drones

MDPI

Article

Expanding Open-Vocabulary Understanding for UAV Aerial Imagery: A Vision–Language Framework to Semantic Segmentation

Bangju Huang [1], Junhui Li [1,*], Wuyang Luan [2], Jintao Tan [1], Chenglong Li [1] and Longyang Huang [1]

[1] College of Air Traffic Management, Civil Aviation Flight University of China, Deyang 618307, China; huangbj@cafuc.edu.cn (B.H.); tjt@cafuc.edu.cn (J.T.); lcl@cafuc.edu.cn (C.L.); longyanghuang@cafuc.edu.cn (L.H.)
[2] School of Computer Science, Civil Aviation Flight University of China, Deyang 618307, China; lwy@cafuc.edu.cn
* Correspondence: junhuili@cafuc.edu.cn

Abstract: The open-vocabulary understanding of UAV aerial images plays a crucial role in enhancing the intelligence level of remote sensing applications, such as disaster assessment, precision agriculture, and urban planning. In this paper, we propose an innovative open-vocabulary model for UAV images, which combines vision–language methods to achieve efficient recognition and segmentation of unseen categories by generating multi-view image descriptions and feature extraction. To enhance the generalization ability and robustness of the model, we adopted Mixup technology to blend multiple UAV images, generating more diverse and representative training data. To address the limitations of existing open-vocabulary models in UAV image analysis, we leverage the GPT model to generate accurate and professional text descriptions of aerial images, ensuring contextual relevance and precision. The image encoder utilizes a U-Net with Mamba architecture to extract key point information through edge detection and partition pooling, further improving the effectiveness of feature representation. The text encoder employs a fine-tuned BERT model to convert text descriptions of UAV images into feature vectors. Three key loss functions were designed: Generalization Loss to balance old and new category scores, semantic segmentation loss to evaluate model performance on UAV image segmentation tasks, and Triplet Loss to enhance the model's ability to distinguish features. The Comprehensive Loss Function integrates these terms to ensure robust performance in complex UAV segmentation tasks. Experimental results demonstrate that the proposed method has significant advantages in handling unseen categories and achieving high accuracy in UAV image segmentation tasks, showcasing its potential for practical applications in diverse aerial imagery scenarios.

Keywords: UAV imagery; semantic segmentation; open vocabulary; vision–language framework

check for updates

Academic Editor: Joaquín Martínez-Sánchez

Received: 25 December 2024
Revised: 14 February 2025
Accepted: 17 February 2025
Published: 19 February 2025

Citation: Huang, B.; Li, J.; Luan, W.; Tan, J.; Li, C.; Huang, L. Expanding Open-Vocabulary Understanding for UAV Aerial Imagery: A Vision–Language Framework to Semantic Segmentation. *Drones* **2025**, *9*, 155. https://doi.org/10.3390/drones9020155

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

Two-dimensional unmanned aerial vehicle (UAV) aerial image understanding plays a vital role in applications such as disaster assessment, precision agriculture, and urban planning. In recent years, deep learning techniques have made significant progress in this field. However, existing deep learning models mainly rely on training data with human annotations, which means that the model can only recognize categories that appear in the training data and cannot handle new categories that have not been seen before. The

predictive power of this closed vocabulary greatly limits the universality and flexibility of the model in practical applications. Especially in aerial imagery, unseen categories may represent rare land cover types or new structural features, and the acquisition of manual annotations is time-consuming and costly.

The research of two-dimensional open-vocabulary understanding has made important progress in the field of computer vision. A lot of work attempts to solve open-vocabulary tasks through vision–language (VL) models, including object detection, semantic segmentation, and instance segmentation. Most of these methods rely on large-scale image–text pair data, with pre-trained VL models such as CLIP and ALIGN embedding visual and text features into a common representation space to enable the recognition of unseen categories [1]. However, these successful cases mainly focus on the understanding of natural images and scenes and do not address the unique challenges of UAV aerial imagery.

Applying the open-vocabulary approach [2–6] to UAV aerial imagery presents several challenges. First, UAV images significantly differ from natural images containing complex geographical and structural features. Second, annotation of UAV imagery is time-intensive and resource-heavy, making it difficult to create large-scale image–text pair datasets. Finally, UAV imagery applications require models with higher accuracy and robustness due to the diversity of terrains and environmental conditions [7,8].

To address these issues, we investigate open-vocabulary understanding of UAV aerial images, enabling the model to locate and identify unseen categories beyond the label space of the training data. Benefiting from Internet-scale image–text data, the VL foundation model has achieved breakthroughs in open-vocabulary understanding. However, directly applying these advances to UAV aerial imagery faces difficulties, including the lack of large-scale UAV image-to-text datasets.

To this end, we propose a method to generate multi-view UAV aerial image descriptions using the GPT model, explicitly associating 2D UAV images with rich semantic text. Furthermore, to facilitate coarse-to-fine visual–semantic representation learning, we design hierarchical image–text pairs that leverage the geometric constraints of UAV aerial images. Through contrastive learning, the model learns to embed language perception linking 2D UAV images and text for open-vocabulary tasks.

Our approach significantly outperforms baseline methods for semantic segmentation and instance segmentation of UAV aerial images and demonstrates strong transfer capabilities in challenging zero-shot domain adaptation tasks. This work demonstrates the potential of vision–language models to expand the vocabulary range of UAV image understanding, greatly improving the applicability of these models to real-world aerial image analysis tasks.

The organization of this paper is as follows. Section 1 introduces the background and motivation for developing an innovative open-vocabulary model tailored for UAV aerial images. Following this, we review existing literature on semantic segmentation techniques, including traditional image-based approaches and the latest methods incorporating textual information, in Section 2. In Section 3, we detail our proposed method, which includes BERT pre-training, Masked Language Modeling (MLM), Next Sentence Prediction (NSP) objectives, key components like Mixup image blending, caption generation and description, grid feature grouping, and the application of vision–language models. We then present extensive experimental results in Section 4, where we conduct experiments on multiple datasets, including the UAV-City dataset and the Cityscapes dataset, comparing the performance of different configurations and showcasing the impact of various modules on overall performance, with detailed ablation study results. And provides an in-depandth analysis of the experimental results, discussing performance differences across different categories and their underlying reasons, while also addressing the limitations of the current

approach and proposing potential directions for future improvements. Finally, in Section 5, we summarize the main contributions of our work and outline future research directions. By structuring the paper in this manner, we aim to clearly present the research background, methodological details, and final outcomes, making it easier for readers to understand and evaluate our work.

