remote sensing

MDPI

Article

Remote Sensing Image Semantic Segmentation Sample Generation Using a Decoupled Latent Diffusion Framework

Yue Xu [1], Honghao Liu [1,2,3], Ruixia Yang [2] and Zhengchao Chen [1,*]

[1] State Key Laboratory of Remote Sensing and Digital Earth, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100101, China; xuyue@aircas.ac.cn (Y.X.); liuhonghao22@mails.ucas.ac.cn (H.L.)
[2] International Research Center of Big Data for Sustainable Development Goals, Beijing 100094, China; yangrx@aircas.ac.cn
[3] University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: chenzc@aircas.ac.cn

Abstract

This paper addresses the challenges of sample scarcity and class imbalance in remote sensing image semantic segmentation by proposing a decoupled synthetic sample generation framework based on a latent diffusion model. The method consists of two stages. In the label generation stage, we fine-tune a pretrained latent diffusion model with LoRA to generate semantic label masks from textual descriptions. A novel proportion-aware loss function explicitly penalizes deviations from the desired class distribution in the generated mask. In the image generation stage, we use ControlNet to train a multi-condition image generation network that takes the synthesized mask, along with its text description, as input and produces a realistic remote sensing image. The base Stable Diffusion model's weights remain frozen during this process, with the trainable ControlNet ensuring that outputs are structurally and semantically aligned with the input labels. This two-stage approach yields coherent image–mask pairs that are well-suited for training segmentation models. Experiments show that models trained on the synthetic samples produced by the proposed method achieve high visual quality and semantic consistency. The proportion-aware loss effectively mitigates the impact of minority classes, boosting segmentation performance on under-represented categories. Results also reveal that adding a suitable proportion of synthetic sample improves segmentation accuracy, whereas an excessive share can cause over-fitting or misclassification. Comparative tests across multiple models confirm the generality and robustness of the approach.

Keywords: remote sensing; semantic segmentation; sample synthesis; diffusion models; class imbalance; latent diffusion; ControlNet; LoRA

check for updates

Academic Editor: Pedro Melo-Pinto

Received: 15 May 2025
Revised: 19 June 2025
Accepted: 20 June 2025
Published: 22 June 2025

Citation: Xu, Y.; Liu, H.; Yang, R.; Chen, Z. Remote Sensing Image Semantic Segmentation Sample Generation Using a Decoupled Latent Diffusion Framework. *Remote Sens.* **2025**, *17*, 2143. https://doi.org/10.3390/rs17132143

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

Semantic segmentation is a key technique in interpreting remote sensing images. By learning a dense mapping between image space and label space, it assigns every pixel in a remote sensing image to a specific semantic class (e.g., building, road, water, vegetation), thereby retrieving the categories and spatial distribution of objects. This information underpins applications such as urban management, environmental monitoring, disaster prevention and mitigation, and resource management [1–5]. In recent years, deep learning algorithms have made notable strides in remote sensing image semantic segmentation. For instance, the method proposed by Diakogiannis et al. [6] achieves high-precision results

on the ISPRS Potsdam and Vaihingen datasets [7]. Nevertheless, mainstream approaches remain supervised learning and depend on large quantities of high-quality annotated samples. Because remote sensing imagery spans vast areas and highly variable land surface environments, annotation demands substantial expertise and is both time-consuming and labor-intensive, leaving the field with an acute shortage of first-rate training samples and limiting broader adoption of segmentation algorithms [8]. Moreover, object categories in land cover data are inherently imbalanced, with some classes covering much larger areas than others, causing sharply differing sample proportions. Models trained under such imbalance often underperform in small sample categories [9]. In sum, the dual challenges of scarce high-quality samples and severe class imbalance have become major bottlenecks for remote sensing image interpretation and its practical applications.

The rapid progress of generative models [10] has opened up new avenues for addressing the scarcity of training samples in semantic segmentation. Generative models learn either the joint distribution $P(X, Y)$ or the marginal distribution $P(X)$, thereby revealing the intrinsic structure and regularities of data and yielding new synthetic samples that resemble the training set. Representative families include Variational Autoencoders (VAEs) [11], Generative Adversarial Networks (GANs) [12–14], autoregressive models [15], and diffusion models [16–18]. Among them, diffusion models simulate a forward "diffusion" (noise-adding) process and its reverse denoising process to learn data distributions and synthesize new samples. Because they capture fine-grained semantic details and excel in image quality, training stability, and controllability, diffusion models are steadily supplanting traditional GANs in image generation tasks. Diffusion approaches fall into two broad categories: pixel-level diffusion models and latent diffusion models. Pixel-level diffusion models operate directly in the high-dimensional pixel space, adding noise step-by-step and then learning the reverse process; DALL·E 2 [19] is a prominent example. However, their computational cost grows exponentially with image resolution, posing severe resource constraints for high-resolution remote sensing imagery.

To alleviate this problem, researchers introduced latent diffusion models (LDMs), exemplified by Stable Diffusion (SD) [20]. Instead of working in pixel space, LDMs apply the diffusion process within a lower-dimensional latent space obtained via a Variational Autoencoder (VAE) [11]. After diffusion in this latent space, a decoder reconstructs the final image, striking a favorable balance between data quality and computational efficiency. Consequently, LDMs have become a focal point of current generative model research. Several optimizations further enhance LDM training and inference. Low-Rank Adaptation (LoRA) reduces the number of trainable parameters through low-rank decomposition, markedly cutting computational complexity [21]. ControlNet injects external conditional control networks without altering the original large-scale pretrained backbone, enabling precise structural and semantic alignment between generated images and input masks, thus making the generation process more predictable [22].

Researchers have begun using generative models to synthesize remote sensing samples, and the approaches can be roughly divided into two categories: end-to-end image-and-label generation, and the decoupled label-then-image generation. End-to-end methods employ a single unified model to produce paired images and labels in one pass. Their main advantage is the ability to leverage end-to-end training to expand datasets at scale and reduce reliance on manual annotation [23]. However, the intricate relationship between remote sensing imagery and its semantic labels makes it difficult for such models to balance pixel-level detail with label consistency, while their computational demands remain high. Decoupled methods, which have attracted growing attention in recent years, first generate semantic masks independently and then synthesize corresponding images conditioned on those masks, thereby disentangling the two stages [24]. This strategy improves con-

trollability and flexibility, allows for targeted optimization at each step, and lowers the overall computational burden. For example, Tang et al. proposed CRS-Diff, a controllable image synthesis framework that integrates text, image, and metadata conditions via ControlNet-based feature fusion, demonstrating strong versatility in remote sensing generation tasks [25]. DiffusionSat extends this line of work by incorporating geospatial metadata such as location and time into a foundational diffusion model, enabling temporally coherent and high-quality satellite imagery generation [26]. Zhu et al. introduced RSDiffSR, a conditional diffusion model that enhances super-resolved remote sensing outputs through content and edge guidance mechanisms [27].

While these approaches have significantly advanced multi-condition controllability and structural guidance, they predominantly adopt tightly coupled architectures, which limit interpretability, flexibility, and adaptation to task-specific requirements. Moreover, most of these models focus on general image synthesis or low-level enhancement tasks like super-resolution, rather than high-level semantic generation. In particular, existing remote sensing sample generation pipelines often follow a text-to-image paradigm, producing synthetic image–text pairs without pixel-level semantic labels. This makes them insufficient for semantic segmentation tasks, which demand densely annotated, pixel-wise training samples. Therefore, a task-specific and structurally decoupled generation framework is urgently needed to meet the growing demands of semantic-level understanding in remote sensing.

To address the scarcity of pixel-level annotated samples for remote sensing semantic segmentation and the pronounced class imbalance within them, this study builds on two complementary principles. First, we leverage the strengths of latent diffusion models, which strike an effective balance between generation quality and computational efficiency while supporting flexible conditional control. Second, we introduce a structurally decoupled framework that separates the label generation and image synthesis stages. This decoupling not only improves semantic consistency between labels and images but also allows for targeted control over class distributions—thereby enhancing the diversity, balance, and usability of the generated training samples for segmentation tasks.

