. Remote Sens.* **2000**, *21*, 353–366. [CrossRef]
4. Mo, Y.; Wu, Y.; Yang, X.; Liu, F.; Liao, Y. Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing* **2022**, *493*, 626–646. [CrossRef]

5. Alshari, E.A.; Gawali, B.W. Development of classification system for LULC using remote sensing and GIS. *Glob. Transit. Proc.* **2021**, *2*, 8–17. [CrossRef]
6. Liu, P. A survey of remote-sensing big data. *Front. Environ. Sci.* **2015**, *3*, 45. [CrossRef]
7. Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **2017**, *202*, 18–27. [CrossRef]
8. Amit, T.; Shaharbany, T.; Nachmani, E.; Wolf, L. SegDiff: Image Segmentation with Diffusion Probabilistic Models. *arXiv* **2022**, arXiv:2112.00390. [CrossRef]
9. Talukdar, S.; Singha, P.; Mahato, S.; Shahfahad; Pal, S.; Liou, Y.-A.; Rahman, A. Land-Use Land-Cover Classification by Machine Learning Classifiers for Satellite Observations—A Review. *Remote Sens.* **2020**, *12*, 1135. [CrossRef]
10. Luo, Y.; Wang, J.; Yang, X.; Yu, Z.; Tan, Z. Pixel Representation Augmented through Cross-Attention for High-Resolution Remote Sensing Imagery Segmentation. *Remote Sens.* **2022**, *14*, 5415. [CrossRef]
11. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440. [CrossRef]
12. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255. [CrossRef]
13. Sumbul, G.; de Wall, A.; Kreuziger, T.; Marcelino, F.; Costa, H.; Benevides, P.; Caetano, M.; Demir, B.; Markl, V. BigEarthNet-MM: A Large Scale Multi-Modal Multi-Label Benchmark Archive for Remote Sensing Image Classification and Retrieval. *IEEE Geosci. Remote Sens. Mag.* **2021**, *9*, 174–180. [CrossRef]
14. Long, Y.; Xia, G.-S.; Li, S.; Yang, W.; Yang, M.Y.; Zhu, X.X.; Zhang, L.; Li, D. On Creating Benchmark Dataset for Aerial Image Interpretation: Reviews, Guidances, and Million-AID. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 4205–4230. [CrossRef]
15. Manas, O.; Lacoste, A.; Giro-i-Nieto, X.; Vazquez, D.; Rodriguez, P. Seasonal Contrast: Unsupervised Pre-Training from Uncurated Remote Sensing Data. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 9394–9403. [CrossRef]
16. Cha, K.; Seo, J.; Lee, T. A Billion-scale Foundation Model for Remote Sensing Images. *arXiv* **2023**, arXiv:2304.05215. [CrossRef]
17. Ayush, K.; Uzkent, B.; Meng, C.; Tanmay, K.; Burke, M.; Lobell, D.; Ermon, S. Geography-Aware Self-Supervised Learning. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 10161–10170. [CrossRef]
18. Fuller, A.; Millard, K.; Green, J.R. SatViT: Pretraining Transformers for Earth Observation. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
19. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 16000–16009. [CrossRef]
20. Reed, C.J.; Gupta, R.; Li, S.; Brockman, S.; Funk, C.; Clipp, B.; Keutzer, K.; Candido, S.; Uyttendaele, M.; Darrell, T. Scale-MAE: A Scale-Aware Masked Autoencoder for Multiscale Geospatial Representation Learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 4088–4099.
21. Li, Z.; Hou, B.; Ma, S.; Wu, Z.; Guo, X.; Ren, B.; Jiao, L. Masked Angle-Aware Autoencoder for Remote Sensing Images. *arXiv* **2024**, arXiv:2408.01946. [CrossRef]
22. Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems, Proceedings of the Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020*; Curran Associates Inc.: Red Hook, NY, USA, 2020; pp. 6840–6851.