The main contributions of this paper are as follows:

- **Open-Vocabulary Model for UAV Aerial Imagery—**We propose an open-vocabulary model for UAV aerial imagery that leverages vision–language methods to establish the relationship between text and images, effectively solving the challenge of dataset annotation and enabling weakly supervised learning. By using pre-trained GPT to generate image descriptions, the model is capable of recognizing and processing unseen categories.
- **UAV Text-Image Bridge—**We introduce the UAV text–image bridge, which employs an image-to-text model and text feature vector transformation, enabling the model to learn rich semantic information from text and images. Specifically, we use a U-Net with Mamba architecture for image encoding and a fine-tuned BERT-based text encoder to generate features for new category recognition, semantic segmentation, and instance segmentation.
- **Vision-Language Fusion for Enhanced Aerial Image Understanding—**Our proposed model leverages a combination of vision–language techniques to significantly expand the lexical range of UAV aerial image understanding and demonstrates strong transfer capabilities in challenging zero-shot tasks. By applying Mixup to multi-view images and extracting features from irregular partitions, the model effectively handles complex aerial scenes and improves recognition accuracy.

2. Related Work

2.1. 2D Open-Vocabulary

In the field of two-dimensional open-vocabulary understanding, many studies have addressed open-vocabulary tasks such as object detection and semantic segmentation by leveraging large-scale image–text data and vision–language (VL) models such as CLIP. These models have proven effective in bridging the gap between visual and textual modalities, allowing for flexible zero-shot learning on new categories that were not seen during training. The early work of Duygulu et al. [9] framed image segmentation as a machine translation problem, where the task of segmenting images was mapped to a language-based framework. This seminal work paved the way for more recent advancements, where researchers adapted pretrained VL models to tackle finer-grained tasks like pixel-level segmentation.

A prominent approach involves adapting image-level vision–language models, originally designed for image classification or image-to-text matching, to the more granular task of pixel-level image–text alignment. This strategy has been widely adopted in open-vocabulary segmentation methods. For instance, MaskCLIP [10] modifies the CLIP image encoder to enhance its pixel-level alignment capabilities without requiring additional retraining. This approach leverages the power of large-scale image–text data to guide the model in generating accurate segmentation maps from textual descriptions. Similarly, TCL [11] uses CLIP for initial grounding of text to image regions, followed by contrastive learning to fine-tune the alignment between the textual embedding and the regions in the image. Other methods, like OpenSeg [12], have explored fine-tuning existing VL models, such as ALIGN [13], with a grounding loss to better align words in captions to segmentation masks. While these methods employ external language models like CLIP to perform the alignment, they typically rely on supervised strategies, including annotated segments or pseudo-masks. Another line of work has focused on training models with ground truth

masks annotated for a limited set of seen categories. These approaches leverage detailed features learned from the annotated classes and use them to generalize to new, unseen categories. ZegFormer [14], for example, trains a MaskFormer by decoupling the task into two sub-tasks: a class-agnostic grouping task and a zero-shot segment classification task. In addition to the above strategies, custom-designed models for open-vocabulary segmentation have also gained attention. These models typically involve novel pixel grouping mechanisms that exploit the relationships between neighboring pixels or feature-based similarities. For instance, GroupViT [15] groups pixels in an image hierarchically using learnable group tokens based on attention scores, while OVSegmentor [16] utilizes Slot Attention [17] to perform pixel grouping based on feature proximity.

While these methods demonstrate powerful capabilities in most scenarios, they often struggle in specific domains, such as UAV aerial imagery. The primary challenges include acquiring large-scale image-to-text datasets and addressing complex scenes with fine-grained objects. Moreover, the direct application of these methods to aerial imagery can lead to performance degradation because they fail to effectively process unique geographical and structural features present in UAV datasets.

2.2. 2D Weakly Supervised Semantic Segmentation

Weakly Supervised Semantic Segmentation (WSSS) has become a prominent research area as it aims to reduce the reliance on pixel-level annotations by leveraging weak supervision signals, such as image-level labels, bounding boxes, or class activation maps (CAMs). Most WSSS methods follow a general three-stage learning process: (1) initial CAM generation, (2) pseudo-mask generation, and (3) training a segmentation model using the generated pseudo-masks. This approach enables the model to learn segmentation maps without needing fine-grained pixel-level annotations.

A key component in the first stage of the process is the generation of Class Activation Maps (CAMs), which highlight the regions of the image most relevant to a particular class. In early works, CAMs were directly used as the initial pseudo-masks for training the segmentation model. However, these CAMs often suffer from incomplete or noisy regions. To address this issue, Hou et al. [18] propose two self-erasing strategies aimed at refining the CAMs by focusing attention only on the most reliable regions. These strategies help produce more complete and precise activation maps, which serve as the foundation for generating better pseudo-masks in the subsequent stages.

To further improve the quality of the generated CAMs, other works have introduced methods to mine more informative object regions. For instance, Chang et al. [19] propose a strategy that investigates object sub-categories to discover additional parts of the object, thereby improving the granularity and completeness of the initial CAMs. Similarly, Jungbeom et al. [20] introduce an anti-adversarial approach, which aims to counteract noise and artifacts in the activation maps by discovering more regions of the target object. This helps to create more accurate initial segmentations, which are critical for effective pseudo-mask generation.

Once the initial CAMs are obtained, the next step is to generate pseudo-masks that are used to train the segmentation model. In this stage, the key challenge is to refine the coarse CAMs into more precise pseudo-masks that capture the object boundaries accurately. Ahn and Kwak [21] address this issue by proposing pixel-level semantic affinity, which encourages the model to refine the boundaries of the pseudo-masks, producing fine-grained segmentations that align better with the true object contours.

To further enhance the quality of the pseudo-masks, many methods incorporate additional cues, such as background information, which helps distinguish the foreground object from the background. Works like [22–25] employ background cues obtained from a

fully supervised saliency detector. These background cues are integrated into the pseudo-mask generation process, helping to mitigate errors caused by ambiguous regions in the CAMs and improving the overall accuracy of the segmentation model.