Building on these, we introduce a decoupled latent diffusion framework capable of producing large batches of synthetic, fully annotated remote sensing segmentation samples. The method consists of two main stages. The first stage is for label generation. We fine-tune a pretrained latent diffusion model with LoRA [21] to generate semantic label masks from textual descriptions extracted with an open-source multimodal large language model (MLLM) [28]. We then design a novel proportion-aware loss function that explicitly penalizes deviations from the desired class distribution in the generated mask. In the second stage (image generation), we use ControlNet [22] to train a multi-condition image generation network that takes the synthesized mask, along with its text description, as input and produces a realistic remote sensing image. The base Stable Diffusion model's weights remain frozen during this process, with the trainable ControlNet ensuring that outputs are structurally and semantically aligned with the input labels.

Unlike those prior pipelines—which either entangle label and image synthesis or lack pixel-wise outputs—our proposed method explicitly decouples the two stages to produce segmentation-ready image–mask pairs that are well-suited for training segmentation models. We tested the proposed sample generation method on the ISPRS Potsdam dataset and compared it with two classic image generation baselines, Pix2Pix [29] and CycleGAN [30]. The results show that our approach outperforms the baselines, producing synthetic samples with superior visual quality and semantic consistency. To verify downstream utility, we trained DeeplabV3+ [31], PSPNet [32], and SegFormer [33] segmentation models with the synthesized data. Across all three networks, overall segmentation accuracy and class bal-

ance metrics improved markedly; gains were especially pronounced for the rare "Clutter" and "Car" categories, underscoring the proposed method's generality and robustness. We further analyzed how the proportion of synthetic samples affects performance. As the ratio of synthetic to real samples increased, mIoU and mF1 first rose and then declined; the best results were obtained when the proportion of synthesized samples approached 40%. This indicates that a moderate amount of synthetic sample can significantly boost segmentation performance, whereas excessive synthetic data risks over-fitting or misclassification.

The main contributions of this paper are as follows: (1) We introduce a decoupled framework that automatically produces paired images and masks with high fidelity, breaking the limitations of conventional text–image sample generation methods and alleviating the sample data scarcity that plagues remote sensing semantic segmentation tasks. (2) We design a color ratio loss tailored to multi-class semantic masks, which enforces precise class proportion constraints during sample generation and offers a practical solution to the long-standing problem of under-represented classes. In the following, we detail the proposed framework (Section 2), experimental results and discussion (Section 3), and conclusions (Section 4).

2. Methodology

2.1. Overview of Latent Diffusion Model

2.1.1. Stable Diffusion Model

The Stable Diffusion (SD) model [20], as a prototypical latent diffusion model, has the basic structure illustrated in Figure 1, which comprises three core components: pixel space, latent space, and condition embeddings.

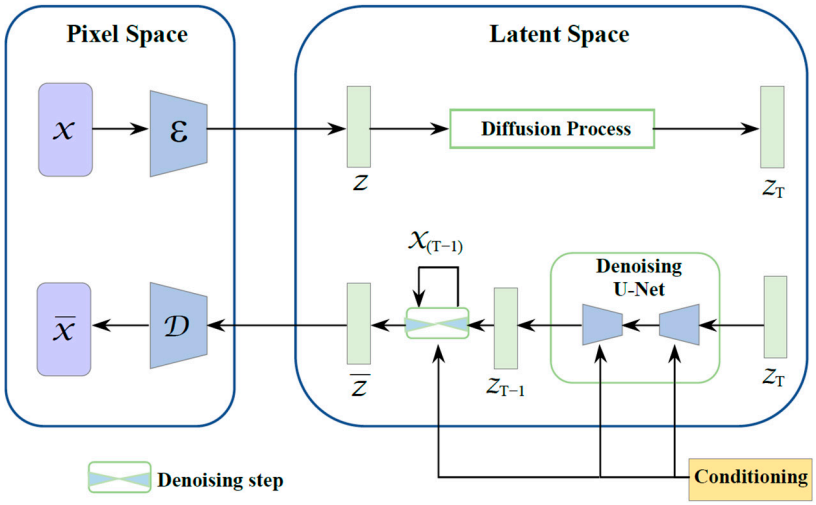


Figure 1. Schematic diagram of latent diffusion model [20].

For an input image X, the SD model operates as follows:

1. Encoding to latent space: A pretrained Variational Autoencoder (VAE) compresses the input image X into latent representations, reducing data dimensionality while extracting essential feature information. The result is a latent code.
2. Forward diffusion: In the latent space, noise is gradually added to Z step by step, producing a sequence of progressively noisier latent codes. Z_T This simulates the degradation of the image from ordered to disordered, laying the groundwork for subsequent denoising.
3. Denoising with U-Net: The noisy latent code Z_T is fed into a U-Net-based denoising network. Here, condition embeddings are used to steer the denoising toward outputs

that satisfy specified conditions. The conditions are derived from textual prompts and other priors, and are injected into the U-Net, guiding it to recover a cleaner latent code $\overline{Z}$ that adheres to the conditioning.

4. Decoding with VAE: The denoised latent code $\overline{Z}$ is decoded back to pixel space by the VAE decoder, yielding the reconstructed image $\overline{X}$.

We employ the VAE to handle both compression and reconstruction, leveraging its powerful encoder–decoder capabilities for bidirectional conversion between pixel and latent spaces. We also use the CLIP text encoder [34] to transform text prompts into condition embeddings, infusing the denoising process with rich semantic guidance and thereby improving the fidelity and relevance of the reconstructed images.

2.1.2. LoRA Method

As mentioned above, LoRA is an optimization method for fine-turning large latent diffusion models. The LoRA structure [21] is shown in Figure 2. LoRA reconstructs the model's parameter update process via low-rank matrix decomposition. For any weight matrix, W in the pretrained model is expressed in Equation (1):

$$W \in \mathbb{R}^{d \times k} \tag{1}$$

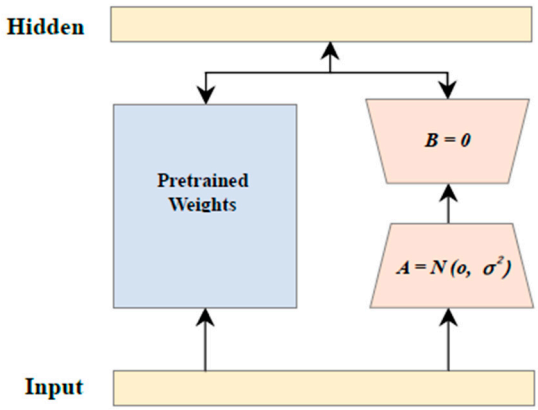


Figure 2. LoRA structure [21].

LoRA decomposes its parameter update ΔW into the product of two low-rank matrices, i.e.,

$$\Delta W = B \cdot A, \quad B \in \mathbb{R}^{d \times r}, \quad A \in \mathbb{R}^{r \times k} \tag{2}$$

where $r \ll min(d, k)$ is a rank hyperparameter. By introducing trainable parameter matrices, A and B, this decomposition reduces the update space from $d \times k$ to $r \times (d + k)$, significantly lowering computational complexity.

In practice, LoRA greatly reduces the number of model parameters that must be updated, making find-tuning much more lightweight. It only requires training a very small number of additional parameters and leaves the original pretrained weights unchanged, preserving the base model's generative capacity. Furthermore, since adjustments are applied via external low-rank matrices rather than by modifying the pretrained parameters directly, LoRA modules can be dynamically plugged in, removed, loaded, or unloaded for different tasks, enabling rapid switching and deployment in multi-task scenarios.

2.1.3. ControlNet Method

ControlNet is another optimization approach that adds trainable condition-specific layers to a diffusion mode [22]. Its structure is illustrated in Figure 3. It comprises the primary diffusion model and an auxiliary conditional control network. The left part of

Figure 3 is the primary diffusion model. It handles the core image generation task with a large, pretrained network such as Stable Diffusion. These base model parameters remain frozen during ControlNet training to preserve the original generative capabilities. The right part of Figure 3 shows the conditional control network, which is a small, trainable sub-network injected into the U-Net backbone of the base model. The ControlNet's weights are initialized to zero so as not to perturb the frozen model at start and are gradually updated during fine-tuning to respond to various control signals (e.g., edge maps, segmentation masks), thereby improving the structural consistency of the outputs. A key advantage of ControlNet is its ability to accept multiple types of input conditions and to dynamically plug in or remove specific control modules, facilitating rapid, efficient switching across different task scenarios.