23. Doersch, C. Tutorial on Variational Autoencoders. *arXiv* **2021**, arXiv:1606.05908. [CrossRef]
24. An, T.; Xue, B.; Huo, C.; Xiang, S.; Pan, C. Efficient Remote Sensing Image Super-Resolution via Lightweight Diffusion Models. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 1–5. [CrossRef]
25. Ayala, C.; Sesma, R.; Aranda, C.; Galar, M. Diffusion Models for Remote Sensing Imagery Semantic Segmentation. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Pasadena, CA, USA, 16–21 July 2023; pp. 5654–5657. [CrossRef]
26. Bandara, W.G.C.; Nair, N.G.; Patel, V.M. DDPM-CD: Remote Sensing Change Detection using Denoising Diffusion Probabilistic Models. *arXiv* **2022**, arXiv:2206.11892. [CrossRef]
27. Thrun, S.; Pratt, L. Learning to Learn: Introduction and Overview. In *Learning to Learn*; Springer: Boston, MA, USA, 1998; pp. 3–17. [CrossRef]
28. Doersch, C.; Gupta, A.; Efros, A.A. Unsupervised Visual Representation Learning by Context Prediction. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1422–1430. [CrossRef]

29. Han, X.; Zhang, Z.; Ding, N.; Gu, Y.; Liu, X.; Huo, Y.; Qiu, J.; Yao, Y.; Zhang, A.; Zhang, L.; et al. Pre-trained models: Past, present and future. *AI Open* **2021**, *2*, 225–250. [CrossRef]
30. Yang, L.; Zhang, Z.; Song, Y.; Hong, S.; Xu, R.; Zhao, Y.; Zhang, W.; Cui, B.; Yang, M.-H. Diffusion Models: A Comprehensive Survey of Methods and Applications. *ACM Comput. Surv.* **2023**, *56*, 1–39. [CrossRef]
31. Reed, C.J.; Gupta, R.; Li, S.; Brockman, S.; Funk, C.; Clipp, B.; Keutzer, K.; Candido, S.; Uyttendaele, M.; Darrell, T. Self-Supervised Pretraining Improves Self-Supervised Pretraining. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 4–8 January 2022; pp. 2584–2594. [CrossRef]
32. Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; Bharath, A.A. Generative Adversarial Networks: An Overview. *IEEE Signal Process. Mag.* **2018**, *35*, 53–65. [CrossRef]
33. Alqahtani, H.; Kavakli-Thorne, M.; Kumar, G. Applications of Generative Adversarial Networks (GANs): An Updated Review. *Arch. Comput. Methods Eng.* **2021**, *28*, 525–552. [CrossRef]
34. Sainath, T.N.; Kingsbury, B.; Ramabhadran, B. Auto-encoder bottleneck features using deep belief networks. In Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012; pp. 4153–4156. [CrossRef]
35. Jiang, X.; Zhang, Y.; Zhang, W.; Xiao, X. A novel sparse auto-encoder for deep unsupervised learning. In Proceedings of the 2013 Sixth International Conference on Advanced Computational Intelligence (ICACI), Beijing, China, 19–21 October 2013; pp. 256–261. [CrossRef]
36. Vincent, P.; Larochelle, H.; Bengio, Y.; Manzagol, P.-A. Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th International Conference on Machine Learning (ICML), Helsinki, Finland, 5–9 July 2008; pp. 1096–1103. [CrossRef]
37. Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; Manzagol, P.-A.; Bottou, L. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.* **2010**, *11*, 3371–3408.
38. Vahdat, A.; Kautz, J. NVAE: A Deep Hierarchical Variational Autoencoder. In *Advances in Neural Information Processing Systems, Proceedings of the Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020*; Curran Associates Inc.: Red Hook, NY, USA, 2020; pp. 19667–19679.
39. Gupta, A.; Wu, J.; Deng, J.; Li, F.-F. Siamese Masked Autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS), Proceedings of the Conference on Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023*; Curran Associates Inc.: Red Hook, NY, USA, 2023; pp. 40676–40693.
40. Singh, M.; Duval, Q.; Alwala, K.V.; Fan, H.; Aggarwal, V.; Adcock, A.; Joulin, A.; Dollár, P.; Feichtenhofer, C.; Girshick, R.; et al. The effectiveness of MAE pre-pretraining for billion-scale pretraining. *arXiv* **2023**, arXiv:2303.13496. [CrossRef]
41. Song, J.; Meng, C.; Ermon, S. Denoising Diffusion Implicit Models. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, 3–7 May 2021; pp. 1–20.
42. Baranchuk, D.; Voynov, A.; Rubachev, I.; Khrulkov, V.; Babenko, A. Label-Efficient Semantic Segmentation with Diffusion Models. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, 3–7 May 2021; pp. 1–15.
43. Lei, J.; Wang, Q.; Cheng, P.; Ba, Z.; Qin, Z.; Wang, Z.; Liu, Z.; Ren, K. Masked Diffusion Models Are Fast Distribution Learners. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, 1 May 2023; pp. 1–23.
44. Huang, S.; Lu, Z.; Cheng, R.; He, C. FaPN: Feature-aligned Pyramid Network for Dense Image Prediction. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 844–853. [CrossRef]
45. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6230–6239. [CrossRef]
46. Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; Sun, J. Unified Perceptual Parsing for Scene Understanding. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 418–434. [CrossRef]
47. Lin, T.Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [CrossRef]
48. Chen, J.; Qin, D.; Hou, D.; Zhang, J.; Deng, M.; Sun, G. Multiscale Object Contrastive Learning–Derived Few-Shot Object Detection in VHR Imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
49. Wang, D.; Zhang, J.; Du, B.; Xia, G.-S.; Tao, D. An Empirical Study of Remote Sensing Pretraining. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–20. [CrossRef]
50. Gidaris, S.; Singh, P.; Komodakis, N. Unsupervised Representation Learning by Predicting Image Rotations. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018; pp. 1–16.
51. Bao, H.; Dong, L.; Piao, S.; Wei, F. BEiT: BERT Pre-Training of Image Transformers. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, 3–7 May 2021; pp. 1–18.