Although the above methods have made progress in the field of open-vocabulary weak supervision for 2D semantic segmentation, they face significant challenges when applied to UAV aerial imagery. First, UAV datasets often lack large-scale image–text pairs, limiting the generalization ability of pre-trained models. Second, the geographical and structural complexity in UAV images requires higher model accuracy and robustness. Finally, effectively utilizing unlabeled UAV imagery to improve the open-vocabulary capability of the model remains an open problem.

2.3. Weak Supervision in UAV Image Segmentation

Weakly supervised learning has gained popularity in UAV image segmentation due to the difficulty and high cost of obtaining large-scale pixel-level annotations. In UAV segmentation, weak supervision typically involves using more accessible signals like image-level labels, bounding boxes, class activation maps (CAMs), or point-level annotations. While these methods reduce the manual annotation effort, challenges such as class imbalance, CAM over-activation, and reliance on additional supervision still remain.WSF-NET [26] addresses class imbalance in UAV segmentation by using a binary segmentation framework and a feature fusion network. While this approach adapts to the unique characteristics of remote sensing targets, it does not tackle CAM over-activation, which can lead to misclassification of non-target regions as part of the object. SPMF-Net [27] uses image-level labels for supervision, along with superpixel pooling and multi-scale feature fusion. While this approach incorporates low-level feature information through superpixels, it is dependent on the accuracy of superpixel segmentation. Poor superpixels can cause misclassification, which is a challenge in complex UAV scenes. U-CAM [28] enhances CAMs by integrating point-level labels to provide object location information. Although this improves segmentation accuracy, it increases the annotation burden, as point-level labels require additional costs, making this method less scalable in large UAV datasets. CDSA [29] uses image-level labels to generate location maps and incorporates structural information from the source domain to address CAM over-activation. However, it still requires pixel-level labels to acquire source domain structure information, which contradicts the principles of weak supervision. Despite these advances, most weakly supervised methods in UAV image segmentation still face significant limitations. They often depend on domain-specific preprocessing techniques, feature engineering, or sensor-specific adjustments, making them less flexible when applied to new datasets or open-vocabulary segmentation tasks. Moreover, while multimodal and self-supervised approaches improve segmentation accuracy, they typically require large amounts of data from different modalities, which are not always available in UAV applications. Further research is needed to overcome these challenges, particularly in developing more robust and generalizable methods that can handle diverse, unlabeled UAV datasets in an open-vocabulary setting.

3. Method

3.1. Overall

We propose an innovative open-vocabulary model for UAV aerial images, which establishes the relationship between text and images through a vision–language approach and utilizes multi-perspective image descriptions, feature extraction, and weakly supervised learning to achieve efficient recognition and segmentation of unseen categories. The entire framework is shown in Figure 1, including multi-view image generation and description,

image encoding and feature extraction, visual-language model (VL Model), and feature vector-based classification and segmentation heads.

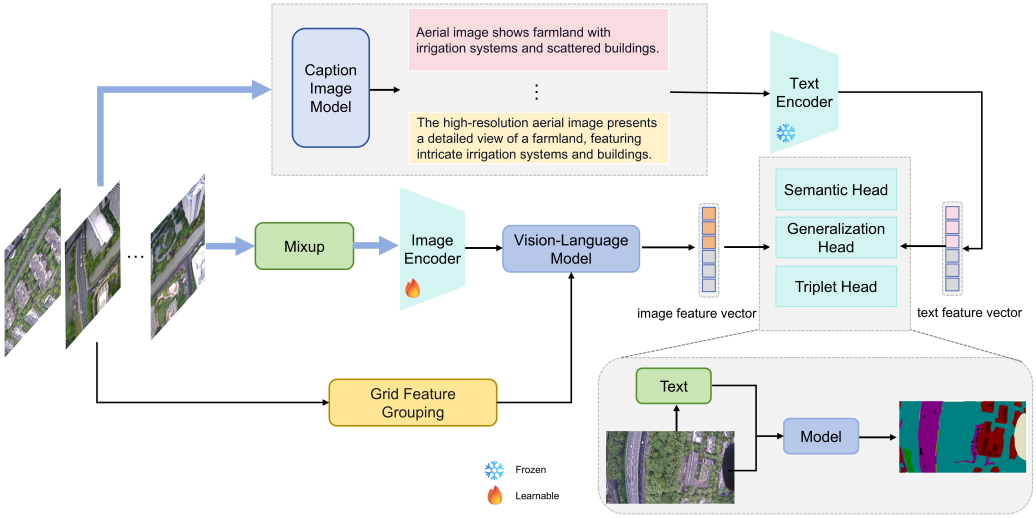


Figure 1. The architecture of our open-vocabulary model for UAV aerial images: First, multi-view image generation and description create diverse image perspectives with corresponding textual descriptions. Next, image encoding and feature extraction modules process these images to obtain high-level visual features. The visual-language (VL) model then connects the visual features with their textual descriptions, enabling multimodal understanding. Finally, classification and segmentation heads use the extracted features to recognize and segment unseen categories efficiently. This streamlined framework enables effective handling of novel categories in UAV aerial imagery through integrated vision–language processing.

3.2. Mix Image Generation

To enhance the recognition and classification ability of the model across multiple categories and avoid falling into local optimal solutions when encountering new categories, we adopt Mixup technology. A single UAV aerial image often covers fewer land-use or structural targets, which limits the model's ability to grasp diverse categories comprehensively. Therefore, we devised a method to stitch non-overlapping regions of multiple aerial images together to generate richer training data. By linearly mixing data from different images, new training samples are generated. Specifically, we splice multiple UAV images into squares according to size, ensuring that targets in each image do not overlap, thus preserving the integrity of each target. This process can be expressed by the following formula:

To enhance the recognition and classification ability of the model across multiple categories and avoid falling into local optimal solutions when encountering new categories, we adopt Mixup technology. A single UAV aerial image often covers fewer land-use or structural targets, which limits the model's ability to grasp diverse categories comprehensively. Therefore, we devised a method to stitch non-overlapping regions of multiple aerial images together to generate richer training data.