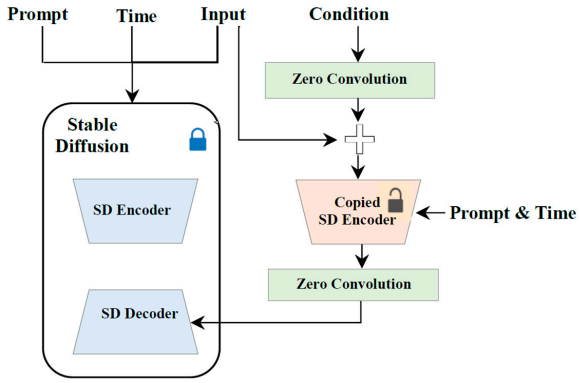


Figure 3. ControlNet structure [22].

2.2. Decoupled Latent Diffusion Framework for Synthetic Sample Generation

This study proposes a decoupled latent diffusion framework for synthesis sample generation, as illustrated in Figure 4. The method is divided into two stages: label generation and image generation. In the label generation stage, Stable Diffusion is combined with LoRA to produce high-quality semantic labels under limited computational resources. A novel class ratio control loss is introduced to balance category distribution. In the image generation stage, Stable Diffusion is paired with ControlNet to train a multi-condition image generation network. By relying on an external, trainable conditioning network rather than modifying the original pretrained Stable Diffusion weights, the generated images remain highly consistent with the input labels in both structure and semantics.

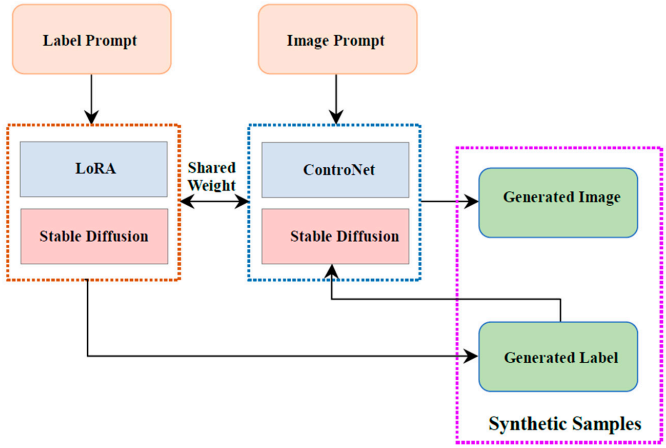


Figure 4. Schematic diagram of decoupled latent diffusion framework for sample synthesis.

2.2.1. Text-Driven Label Generation with LoRA

We employ Low-Rank Adaptation (LoRA) to perform targeted fine-tuning of a pre-trained generative model, improving both efficiency and scalability for text-driven label generation. A lightweight adapter network is rapidly trained to produce segmentation masks. Building on this, we introduce a proportion-aware loss function that incorporates both color ratio control and textual label information into the training objective. The loss constrains the category proportions in the generated masks, allowing for us to regulate the distribution of different classes in the output mask and dynamically adjust class ratios during training. In this way, minority classes are reinforced in the synthetic labels, improving segmentation accuracy for rare categories. A dedicated proportion control module is detailed in Section 2.2.2.

The structure of the LoRA-based text-driven label generation algorithm is illustrated in Figure 5. A textual prompt serves as the input to both the pretrained Stable Diffusion model and the lightweight LoRA module, enabling flexible control of the label generation process without modifying the base model's parameters (indicated by the lock icon). The Stable Diffusion model generates semantic label masks conditioned on the prompt, while the LoRA module enhances adaptability and efficiency during fine-tuning. To guide optimization, two loss functions are applied: Cross-entropy (CE) loss compares the generated label with the ground truth label derived from the real image, ensuring pixel-level semantic accuracy, while the proportion-aware loss enforces the alignment between the category distribution in the generated labels and a predefined class proportion. This dual-loss design ensures both semantic consistency and class balance in the generated masks.

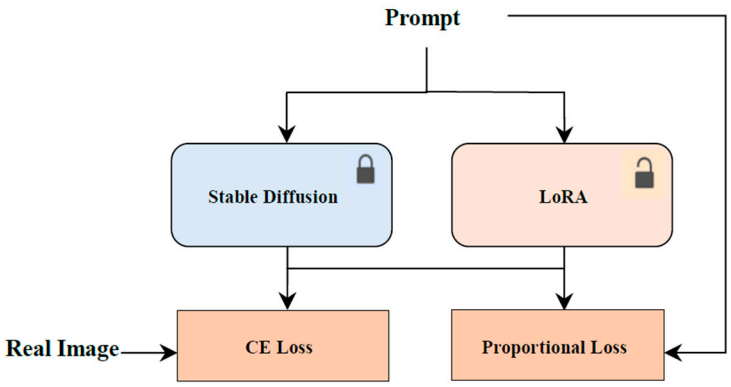


Figure 5. Structure of LoRA-based text-driven label generation algorithm.

2.2.2. Proportion-Aware Loss Function

Most existing text-driven label generation methods rely solely on label classes to create segmentation masks and do not regulate the proportion of each class in the output. As a result, minority classes can be severely under-represented, degrading segmentation accuracy for those classes. To remedy this, we propose a proportion-aware loss that combines color-based proportion control with textual label information to optimize label generation. Specifically, the input text prompt is augmented with explicit class proportion constraints that define expected share of each class in the segmentation mask. During training, an additional loss term is calculated based on the deviation between the generated mask's class proportions and the target proportions, ensuring that the final synthetic mask obeys the specified class distribution.

We further refine this proportion control loss function by introducing a proportion-aware loss that constrains the class proportions in the generated output, ensuring they

align with the intended textual description and target class distribution. The calculation formula is expressed in Equation (3):

$$\mathcal{L}_{prop} = \frac{1}{n}\Sigma_{i=1}^{n}\left|p_{gen,i} - p_{prompt.i}\right| \quad (3)$$

where $p_{gen,i}$ denotes the actual proportion of class i in the generated mask, while $p_{prompt.i}$ represents the target proportion specified in the input prompt. By minimizing this loss, the model adjusts the class distribution in the generated labels so that it more closely matches the desired proportions.

Combining this term with the standard cross-entropy loss ($\mathcal{L}_{CE}$), the overall training objective for label generation used in this study is defined by Equation (4):

$$\mathcal{L} = 0.5\mathcal{L}_{CE} + 0.5\mathcal{L}_{prop} \quad (4)$$

where λ is a weighting hyperparameter that controls the influence of the proportion-aware loss. In our experiments, $\mathcal{L}_{prop}$ was given a moderate weight (tuned empirically) to balance class ratio enforcement with the primary segmentation loss.

2.2.3. Image Generation Method Based on ControlNet

We design an image generation algorithm whose structure (see Figure 6) uses both text prompts and segmentation masks as control signals. The generative model is trained to reconstruct realistic remote sensing images directly from the masks, harnessing the structural information encoded in the labels while also leveraging the semantic guidance provided by the text. This dual conditioning significantly enhances the realism and detail fidelity of the generated imagery.

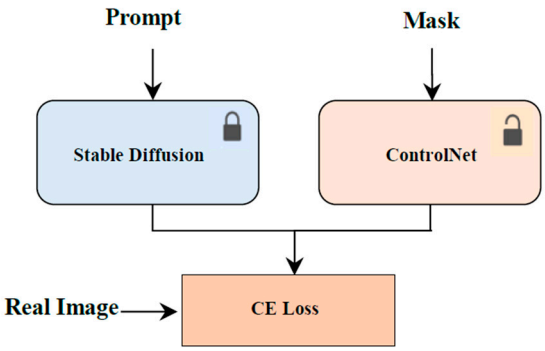


Figure 6. Structure of image generation algorithm.

The synthetic sample generation and segmentation pipeline algorithm is listed below (Algorithm 1), where

- Lines 5–13 implement LoRA fine-tuning with the proportion-aware loss (Lprop) that enforces class ratio targets Ptarget;
- Lines 15–22 train a ControlNet conditioned on both the generated mask and its caption, while keeping the Stable Diffusion backbone frozen;
- Lines 23–29 generate a user-defined number of synthetic image–mask pairs;
- Line 30 mixes them with real data at a configurable ratio α (e.g., 40%);
- Lines 31–34 perform standard supervised training of any segmentation model (e.g., DeepLabV3+, PSPNet, SegFormer) on the augmented dataset.

During training, we freeze all original Stable Diffusion weights and only train the ControlNet parameters, providing the network with pairs of ground truth masks and images (along with their text descriptions) as supervision. In this way, the ControlNet learns to map

label structures to image content without altering the base model's weights, preserving the pretrained model's knowledge and ensuring that output quality remains high. The result of this stage is a coherent image–mask pair: a realistic synthetic remote sensing image that is structurally and semantically aligned with its corresponding segmentation label.