52. Wei, C.; Mangalam, K.; Huang, P.-Y.; Li, Y.; Fan, H.; Xu, H.; Wang, H.; Xie, C.; Yuille, A.; Feichtenhofer, C. Diffusion Models as Masked Autoencoders. *arXiv* **2023**, arXiv:2304.03283. [CrossRef]
53. Pan, Z.; Chen, J.; Shi, Y. Masked Diffusion as Self-supervised Representation Learner. *arXiv* **2023**, arXiv:2308.05695. [CrossRef]
54. Zhao, Z.; Bai, H.; Zhu, Y.; Zhang, J.; Xu, S.; Zhang, Y.; Zhang, K.; Meng, D.; Timofte, R.; Van Gool, L. DDFM: Denoising Diffusion Model for Multi-Modality Image Fusion. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 8082–8093. [CrossRef]
55. Xu, Y.; Yu, W.; Ghamisi, P.; Kopp, M.; Hochreiter, S. Txt2Img-MHN: Remote Sensing Image Generation from Text Using Modern Hopfield Networks. *IEEE Trans. Image Process.* **2023**, *32*, 5737–5750. [CrossRef] [PubMed]
56. Czerkawski, M.; Tachtatzis, C. Exploring the Capability of Text-to-Image Diffusion Models with Structural Edge Guidance for Multispectral Satellite Image Inpainting. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 1–5. [CrossRef]
57. Xiao, Y.; Yuan, Q.; Jiang, K.; He, J.; Jin, X.; Zhang, L. EDiffSR: An Efficient Diffusion Probabilistic Model for Remote Sensing Image Super-Resolution. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–14. [CrossRef]
58. Jia, J.; Lee, G.; Wang, Z.; Lyu, Z.; He, Y. Siamese Meets Diffusion Network: SMDNet for Enhanced Change Detection in High-Resolution RS Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 8189–8202. [CrossRef]
59. Kolbeinsson, B.; Mikolajczyk, K. Multi-Class Segmentation from Aerial Views Using Recursive Noise Diffusion. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2024; pp. 8439–8449.
60. Chen, N.; Yue, J.; Fang, L.; Xia, S. SpectralDiff: A Generative Framework for Hyperspectral Image Classification with Diffusion Models. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–16. [CrossRef]
61. Ma, J.; Xie, W.; Li, Y.; Fang, L. BSDM: Background Suppression Diffusion Model for Hyperspectral Anomaly Detection. *arXiv* **2023**, arXiv:2307.09861. [CrossRef]
62. Li, T.; Katabi, D.; He, K. Return of Unconditional Generation: A Self-supervised Representation Generation Method. *arXiv* **2023**, arXiv:2312.03701. [CrossRef]
63. Misra, I.; van der Maaten, L. Self-Supervised Learning of Pretext-Invariant Representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 6707–6717. [CrossRef]
64. Zhang, D.; Li, C.; Li, H.; Huang, W.; Huang, L.; Zhang, J. Rethinking Alignment and Uniformity in Unsupervised Image Semantic Segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 11709–11717. [CrossRef]
65. Kornblith, S.; Shlens, J.; Le, Q.V. Do Better ImageNet Models Transfer Better? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 2661–2671. [CrossRef]
66. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 580–587. [CrossRef]
67. Mukhopadhyay, S.; Gwilliam, M.; Agarwal, V.; Padmanabhan, N.; Swaminathan, A.; Hegde, S.; Zhou, T.; Shrivastava, A. Diffusion Models Beat GANs on Image Classification. *arXiv* **2023**, arXiv:2307.08702. [CrossRef]
68. Melas-Kyriazi, L.; Rupprecht, C.; Laina, I.; Vedaldi, A. Finding an Unsupervised Image Segmenter in each of your Deep Generative Models. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, 3–7 May 2021; pp. 1–18.
69. Castillo-Navarro, J.; Le Saux, B.; Boulch, A.; Audebert, N.; Lefèvre, S. Semi-supervised semantic segmentation in Earth Observation: The MiniFrance suite, dataset analysis and multi-task network study. *Mach. Learn.* **2022**, *111*, 3125–3160. [CrossRef]
70. Tong, X.-Y.; Xia, G.-S.; Lu, Q.; Shen, H.; Li, S.; You, S.; Zhang, L. Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sens. Environ.* **2020**, *237*, 111322. [CrossRef]
71. Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 11976–11986. [CrossRef]
72. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.; et al. Rethinking Semantic Segmentation From a Sequence-to-Sequence Perspective With Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual Event, 19–25 June 2021; pp. 6881–6890. [CrossRef]
73. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Virtual Event, 6–14 December 2021; pp. 12077–12090.

74. Li, X.; He, H.; Li, X.; Li, D.; Cheng, G.; Shi, J.; Weng, L.; Tong, Y.; Lin, Z. PointFlow: Flowing Semantics Through Points for Aerial Image Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual Event, 19–25 June 2021; pp. 4217–4226. [CrossRef]
75. Li, Y.; Li, X.; Dai, Y.; Hou, Q.; Liu, L.; Liu, Y.; Cheng, M.-M.; Yang, J. LSKNet: A Foundation Lightweight Backbone for Remote Sensing. *Int. J. Comput. Vis.* **2025**, *133*, 1410–1431. [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.