To generate new training samples by linearly mixing data from different images, we employ a concatenation method. Specifically, we concatenate multiple UAV images into a square arrangement, ensuring that targets in each image do not overlap, thereby preserving the integrity of each target. This process can be expressed by the following formula:

$$\hat{x} = \text{Concat}(x_1, x_2, \ldots, x_n) \tag{1}$$

where $\hat{x}$ represents the concatenated image and $x_1, x_2, \ldots, x_n$ represent the different input images, which are concatenated together to form a new image array.

The `Concat` function is used to concatenate multiple images along a specified dimension. For example, if each input image x_i has the shape (H, W, C), the shape of the concatenated image $\hat{x}$ will depend on the concatenation method. If we concatenate along the width dimension, the resulting image shape might be (H, nW, C), where n is the number of images being concatenated.

By generating new images in this way, not only do they contain more targets, but they also enhance the robustness of the model, making it less likely to fall into local optima when recognizing new categories. Additionally, the Mixup method increases the diversity of training data and further improves the generalization ability of the model.

3.3. Caption Generation and Description

In the field of UAV aerial imagery, due to the lack of specific models designed for generating text descriptions of aerial images, we use the GPT model to process image data. While existing image description models perform well on natural images, their application to aerial images is limited and often struggles to generate accurate and contextually relevant descriptions. To address this, we use the GPT model to manually generate image descriptions to ensure accuracy and domain relevance.

This process can be expressed by the following formula:

$$T_i = \mathrm{GPT}(x_i) \tag{2}$$

where T_i represents the text description of the image x_i and GPT represents the text generation process of the GPT model. The text descriptions generated by the GPT model, such as "Aerial image shows farmland with irrigation systems and scattered buildings", provide rich semantic information for the subsequent vision–language model.

By ensuring the accuracy and domain relevance of aerial image descriptions, we compensate for the shortcomings of existing image description models in the UAV field. These high-quality text descriptions enhance the model's semantic understanding capabilities and provide additional training data, further improving the model's performance in recognizing unseen categories.

3.4. Image Encoder

We use U-Net with Mamba architecture for image coding to process complex UAV aerial imagery data and enhance the feature extraction capability of the model. The structure of the image encoder is shown in Figure 2, including input embedding, feature transformation, State Space Model (SSM) module, activation function, and linear layer processing, and finally feature normalization.

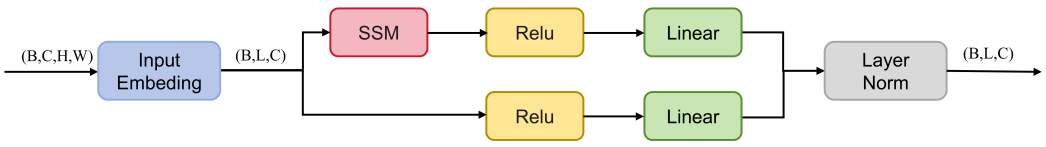


Figure 2. The detailed architecture of our Image Encoder, which combines a U-Net framework with Mamba-based State Space Model modules and dual-path processing to efficiently extract and normalize features from complex UAV aerial images.

The input embedding part involves transforming the input image data into feature vectors through a series of operations. Typically, we first apply a pooling operation followed by one-dimensional convolution to reduce the dimensionality. The specific formula for this process is as follows:

$$X = \text{Conv1D}(\text{Pooling}(\text{Input})) \tag{3}$$

where Input represents the input image tensor with shape (H, W, D), where H and W are the height and width of the image, and D is the number of channels. Pooling represents the pooling operation, which reduces the spatial dimensions of the input image. Conv1D represents the one-dimensional convolution operation, which processes the input along one spatial dimension. X represents the processed feature vector with shape (B, L, C), where B is the batch size, L is the length of the sequence (or reduced spatial dimension), and C is the number of channels in the feature map.

The image encoder consists of two paths: one uses SSM (Mamba) to process features, while the other uses ReLU activation functions to process features and then concatenates the output of the two. SSM modules are designed for efficient processing of sequence data, with linear time complexity and the ability to handle long sequence dependencies. The specific formulas are as follows:

$$h_{t+1} = Ah_t + Bx_t \tag{4}$$

$$y_t = Ch_t \tag{5}$$

where h_t represents the hidden state at time step t, x_t represents the input features at time step t, A is a matrix that governs the transition of the hidden state from one time step to the next, B is a matrix that scales the input features x_t before adding them to the current hidden state h_t, y_t is the output at time step t, and C is a matrix that maps the hidden state h_t to the output y_t.

It is important to note that the symbols B and C used here are different from those in Equation (3). In Equation (3), B represents the batch size, L represents the length of the sequence (or reduced spatial dimension), and C represents the number of channels in the feature map. In Equations (4) and (5), B and C are matrices used in the hidden state update and output generation processes, respectively.

The other path uses the ReLU activation function and a linear layer to process the features, as follows:

$$F_{\text{ReLU}} = \text{Linear}(\text{ReLU}(x)) \tag{6}$$

The output features of the two paths are spliced and normalized. The formulas are as follows:

$$F_{\text{concat}} = \text{concat}(F_{\text{SSM}}, F_{\text{ReLU}}) \tag{7}$$

$$F_{\text{output}} = \text{LayerNorm}(F_{\text{concat}}) \tag{8}$$

Among them, concat represents the feature concatenation operation, LayerNorm represents the layer normalization operation, and F_{output} represents the final output feature. In this way, our image encoders are able to efficiently process and extract important features from UAV aerial images, providing high-quality feature input for subsequent visual-language models.

3.5. Grid Feature Grouping

In order to improve the effectiveness of the feature information and enable the model to connect with text query vectors at a larger level, a Grid Feature Grouping method is designed. The method first detects the edge information in the image, and then performs partitioning pooling and clustering to obtain the key point information. This approach not only enhances the feature representation but also makes the model more efficient when aligned with text descriptions.

First, we detect the edge of the input image to identify the important edge and contour information in the image. Edge detection can be implemented using the Sobel operator or the Canny edge detection algorithm, as shown in the following formula:

$$E = \text{EdgeDetect}(I) \tag{9}$$

where E indicates the image after edge detection, I indicates the input image, and EdgeDetect represents the edge detection operation.