Algorithm 1: Synthetic–Sample Generation and Segmentation Pipeline

Require: $\mathcal{D}_{\text{real}} = \{(I_k, M_k)\}$ ▷ real images and masks
Require: $\mathcal{P}_{\text{target}} = \{p_c\}$ ▷ target class ratios
Require: TextGen$(\cdot)$, SD_{base} (frozen)
Hyper-parameters: $\text{rank}_{\text{LoRA}}$, λ_{prop}

/* Stage 1 – Label generation */
for all $(I_k, M_k) \in \mathcal{D}_{\text{real}}$ do
    $\text{caption}_k \leftarrow \text{TextGen}(I_k, M_k)$
    buffer $\leftarrow (\text{caption}_k, M_k, \mathcal{P}_{\text{target}})$
end for
Initialise LoRA adapters on SD_{base} ($\text{rank}_{\text{LoRA}}$)
while training do
    Sample mini-batch $(\text{cap}, M_{\text{gt}}, \mathcal{P}_{\text{target}})$
    $M_{\text{pred}} \leftarrow \text{SD}_{\text{mask}}(\text{cap})$
    $L_{\text{CE}} \leftarrow \text{CrossEntropy}\left(M_{\text{pred}}, M_{\text{gt}}\right)$
    $L_{\text{prop}} \leftarrow \frac{1}{|C|}\sum_c |\text{prop}(M_{\text{pred}}, c) - p_c|$
    $L_{\text{tot}} \leftarrow L_{\text{CE}} + \lambda_{\text{prop}} L_{\text{prop}}$
    Update LoRA with ∇L_{tot}
end while
$\text{SD}_{\text{mask}} \leftarrow$ trained mask generator

/* Stage 2 – Image synthesis */
Initialise ControlNet weights $\Phi \leftarrow 0$
while training do
    Sample $(I_k, M_k, \text{caption}_k)$
    $I_{\text{pred}} \leftarrow \text{SD}_{\text{base}} + \text{CN}_{\Phi}(M_k, \text{caption}_k)$
    $L_{\text{img}} \leftarrow \| I_{\text{pred}} - I_k \|_2^2 + \text{PerceptualLoss}$
    Update Φ with ∇L_{img}
end while
$\text{SD}_{\text{img}} \leftarrow$ trained image generator

/* Stage 3 – Segmentation training */
$\mathcal{D}_{\text{syn}} \leftarrow \varnothing$
for $n = 1$ to N_{syn} do
    $\text{caption}_{\text{rnd}} \leftarrow$ random prompt
    $M_{\text{syn}} \leftarrow \text{SD}_{\text{mask}}(\text{caption}_{\text{rnd}})$
    $I_{\text{syn}} \leftarrow \text{SD}_{\text{img}}(M_{\text{syn}}, \text{caption}_{\text{rnd}})$
    $\mathcal{D}_{\text{syn}} \leftarrow \mathcal{D}_{\text{syn}} \cup (I_{\text{syn}}, M_{\text{syn}})$
end for
$\mathcal{D}_{\text{train}} \leftarrow \mathcal{D}_{\text{real}} \cup \text{Subsample}(\mathcal{D}_{\text{syn}}, \alpha)$
Train segmentation net f_{seg} on $\mathcal{D}_{\text{train}}$
Evaluate on validation set; return θ_{seg}

3. Results and Discussion

3.1. Dataset

In this study, we use the ISPRS Potsdam dataset [7], as shown in Figure 7. Released by the International Society for Photogrammetry and Remote Sensing (ISPRS) in 2014, it comprises high-resolution, true-color (RGB) aerial images covering a 4 km^2 area of Potsdam, Germany. These near-vertical images come with high-quality manual labels for six classes: impervious surfaces, buildings, low vegetation, trees, cars, and clutter/background. The dataset contains 38 images, each approximately 6000 × 6000 pixels. To prepare for model training, we partitioned each large image into 512 × 512 patches and randomly split them into training and validation sets in a 9:5 ratio.



Figure 7. ISPRS Potsdam dataset example images [7].

To supplement the missing textual labels in the ISPRS Potsdam dataset, we analyzed the 3456 training patches in detail. Using a custom script, we computed the area proportions of each color-coded class region within the segmentation masks. We then standardized these proportions in the format "[Class: Rate]" to generate corresponding text labels. For example, if buildings occupy 30% of a patch, the text label includes "building: 30%." Figure 8 shows examples of these generated label–text pairs. Simultaneously, we employed a multimodal large language model (MLLM) to generate descriptive captions for the Potsdam images and further refined them with a language model to ensure consistent formatting and professional terminology. Figure 9 presents samples of our images alongside their final textual descriptions.

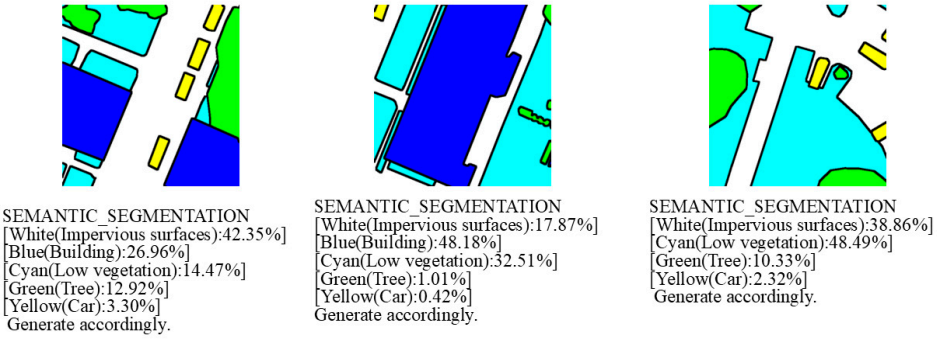


Figure 8. Example of label text annotation.

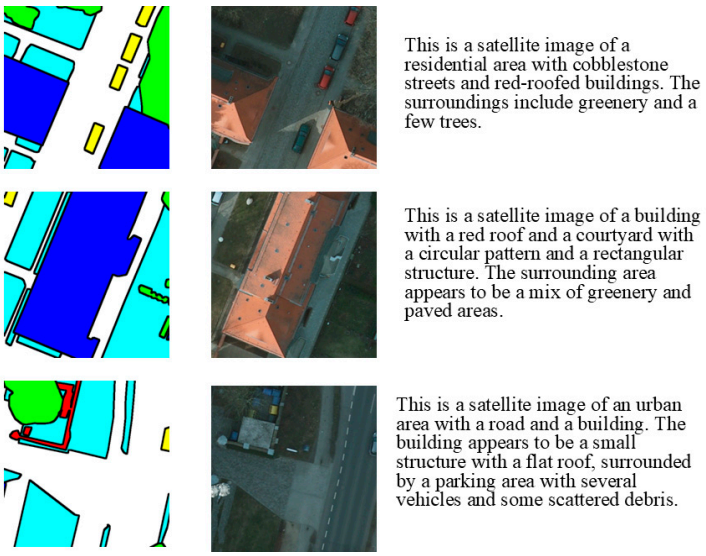


Figure 9. Example image and its synthesized text description.

3.2. Evaluation Metrics and Parameter Settings

3.2.1. Evaluation Metrics

To comprehensively evaluate the performance of the generated samples and the downstream semantic segmentation models, we employ the following evaluation metrics:

1. Mean Intersection Over Union (mIoU): mIoU is a primary metric for segmentation performance, defined as the average of the IoU across all classes. The mIoU can be defined by Equation (5):

$$\text{mIoU} = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FP_i + FN_i} \tag{5}$$

   where N is the total number of categories, TPi is the number of pixels correctly categorized as positive instances in category i, FPi is the number of pixels incorrectly categorized as positive instances in category i, and FNi is the number of pixels incorrectly categorized as negative instances in category i.
2. Mean $F1$ score (FI): The average $F1$ score can be defined by Equation (6):

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{6}$$

   where $Precision$ denotes the proportion of positive cases predicted by the model that are actually positive, calculated as TP/(TP + FP), and $Recall$ denotes the proportion of positive cases that are actually positive and successfully predicted by the model, calculated as TP/(TP + FP). The mean $F1$ score is the average of the multi-classes, i.e., $\text{mF1} = \frac{1}{N}\sum_{1}^{N} F1_i$.
3. Fréchet Inception Distance (FID): This measure the distributional difference between generated images and real images; a lower FID indicates higher generation quality. Its computation is defined by Equation (7):

$$FID = \sqrt{(\mu_r - \mu_g)^2 + T_r\left(\sum_r \sum_g\right)} \tag{7}$$

   Here, μ_r and μ_g denote the feature means of the real and generated images, respectively; $\sum_r$ and $\sum_g$ denote the feature covariance matrices of the real and generated images, respectively.

4. Mean Class Proportion Error (mCPE): This metric quantifies the deviation between the class proportions in the generated labels and the predefined target proportions. Lower mCPE values indicate that the generated labels align more closely with the expected class distribution. mCPE is represented by Equation (8):

$$mCPE = \frac{1}{nm} \sum_{i=1}^{n} \left| p_{gen,i} - p_{prmopt,i} \right| \quad (8)$$

where n is the total number of samples and m is the total number of classes; $p_{gen,i}$ is the proportion of class i in the generated labels; and $p_{prmopt,i}$ is the proportion of class i in the text-annotated labels. The absolute difference $\left| p_{gen,i} - p_{prmopt,i} \right|$ represents the proportion error for class i. The chief advantage of mCPE is that it quantifies how far the generated label distribution deviates from the desired class proportions and, in a straightforward numeric form, reflects the model's effectiveness at proportion control.

3.2.2. Parameter Settings

All experiments in this study were conducted with the PyTorch deep learning framework (version 2.1.2 with CUDA 11.8) on Ubuntu 22.04. The LoRA and ControlNet modules were trained on a single NVIDIA A100 GPU (40 GB VRAM). LoRA is trained at half precision (FP16) with a batch size of 16, using the AdamW optimizer (weight decay enabled). The initial learning rate was 1×10^{-4}; β was set to (0.9, 0.999); and the LoRA rank dimension was 32. This rank value was chosen as a trade-off between adaptation capacity and efficiency: we found that smaller ranks (e.g., 8 or 16) led to underfitting, while higher ranks showed diminishing returns. A fixed random seed 42 was used, and training ran for 50 epochs. ControlNet is trained in full precision (FP32) with a batch size of 12. The initial learning rate was reduced to 1×10^{-5} (an order of magnitude lower than LoRA's to ensure stable convergence), while all other optimizer settings matched those of LoRA. Training ControlNet for fine-tuning lasted 30 epochs.

For the downstream segmentation task, limited hardware resources dictated the use of an NVIDIA RTX 4090 GPU (24 GB VRAM). The batch size was set to 4; AdamW (weight decay enabled) was again employed, with the initial learning rate raised to 5×10^{-4} and β kept at (0.9, 0.999).

All experiments used a constant learning-rate schedule. Reproducibility was ensured by fixing the random seed at 42 in both the LoRA and ControlNet runs. Hardware selection was tailored to VRAM requirements: the A100 handled large-model training, while the RTX 4090 sufficed for the lighter downstream tasks.

On our hardware, generating one 512×512 synthetic sample (a segmentation mask together with its corresponding image) takes roughly 1.5 s in inference. In particular, label generation requires about 0.6 s and image generation about 0.9 s per sample on an NVIDIA A100 GPU, demonstrating the practicality of the framework for large-scale data augmentation.

3.3. Comparative Experiments

3.3.1. Comparison of Different Models

Using the ISPRS Potsdam label–text–image triplet dataset, we compared the sample-generation quality of the proposed method with two widely used baselines, Pix2Pix and CycleGAN. As reported in Table 1, our method achieves the lowest FID, indicating that its generated images are more realistic in both fine details and global structure. Figure 10 shows sample outputs, where the proposed approach clearly surpasses the baselines in clarity and structural plausibility.

Table 1. FID metrics of different generative models.

Method	FID
Pix2Pix	129.09
CycleGAN	144.72
Ours	92.59

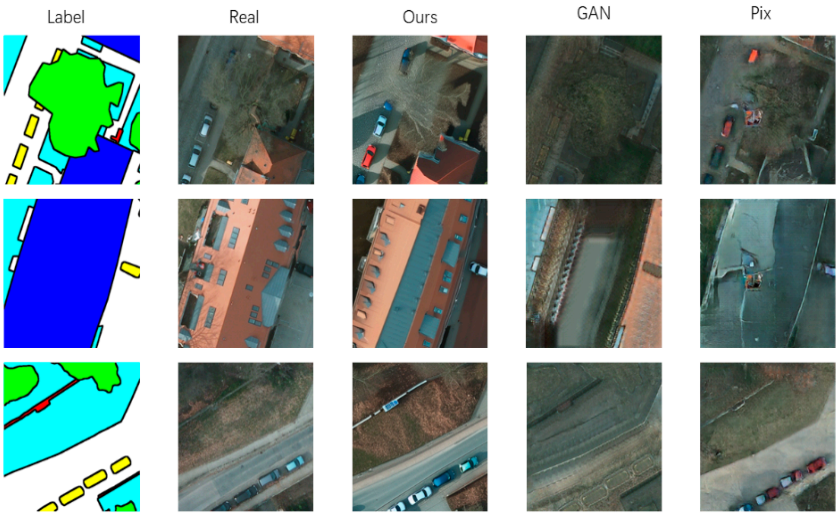


Figure 10. Examples of synthetic data generated by different models. Label colors: impervious surfaces (white), buildings (blue), low vegetation (cyan), trees (green), cars (yellow), and clutter (red).

3.3.2. Evaluation on Additional Dataset

To validate the generalizability of our approach, we conducted experiments on the ISPRS Vaihingen dataset [2], which consists of 33 aerial images with the same six semantic classes as Potsdam. We applied our entire pipeline to Vaihingen using similar training settings. Qualitatively, the generated samples maintained good realism and accurate alignment with the conditioning labels. Quantitatively, our method again achieved substantially better image quality than the baselines (FID = 125.29 for ours vs. 165.09 for CycleGAN), indicating that the synthetic images are highly faithful to real data distribution; the FID lower than Potsdam datasets is mainly because of the small sample size of the Vaihingen dataset (only 344 images for training where Potsdam has 3456 images). In downstream segmentation, adding synthetic samples led to notable performance gains. Using 40% synthetic data improved the DeepLabV3+ model's mIoU on Vaihingen from 69.28% to 70.67%. The smallest class ("Clutter") in Vaihingen saw its IoU increase from about 37.84% (no augmentation) to 41.26% with our synthetic augmentation. These improvements, similar to those observed for Potsdam, confirm that the proposed framework is broadly effective and not limited to a single dataset or sensor.

3.3.3. Comparison with Random Oversampling

In addition, we compare our approach with a simpler baseline that addresses class imbalance by random oversampling of the existing training data. In this baseline, we increased the representation of minority-class examples by duplicating images/patches of those classes (without generating new imagery). We found that oversampling provided only limited benefits. For example, for DeepLabV3+ on Potsdam, the overall mIoU improved from ~72.1% to ~74.2%, was below 74.7% and 76.0% when adding 20% and 40% synthetic data, and the Clutter class IoU rose from 20.4% to about 38.6% with oversampling, below the 39.9% Clutter IoU achieved by our synthetic data approach (at 40% synthetic data). Furthermore, on the Vaihingen dataset when oversampling 40% samples, the mIoU even

decreased from 69.2% to 68.9%, which indicates oversampling simply replicates existing data and can lead to over-fitting to those samples, without introducing new variations. By contrast, our diffusion-based method creates novel examples that expand the data distribution. These results demonstrate that naïve oversampling is much less effective at mitigating class imbalance than the proposed synthetic sample generation strategy.

3.4. Ablation Study

To quantify the contribution of each component to overall performance, we designed a series of ablation experiments.

3.4.1. Effect of Proportion-Aware Loss

In the label generation stage, we introduced a proportion-aware loss to mitigate the impact of class imbalance. Table 2 compares the mean class proportion error (mCPE) before and after adding this loss. The results show a marked reduction in class proportion error, indicating more accurate label generation. Figure 11 presents example masks produced from text prompts, visually demonstrating the loss function's effect on the generated labels. After incorporating the proportion-aware loss, the class distribution in the outputs becomes noticeably more balanced, yielding masks that better match the target class proportions.

Table 2. Quantitative effect of the proportion-aware loss on label generation (lower MCPE is better).