Next, we perform partitioning pooling and clustering operations on the image after edge detection to extract key point information. Partitioning pooling divides the image into several regions and performs pooling operations in each region to obtain the feature representation of each region. The formula for this operation is the following:

$$P_{ij} = \text{Pooling}(E_{ij}) \tag{10}$$

where P_{ij} indicates the pooling result of column j in row i, E_{ij} indicates the edge information in the area, and the pooling operation is applied. Then, we cluster the features of each pooled region to further extract key point information. The K-Means algorithm can be used for clustering, and the formula is as follows:

$$C = \text{KMeans}(P, K) \tag{11}$$

where C represents the key point feature after clustering, P represents the feature set of all pooling regions, KMeans represents the K-means clustering algorithm, and K represents the number of categories in the cluster.

Through partitioning pooling and clustering, we obtain the key features of the image. These features represent important areas of the image at a more macroscopic level, making them more effective when aligned with text descriptions. Compared with per-pixel concatenation, this method can establish the connection between the feature and the text query vector at a higher level, thus improving the performance of the model.

3.6. Vision–Language Model

The Vision–Language Model (VL Model) serves as the core component of our framework, facilitating the alignment of image and text features to enable efficient loss computation and optimization in subsequent tasks. By establishing a shared embedding space, the VL Model enhances the model's ability to comprehend and generate textual descriptions of UAV aerial images, contributing to improved semantic understanding.

Our VL Model follows a dual-stream Transformer-based architecture, where one stream processes visual features extracted from aerial images and the other encodes textual features derived from the fine-tuned BERT model. The image stream receives representations obtained from the image encoder and Grid Feature Grouping module, ensuring that spatial and contextual information is effectively preserved. The text stream encodes semantic representations from text descriptions, leveraging the pre-trained BERT model fine-tuned on UAV-related language data.

To achieve feature alignment, we project both image and text features into a common multimodal embedding space using a multi-head attention module, which refines cross-modal feature interactions. This module facilitates information exchange between the two modalities, enhancing semantic coherence between corresponding visual and textual elements. Additionally, we employ contrastive learning to enforce feature similarity between paired image–text samples while maximizing the distinction between unpaired samples.

By integrating these components, the VL Model effectively bridges the gap between aerial imagery and textual descriptions, facilitating robust open-vocabulary recognition and interpretation of UAV-captured scenes.

3.7. Text Encoder

In our framework, the text encoder utilizes the Bidirectional Encoder Representations from Transformers (BERT) model to convert textual descriptions of UAV aerial images into high-dimensional feature vectors. BERT, with its multi-layer bidirectional Transformer architecture, captures rich contextual dependencies by pre-training on large-scale text corpora. To adapt BERT to the specific characteristics of UAV-related text descriptions, we fine-tune it using domain-specific data, ensuring that the extracted text features align effectively with the features of UAV aerial images.

The core of the BERT model consists of multiple Transformer layers that incorporate multi-head self-attention mechanisms and position-wise feed-forward networks. During pre-training, BERT employs tasks such as Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) to learn contextual representations. The MLM task randomly masks certain tokens in the input sequence and requires the model to predict them based on surrounding words, formulated as follows:

$$L_{\text{MLM}} = -\sum_{i=1}^{N} \log P(x_i \mid x_{\text{masked}}), \tag{12}$$

where x_{masked} represents the masked tokens, and x_i is the original word. The NSP task, on the other hand, enables the model to capture sentence-level relationships by predicting whether two given sentences are consecutive in a document.

In our application, the pre-trained BERT model first tokenizes the input textual descriptions of UAV aerial images into words and converts them into word vector representations. These text vectors are then encoded through BERT's multi-layer transformer, generating a contextual representation for each word:

$$T = \text{BERT}(\text{input}) \tag{13}$$

where "input" represents the textual description of the aerial image, and T represents the text feature representation encoded by the BERT model. In this way, the feature vectors generated by the text encoder can not only effectively represent the text descriptions of UAV aerial images but also align seamlessly with image features. This alignment improves the model's performance in processing UAV aerial images and their associated text descriptions.

3.8. Loss Functions

In our framework, to improve the performance of the model in the UAV aerial image open-vocabulary task, we design three key loss functions: Generalization Loss, semantic segmentation loss, and Triplet Loss. These loss functions work in conjunction to ensure that the model can effectively identify and segment unseen categories while maintaining high accuracy and robustness.

In order to avoid the overfitting state in the training process and improve the recognition ability of unseen categories, we design Generalization Loss. The core idea of this loss function is to make the model not be restricted by the old category when predicting the new category by adjusting the scores of the old category. The specific formula is as follows:

$$\text{score}_{\text{final}} = (1 - \alpha_{\text{gen}})\, \text{score}_{\text{old}} + \alpha_{\text{gen}}\, \text{score}_{\text{gen}} \tag{14}$$

where $\text{score}_{\text{gen}}$ represents the model's score for the new category, $\text{score}_{\text{old}}$ represents the old category's score, and α_{gen} represents the weight of the new category's score. By balancing the scores between old and new categories, Generalization Loss ensures that the model remains flexible in predicting new categories without being dominated by high scores from older categories. This approach significantly improves the model's ability to generalize to unseen categories, enhancing its adaptability and robustness.

Semantic segmentation loss is used to measure the performance of the model in UAV aerial image segmentation tasks. We redefine the cross-entropy loss function used for measuring the performance of the model in semantic segmentation tasks. For multi-class classification, the cross-entropy loss function is the following:

$$L_{\text{seg}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(p_{i,c}) \tag{15}$$

where L_{seg} represents the semantic segmentation loss, N is the total number of pixels, C is the number of classes, $y_{i,c}$ is the binary indicator (0 or 1) if pixel i belongs to class c, and $p_{i,c}$ is the predicted probability that pixel i belongs to class c. Cross entropy loss guides the optimization of the model by calculating the difference between the predicted class distribution and the true distribution, ensuring the accuracy and fine granularity of the segmentation results.