Prompt	Loss	Impervious Surface	Building	Low Vegetation	Tree	Car	Clutter	mCPE
√	√	2.57	23.76	4.24	18.98	1.44	1.24	8.705
√	-	3.18	26.01	5.75	24.29	1.66	3.69	10.76

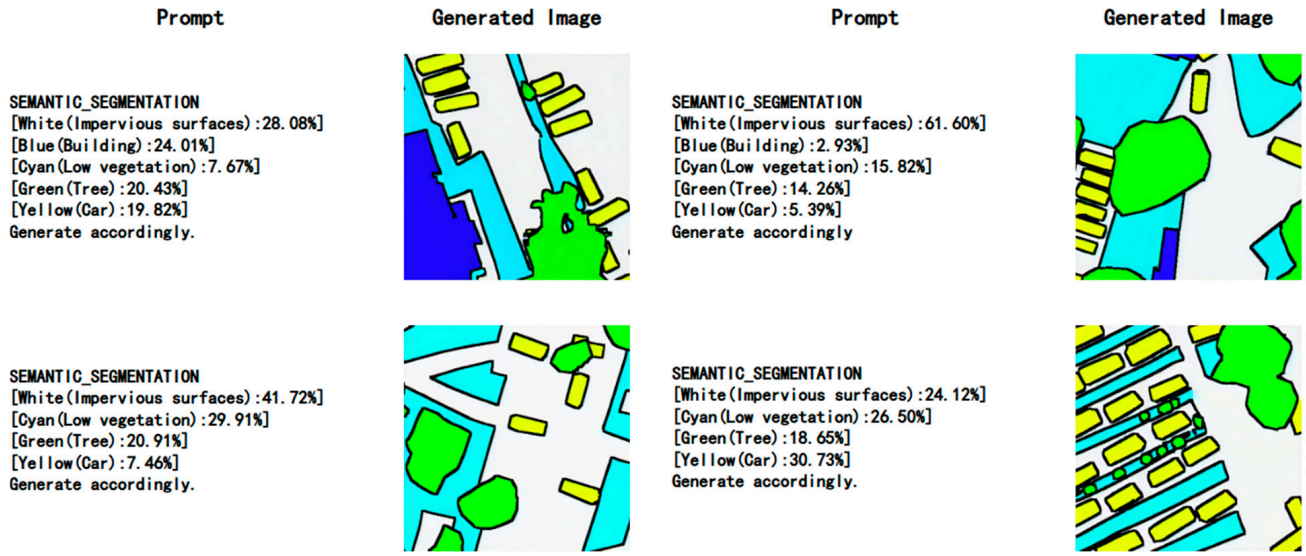


Figure 11. Sample text label pair produced.

3.4.2. Effect of Synthetic Sample Proportion

We conducted an experiment series to examine how the number of synthetic samples affects segmentation performance. Figure 12 depicts the class distribution in the original ISPRS Potsdam dataset, while Figure 13 shows the class distribution in the synthetic samples generated by our method. Notably, classes with low representation (Car and Clutter) in the original data are substantially enriched in the synthetic dataset.

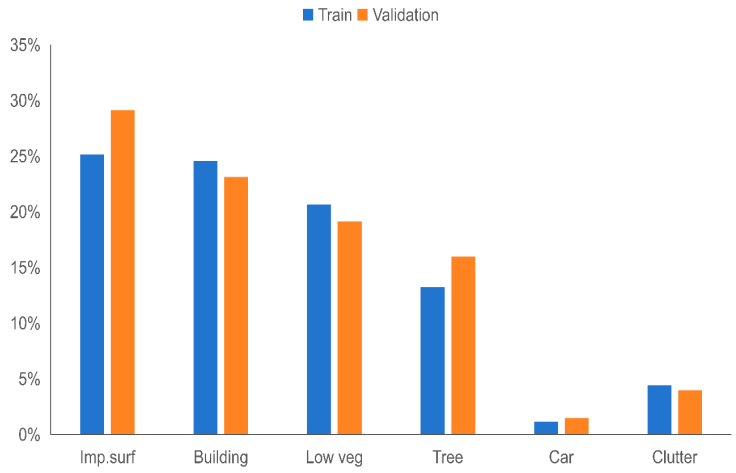


Figure 12. The percentage distribution of samples for each class in ISPRS Potsdam dataset.

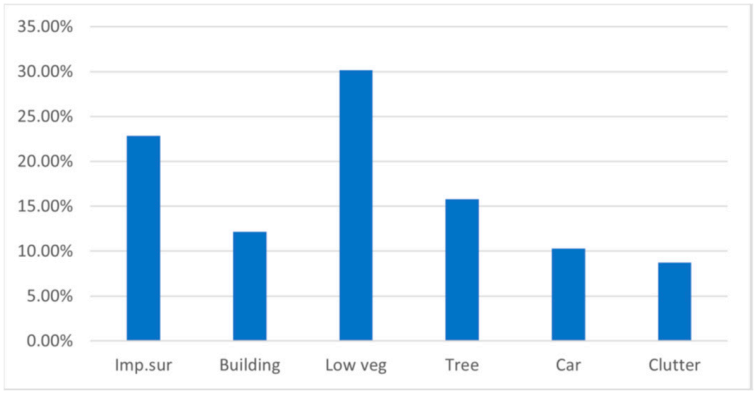


Figure 13. The percentage distribution of samples for each class in the synthetic samples.

Using the DeepLabV3+ model, we trained on the original dataset augmented with different proportions of synthetic samples. Figure 14 charts the quantitative metrics (mIoU and mF1): as the share of synthetic samples increases, both metrics first rise and then fall, peaking when synthetic samples comprise about 40% of the training set.

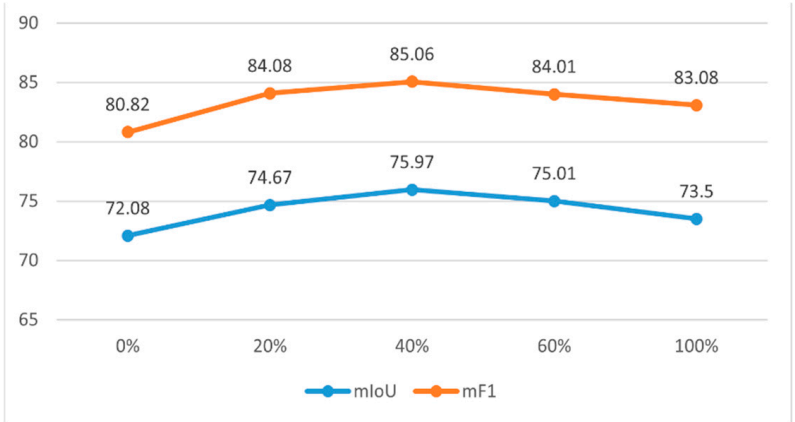


Figure 14. Changes in model metrics under different synthetic sample ratios.

Table 3 details class-wise segmentation performance at each synthetic data ratio. Adding synthetic samples significantly improves the segmentation of minority classes (Car and Clutter) while having virtually no adverse effect on the other classes.

Table 3. Metrics of DeeplabV3+ under different synthetic sample ratios.

Synthetic Sample Ratio	Impervious Surface (mIoU/F1)	Building	Low Vegetation	Tree	Car	Clutter
0%	84.15/91.39	90.87/95.22	73.05/84.42	76.66/86.79	87.35/93.25	20.39/33.87
20%	84.24/91.45	89.02/94.19	72.71/84.20	75.25/82.31	89.55/94.49	37.25/54.28
40%	84.56/91.63	90.30/94.90	73.37/84.64	77.03/87.02	90.62/95.08	39.92/57.06
60%	84.52/91.61	90.47/95.00	74.21/85.19	78.37/87.87	88.93/94.14	33.55/50.25
100%	81.96/90.09	88.07/93.66	71.76/83.56	76.71/86.82	89.00/94.18	33.51/50.20

Figure 15 shows qualitative segmentation results across the same ratios. The visual quality mirrors the quantitative trend: boundaries become sharper and overall accuracy improves up to roughly 40% synthetic data, after which over-fitting and occasional mis-segmentations emerge in some regions.

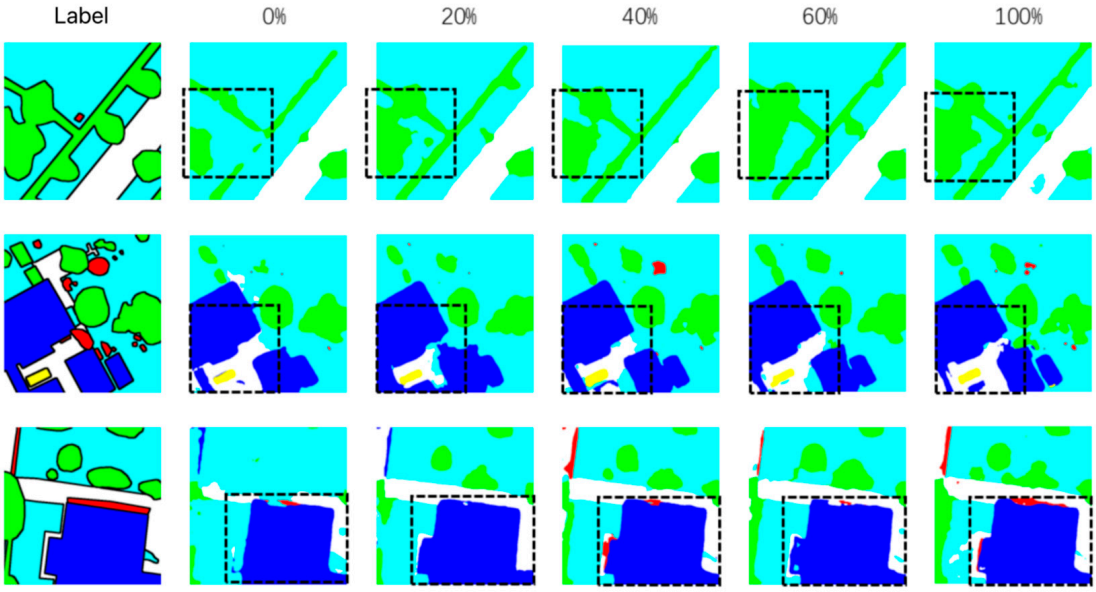


Figure 15. Segmentation results of DeeplabV3+ under different synthetic sample ratios.

3.4.3. Performance Across Different Models

We further evaluated the approach using PSPNet and SegFormer. In these experiments, synthetic samples were set to 40% of the real sample data volume. As summarized in Table 4, incorporating the synthetic samples improves the overall performance of every model tested, confirming the robustness of the proposed method. Notably, the largest gains are again seen in the smallest classes (e.g., Clutter), consistent with our earlier observations.

Table 4. Metrics of PSPNet and SegFormer under different synthetic sample ratios.

Model/Synthetic Sample Ratio	IoU/F1						mIoU	mF1
	Impervious Surface	Building	Low Vegetation	Tree	Car	Clutter		
PSPNet/0%	83.91/91.25	87.86/93.54	70.58/82.75	75.70/86.17	88.27/93.17	33.24/49.89	73.26	82.89
PSPNet/40%	84.48/91.59	89.61/94.52	72.79/84.25	76.73/86.83	89.26/94.32	33.32/49.98	74.36	83.58

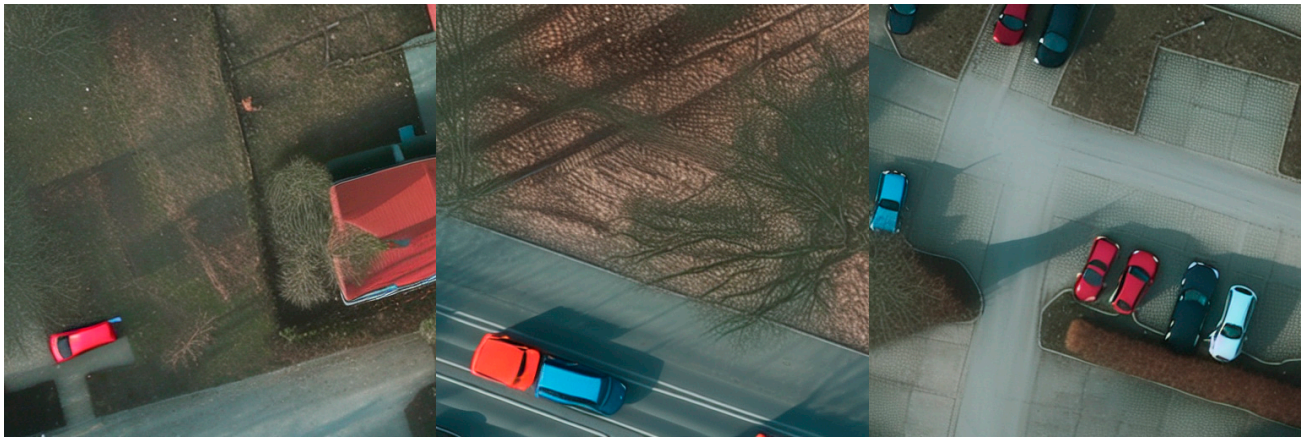
4. Conclusions

This study tackles the challenges of sample scarcity and class imbalance in remote sensing semantic segmentation by presenting a decoupled synthetic sample generation framework built on a latent diffusion model. Separating label generation from image generation, the approach eliminates the mutual constraints between the two processes. For

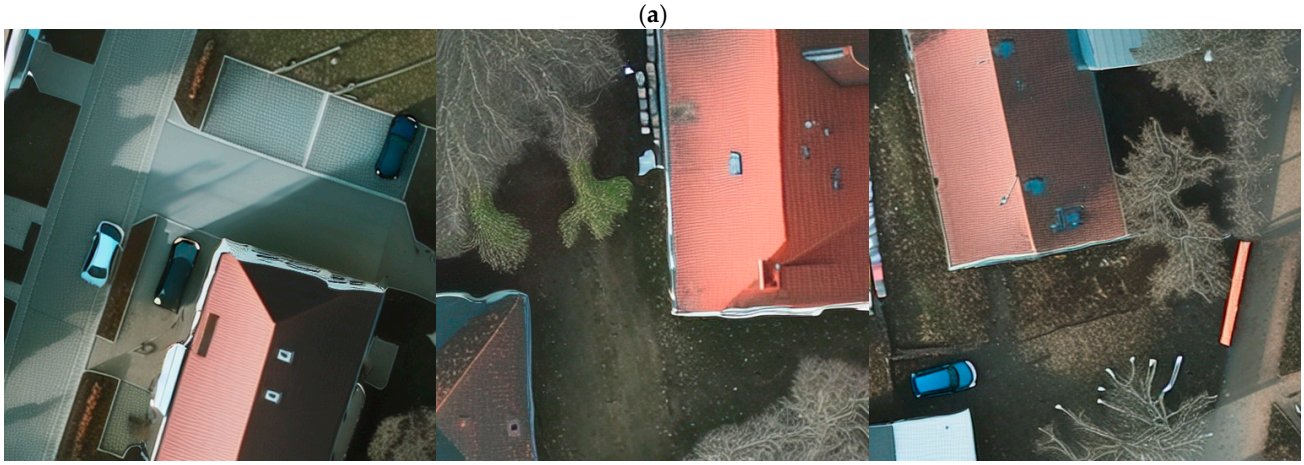
label generation, high-quality semantic masks are produced under limited computational resources by leveraging LoRA. A novel proportion-aware loss is introduced to enforce balanced class distributions. For image generation, a multi-condition generator based on ControlNet is trained without altering the original Stable Diffusion weights, and an external condition network ensures that the generated images remain structurally and semantically aligned with the input masks.

Experiments demonstrate that the synthetic samples produced by the proposed method achieve high visual fidelity and semantic consistency. The proportion-aware loss effectively mitigates the influence of minority classes, improving segmentation performance on under-represented categories. Results further show that introducing a moderate share of synthetic samples boosts overall segmentation accuracy (peaking at roughly 40%) whereas an excessive proportion can lead to over-fitting or misclassification. Cross-model tests with PSPNet, DeepLabV3+, and SegFormer confirm the method's robustness and generality. Overall, the framework can automatically generate high-quality, class-balanced image–label pairs, substantially alleviating data shortages in segmentation tasks.

However, we also observed some failure cases, which may be attributed to the method's lack of domain knowledge specific to oblique remote sensing imagery and the absence of instance-level structural constraints. As illustrated in Figure 16, the first three examples reveal incorrect internal physical scales—for example, distorted and disproportionate vehicle structures. The middle three exhibit unsmooth boundaries, likely due to missing side-view parallax information essential for accurate building delineation. The last three highlight the absence of intra-class structural cues, such as court lines on sports fields or lane markings on roads.