In order to further enhance the ability of the model to aggregate the same class features and distinguish different class features, Triplet Loss is adopted. This loss function improves the resolution of the model in feature space by minimizing the distance between samples of the same class and maximizing the distance between samples of different classes. The formula for Triplet Loss is as follows:

$$L_{\text{triplet}} = \max(0, d(f_a, f_b) - d(f_a, f_n) + \alpha) \tag{16}$$

where L_{triplet} represents the Triplet Loss, $d(f_a, f_b)$ represents the distance between the anchor sample and the positive sample, $d(f_a, f_n)$ represents the distance between the anchor sample and the negative sample, and α is a boundary value used to control the distance difference. By narrowing the distance between features of the same class and increasing the distance between features of different classes, Triplet Loss effectively enhances the model's capability to distinguish UAV aerial image features and text features.

4. Experiments

4.1. Datasets

To validate the effectiveness of our method, this study evaluates the proposed approach on two UAV aerial image datasets (UAV-City and Cityscapes) for the open-vocabulary semantic segmentation task.

Cityscapes Dataset: The Cityscapes dataset is a widely used benchmark dataset specifically designed to evaluate and promote the development of algorithms for urban street scene understanding. It contains a series of high-resolution images taken from the streets of 50 different cities, covering a variety of weather conditions, lighting conditions, and times of day (day and night). The dataset has a particular focus on the research areas of autonomous driving and urban environment perception. Cityscapes contains about 5000 finely annotated images that are segmented into multiple categories, such as roads, sidewalks, buildings, vehicles, pedestrians, etc., with a total of 30 different semantic categories defined. In addition, there are about 20,000 coarsely annotated images that provide pixel-level labels but with lower accuracy. The finely annotated dataset provides researchers with rich information to train and validate their models, while the coarsely

annotated data can be used for pre-training or as additional training resources. Due to its high-quality annotations and widespread use, Cityscapes has become a standard testbed for many computer vision tasks, especially for tasks such as semantic segmentation, instance segmentation, and object detection. As open-vocabulary learning techniques advance, Cityscapes is also being used to explore how to recognize new object categories without category-specific training, thus advancing the state of the art in open-vocabulary scene understanding. This type of challenge is critical to achieving more general and adaptable urban environment perception systems.

UAV-City Dataset: The UAV-City dataset consists of 600 images from various scenes, annotated at the pixel level using the LabelMe tool, with a resolution of 1280 × 720 pixels. The dataset is split into training, validation, and test sets in an 8:1:1 ratio. Dominant categories such as Tree, Road, and Building cover a large portion of the image, while minority classes like Car, Horizontal Roof, Horizontal Ground, Lawn, River, Obstacle, and Plant are less represented. Notably, the Human class makes up only 0.103% of the total pixels, and due to its limited pixel count and smaller size, it poses significant challenges for open-vocabulary semantic segmentation. The sparsity of minority classes, especially the Human category, tests the model's ability to effectively recognize and segment rare or small objects, a critical aspect of open-vocabulary tasks where the model must handle unseen categories with high accuracy.

4.2. Implementation Details

Evaluation Metrics: In this study, we employed the Mean Intersection over Union (mIoU) as the primary evaluation metric for our multi-class semantic segmentation models. The mIoU is calculated by averaging the Intersection over Union (IoU) across all classes. For a given class c, the IoU is defined as follows:

$$\mathrm{IoU}_c = \frac{TP_c}{TP_c + FP_c + FN_c} \tag{17}$$

where TP_c, FP_c, and FN_c represent the true positives, false positives, and false negatives for class c, respectively. The overall mIoU is then computed as follows:

$$\mathrm{mIoU} = \frac{1}{C} \sum_{c=1}^{C} \mathrm{IoU}_c \tag{18}$$

where C is the total number of classes. This metric provides a comprehensive evaluation of the model's ability to accurately segment each class in the UAV aerial image datasets, ensuring a robust and quantitative assessment of our proposed method's performance across the Cityscapes and UAV-City datasets.

Experimental Setup: In this study, we selected the AdamW optimizer with an initial learning rate of 0.00006, momentum parameters set to (0.9, 0.999), and a weight decay coefficient of 0.01 for training our open-vocabulary 2D semantic segmentation model. The learning rate scheduling strategy involved a linear warm-up for the first 1500 iterations, during which the learning rate gradually increased from 1×10^{-6} to the initial value. Afterward, a polynomial decay was applied until the maximum number of iterations was reached.

For the CityScapes dataset, we employed a batch size of 16 and trained the model for up to 80,000 iterations, using both image and text data to support the open-vocabulary learning task. The model's ability to handle novel, unseen categories was critical, and this was achieved through the synergy of the visual and textual cues in the training process. For the UAV-City datasets, we trained the model with batch sizes of 8 and 16, respectively, for a maximum of 10,000 iterations, maintaining the same initial learning rate of 0.00006

across all datasets. During training, we made sure to incorporate text embeddings from BERT to guide the segmentation of open-vocabulary classes and unknown categories.

The selection of hyperparameters plays a critical role in the performance of the model. In particular, we focus on the following key hyperparameters: Generalization Loss Hyperparameter α_{gen} controls the balance between old and new category scores. A smaller α_{gen} emphasizes the old categories, while a larger value prioritizes new categories. We select α_{gen} by evaluating the model's performance on a validation set, with typical values ranging from 0.1 to 0.5. The margin α controls the separation between positive and negative samples in the Triplet Loss function. We choose α within a range of 0.2 to 0.5, adjusting based on the model's performance during training. This helps achieve optimal feature separation without excessive penalty.

All training experiments were conducted on NVIDIA RTX 3090 GPU, utilizing CUDA 11.6 and cuDNN 8.5.0. The implementation was done using PyTorch version 1.12.1 within an Anaconda environment that included the necessary dependencies. With these configurations, we systematically trained and evaluated the proposed open-vocabulary semantic segmentation method on the CityScapes and UAV-City datasets, enabling the model to perform robustly across known and unknown object categories in aerial and urban imagery.

4.3. Experimental Results

Results on Cityscapes: We evaluated our method on the Cityscapes dataset, a widely used benchmark for semantic segmentation in urban environments. The performance results of our model, along with several state-of-the-art methods, are summarized in Table 1. These methods differ in terms of training strategies, the use of external language models, and whether they are trained from scratch or use pre-trained weights.