(**a**)



(**b**)

Figure 16. *Cont.*



(**c**)

Figure 16. Examples of failure cases. (**a**) Incorrect internal physical scales: distorted and disproportionate vehicle structures; (**b**) Unsmooth boundaries due to missing side-view parallax information; (**c**) Absence of intra-class structural cues: unreasonable court lines on sports fields or lane markings on roads.

Despite the encouraging results on high-resolution imagery, several limitations remain to be addressed in future work, such as (1) Applicability to lower resolutions: The current study targets high-resolution data, while further validation is needed for lower-resolution imagery. Integrating super-resolution techniques could help enhance fine details when applying the model to coarse inputs. (2) Hyperparameter sensitivity: The weight of the proportion-aware loss and the ratio of synthetic to real samples strongly affect outcomes and are presently tuned empirically. Future research should explore adaptive schemes or meta-learning strategies to automatically set these parameters. (3) Computational cost: Although LoRA and ControlNet reduce resource demands, training our diffusion models still requires substantial GPU time. More efficient fine-tuning strategies or knowledge distillation could cut the cost of sample generation. (4) Incorporation of geographic priors: Adding external knowledge (e.g., location, season, or meteorological conditions) may further improve the semantic plausibility and diversity of generated data. Exploring these avenues could lead to even more effective and scalable synthetic data generation methods for remote sensing semantic segmentation.

Author Contributions: Conceptualization, methodology, writing—review and editing, Y.X.; methodology, validation, writing—original draft preparation, H.L.; supervision R.Y.; project administration, Z.C. All authors have read and agreed to the published version of the manuscript.

Funding: The research was supported by the Science and Disruptive Technology Program, Aerospace Information Research Institute, Chinese Academy of Sciences, under Grant E3Z219010F.

Data Availability Statement: The data presented in this study are available on request.

Acknowledgments: The authors are grateful to the editors and anonymous reviewers for their informative suggestions.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

GAN	Generative Adversarial Network
LDM	Latent Diffusion Model
LoRA	Low-Rank Adaptation
MLLM	Multimodal Large Language Model

mIoU	Mean Intersection over Union
mF1	Mean F1 Score (average per-class F1)
mCPE	Mean Class Proportion Error
VAE	Variational Autoencoder

References

1. Zhu, M.; Zhu, D.; Huang, M.; Gong, D.; Li, S.; Xia, Y.; Lin, H.; Altan, O. Assessing the Impact of Climate Change on the Landscape Stability in the Mediterranean World Heritage Site Based on Multi-Sourced Remote Sensing Data: A Case Study of the Causses and Cévennes, France. *Remote Sens.* **2025**, *17*, 203. [CrossRef]
2. Qiu, X.; Zhang, Z.; Luo, X.; Zhang, X.; Yang, Y.; Wu, Y.; Su, J. Semantic Uncertainty-Awared for Semantic Segmentation of Remote Sensing Images. *IET Image Process.* **2025**, *19*, e70045. [CrossRef]
3. Wang, J.; Chen, T.; Zheng, L.; Tie, J.; Zhang, Y.; Chen, P.; Luo, Z.; Song, Q. A Multi-Scale Remote Sensing Semantic Segmentation Model with Boundary Enhancement Based on UNetFormer. *Sci. Rep.* **2025**, *15*, 14737. [CrossRef] [PubMed]
4. Danish, M.U.; Buwaneswaran, M.; Fonseka, T.; Grolinger, K. Graph Attention Convolutional U-Net: A Semantic Segmentation Model for Identifying Flooded Areas. In Proceedings of the IECON 2024-50th Annual Conference of the IEEE Industrial Electronics Society, Chicago, IL, USA, 3–6 November 2024; IEEE: New York, NY, USA, 2024; pp. 1–6.
5. Gong, D.; Huang, M.; Ge, Y.; Zhu, D.; Chen, J.; Chen, Y.; Zhang, L.; Hu, B.; Lai, S.; Lin, H. Revolutionizing Ecological Security Pattern with Multi-Source Data and Deep Learning: An Adaptive Generation Approach. *Ecol. Indic.* **2025**, *173*, 113315. [CrossRef]
6. Diakogiannis, F.I.; Waldner, F.; Caccetta, P.; Wu, C. ResUNet-a: A Deep Learning Framework for Semantic Segmentation of Remotely Sensed Data. *ISPRS J. Photogramm. Remote Sens.* **2020**, *162*, 94–114. [CrossRef]
7. ISPRS. ISPRS Test Project on Urban Classification, 3D Building Reconstruction and Semantic Labeling. Available online: https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/semantic-labeling.aspx (accessed on 20 April 2025).
8. Wang, J.; Zheng, Z.; Ma, A.; Lu, X.; Zhong, Y. LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation. *arXiv* **2021**, arXiv:2110.08733.
9. Zermatten, V.; Lu, X.; Castillo-Navarro, J.; Kellenberger, T.; Tuia, D. Land Cover Mapping from Multiple Complementary Experts under Heavy Class Imbalance. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 6468–6477. [CrossRef]
10. Harshvardhan, G.M.; Gourisaria, M.K.; Pandey, M.; Rautaray, S.S. A Comprehensive Survey and Analysis of Generative Models in Machine Learning. *Comput. Sci. Rev.* **2020**, *38*, 100285.
11. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv* **2013**, arXiv:1312.6114.
12. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2672–2680.
13. Hinz, T.; Fisher, M.; Wang, O.; Wermter, S. Improved Techniques for Training Single-Image Gans. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2021; pp. 1300–1309.
14. Niu, Z.; Li, Y.; Gong, Y.; Zhang, B.; He, Y.; Zhang, J.; Tian, M.; He, L. Multi-Class Guided GAN for Remote-Sensing Image Synthesis Based on Semantic Labels. *Remote Sens.* **2025**, *17*, 344. [CrossRef]
15. Graves, A. Generating Sequences With Recurrent Neural Networks. *arXiv* **2013**, arXiv:1308.0850.
16. Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; Ganguli, S. Deep Unsupervised Learning Using Nonequilibrium Thermodynamics. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2256–2265.
17. Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
18. Song, J.; Meng, C.; Ermon, S. Denoising Diffusion Implicit Models. *arXiv* **2020**, arXiv:2010.02502.
19. Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; Chen, M. Hierarchical Text-Conditional Image Generation with Clip Latents. *arXiv* **2022**, arXiv:2204.06125.
20. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10684–10695.
21. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. Lora: Low-Rank Adaptation of Large Language Models. *ICLR* **2022**, *1*, 3.
22. Zhang, L.; Rao, A.; Agrawala, M. Adding Conditional Control to Text-to-Image Diffusion Models. In Proceedings of the IEEE/CVF International Conference On Computer Vision, Paris, France, 1–6 October 2023; pp. 3836–3847.
23. Toker, A.; Eisenberger, M.; Cremers, D.; Leal-Taixé, L. Satsynth: Augmenting Image-Mask Pairs through Diffusion Models for Aerial Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 27695–27705.
24. Zhao, C.; Ogawa, Y.; Chen, S.; Yang, Z.; Sekimoto, Y. Label Freedom: Stable Diffusion for Remote Sensing Image Semantic Segmentation Data Generation. In Proceedings of the 2023 IEEE International Conference on Big Data (BigData), Sorrento, Italy, 15–18 December 2023; IEEE: New York, NY, USA, 2023; pp. 1022–1030.

25. Tang, D.; Cao, X.; Hou, X.; Jiang, Z.; Liu, J.; Meng, D. CRS-Diff: Controllable Remote Sensing Image Generation with Diffusion Model. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5638714. [CrossRef]
26. Khanna, S.; Liu, P.; Zhou, L.; Meng, C.; Rombach, R.; Burke, M.; Lobell, D.; Ermon, S. DiffusionSat: A Generative Foundation Model for Satellite Imagery. *arXiv* **2023**, arXiv:2312.03606.
27. Zhu, C.; Liu, Y.; Huang, S.; Wang, F. Taming a Diffusion Model to Revitalize Remote Sensing Image Super-Resolution. *Remote Sens.* **2025**, *17*, 1348. [CrossRef]
28. Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W. Qwen2-vl: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv* **2024**, arXiv:2409.12191.
29. Isola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
30. Zhu, J.-Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
31. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
32. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
33. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
34. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models from Natural Language Supervision. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 8748–8763.

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.