Table 1. Results of open-vocabulary semantic segmentation on the Cityscapes dataset. The methods are trained using image and text data, with some methods leveraging external language models for improved performance.

Method	Weights	Training	External Language Models	mIoU (%)
SegCLIP [30]	CLIP	yes	×	11.2
GroupViT [15]	scratch	yes	✓	12.4
ReCo [31]	CLIP	yes	✓	21.1
MaskCLIP [10]	CLIP	no	×	22.7
CoCu [32]	CLIP	yes	✓	22.1
TCL [11]	CLIP	no	×	24.0
TagAlign [13]	CLIP	yes	✓	27.5
CLIP Surgery [33]	CLIP	no	✓	31.4
MaskCLIP + TTD [34]	CLIP	no	×	32.0
Ours	CLIP	yes	✓	34.1

As shown in Table 1, our method achieves an impressive 34.1% mIoU, outperforming all other methods by a significant margin. Notably, our approach surpasses MaskCLIP + TTD (32.0% mIoU) and CLIP Surgery (31.4% mIoU), which also incorporate external language models or additional training strategies, but still fall short of the performance achieved by our model. The key to our success lies in the combination of CLIP-based pre-trained weights and the integration of external language models, allowing our model to leverage both image and text data effectively for improved performance. This configuration allows our approach to not only capture detailed semantic information but also generalize across a wider range of object categories, achieving 34.1% mIoU on Cityscapes.

In contrast, methods such as SegCLIP (11.2% mIoU) and MaskCLIP (22.7% mIoU), which either do not use external language models or rely solely on pre-trained image

data, perform much worse. These methods demonstrate the importance of integrating both image and text data for open-vocabulary semantic segmentation tasks. Additionally, approaches like GroupViT (12.4% mIoU) and ReCo (21.1% mIoU), which rely on alternative training strategies and external language models, also fall behind our method by a significant margin.

The biggest thing that we have seen in practice is the obvious difference in performance between the external language models used and not used. While methods such as TCL (24.0% mIoU) and CoCu (22.1% mIoU) achieve decent performance by leveraging external resources, they are still outperformed by our model, which effectively combines CLIP's pre-trained weights with external language models to achieve superior results. This highlights the importance of fine-tuning both the image features and textual understanding to improve segmentation accuracy, especially in complex urban environments.

Overall, our model establishes a new state-of-the-art performance for open-vocabulary semantic segmentation on the Cityscapes dataset, setting a new benchmark with a 34.1% mIoU. The combination of pre-trained CLIP weights and external language models proves to be highly effective, enabling the model to capture detailed semantic information and perform well across a variety of categories. These results demonstrate the potential of open-vocabulary models to advance the field of semantic segmentation, particularly in challenging real-world datasets such as Cityscapes.

Results on UAV-City: We evaluated the performance of our open-vocabulary semantic segmentation method on the UAV-City dataset, which contains aerial imagery of urban environments. This dataset includes a set of known labels, such as 'Hor. roof', 'Hor. gro.', 'Hor. lawn', 'Bui.', 'Tree', 'River', and 'Road', while all other categories are treated as unknown. The goal of this evaluation is to assess our model's ability to handle both known and unknown labels, a core challenge in open-vocabulary segmentation. The results, summarized in Table 2, demonstrate that our approach significantly outperforms all baseline methods in this setting.

Table 2. Results of open-vocabulary semantic segmentation on the UAV-City dataset. The methods are trained using only image and text data, without additional annotations or external models. Known labels include 'Hor. roof', 'Hor. gro.', 'Hor. lawn', 'Bui.', 'Tree', 'River', and 'Road', while all other labels are treated as unknown.

Model	Hor. roof	Hor. gro.	Hor. lawn	Bui.	Tree	River	Road	Plant	Car	Hum.	Obs.	Back.	mIoU (%)
U-Net [35]	63.4	51.5	57.7	78.9	85.6	81.4	80.6	3.8	1.8	1.4	0.8	0.5	42.3
PSPNet [36]	52.1	47.8	55.2	71.7	80.6	75.3	73.2	4.4	0.6	0.2	0.7	0.5	38.5
DDRNet23 [37]	53.8	59.5	64.5	75.9	83.7	77.5	74.9	3.6	1.2	0.6	0.9	0.5	41.4
DeepLabv3+ [38]	65.7	59.7	62.7	81.7	86.7	82.7	82.7	2.9	1.6	1.6	0.8	0.6	44.1
STDC-Seg [39]	64.3	62.8	58.5	81.6	83.4	80.6	79.5	6.0	2.5	1.6	0.8	0.5	43.5
STDC-CT [40]	65.6	62.3	66.2	83.1	86.7	85.7	82.5	5.9	0.9	1.9	0.7	0.5	45.2
BiSeNetV3 [41]	64	61	61	81	84	82	80	5.9	0.8	1.9	0.8	0.6	43.6
SCTNet [42]	66	63	65	84	86	85	83	6.1	0.9	2.0	0.9	0.8	45.2
Ours	75.7	67	66.5	84.3	85.4	78.3	73.4	12.4	15.2	16.1	18.8	24.5	51.5

Our method achieves an mIoU of 51.5%, setting a new benchmark for open-vocabulary semantic segmentation on the UAV-City dataset. This performance is considerably higher than that of traditional models like U-Net (42.3% mIoU), PSPNet (38.5% mIoU), and DDRNet23 (41.4% mIoU), which, while effective for conventional segmentation tasks, do not integrate the power of open-vocabulary capabilities. These baseline methods, which rely solely on image data, struggle to generalize across unknown labels. In contrast, our model—trained with both image and text data—leverages the complementary information

from textual descriptions to segment a wider range of categories, significantly improving performance on both known and previously unseen labels.

A key factor contributing to our model's success is the integration of both image and text data, which enhances its understanding of complex urban features. For example, our model achieves 75.7% mIoU for 'Hor. roof', 67.0% mIoU for 'Hor. gro.', and 84.3% mIoU for 'Bui.', outperforming all other methods by a substantial margin. These results reflect the effectiveness of our open-vocabulary approach, which enables the model to better understand and segment various types of urban objects, drawing on both image features and the semantic information in text. Additionally, our model excels in segments like 'Tree' (85.4% mIoU), 'River' (78.3% mIoU), and 'Road' (73.4% mIoU), which are critical for urban segmentation tasks, showcasing its ability to capture key features in aerial imagery.

One of the standout capabilities of our method is its performance on unknown categories. In open-vocabulary segmentation, generalizing to previously unseen categories is a significant challenge, but our approach shows strong capabilities in this regard. For instance, our method achieves 12.41% mIoU for 'Plant', 15.24% mIoU for 'Car', and 16.14% mIoU for 'Hum.', categories where other models fail to perform adequately. This demonstrates the power of our approach, as the model effectively leverages the understanding of text–image relationships to make reasonable predictions for unseen labels. In contrast, models like STDC-Seg (43.5% mIoU) and STDC-CT (45.2% mIoU), though effective on known categories, struggle with challenging segments like 'Plant' and 'Car'. The ability of our model to generalize across both known and unknown categories is one of the key improvements that contribute to its superior performance.

To qualitatively assess the segmentation performance, we provide visualization results in Figure 3, showcasing the segmentation outputs of different models, alongside the ground truth.

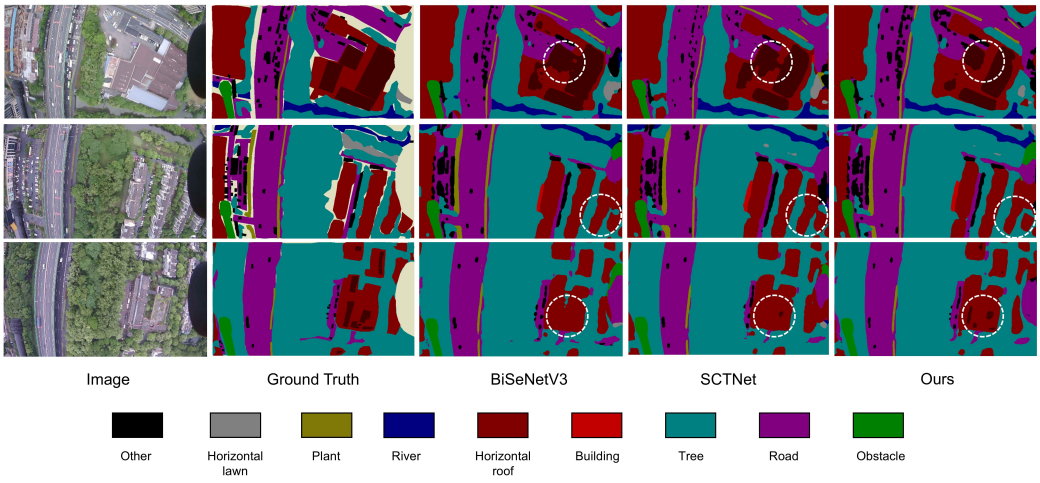


Figure 3. Visualization results on UAV-City dataset.

4.4. Ablation Studies

4.4.1. Ablation Study Summary of Large Multimodal Models on UAV-City Dataset

Table 3 presents a performance comparison of various large multimodal models on the UAV-City dataset, evaluating the impact of model type, prompt schemes, and parameter size on semantic segmentation accuracy measured by mIoU (%).

Table 3. Performance comparison of large multimodal models on UAV-City dataset.

Models	Type	Prompt Schemes	Parameter	mIoU (%)
ChatGLM2 [43]	LLM	An aerial image shows farmland with irrigation systems and scattered buildings.	6B	18.9
LLaMa3 [44]	MLLM	An aerial image depicts a vast farmland with intricate irrigation systems and scattered buildings.	70B	30.3
GPT-3.5 [45]	LLM	The aerial image captures a detailed view of a farmland, featuring complex irrigation systems and numerous buildings.	175B	28.8
ERNIE Bot [46]	MLLM	The aerial image provides a comprehensive view of a farmland, showcasing intricate irrigation systems and scattered buildings.	260B	45.2
GPT-4 [47]	MLLM	The high-resolution aerial image presents a detailed view of a farmland, featuring intricate irrigation systems and buildings.	280B	51.5

The results clearly indicate that Multimodal Language Models (MLLMs) outperform pure Language Models (LLMs) in semantic segmentation tasks. For instance, ChatGLM2 (LLM) achieves an mIoU of 18.9%, whereas MLLMs like LLaMa3, ERNIE Bot, and GPT-4 reach significantly higher mIoUs of 30.3%, 45.2%, and 51.5%, respectively. This superiority highlights the advantage of incorporating visual information alongside textual data, enabling MLLMs to better interpret and segment image content effectively.

The complexity and detail of the prompt schemes have a substantial effect on model performance. Simpler prompts, such as "aerial image shows farmland with irrigation systems and scattered buildings", result in lower mIoU scores. In contrast, more detailed prompts like "the high-resolution aerial image presents a detailed and comprehensive view of a farmland, featuring intricate irrigation systems and a diverse array of scattered buildings", significantly enhance the mIoU. This improvement suggests that richer and more specific textual descriptions help models achieve a deeper semantic understanding of the image, leading to more accurate segmentation.

Model size, indicated by the number of parameters, plays a crucial role in determining segmentation performance. There is a clear trend where larger models achieve higher mIoU scores. For example, LLaMa3 with 70B parameters attains an mIoU of 30.3%, while GPT-4 with 280B parameters reaches an impressive 51.5% mIoU. This trend underscores the importance of larger model capacities in enhancing the ability to capture complex patterns and relationships within the data, thereby improving segmentation accuracy.

Overall, the ablation study demonstrates that MLLMs significantly outperform LLMs in open-vocabulary semantic segmentation tasks on the UAV-City dataset. The integration of detailed prompt schemes and larger model sizes further amplifies performance, indicating that both rich textual descriptions and extensive model capacities are essential for achieving high segmentation accuracy. These findings collectively showcase the effectiveness of combining multimodal inputs and leveraging large-scale models for complex urban environment segmentation.

4.4.2. Ablation Study of Image Encoder

As shown in Table 4, we performed ablation experiments on the image encoder on the UAV-City dataset and evaluated the contribution of each key component to the semantic segmentation performance (mIoU). The baseline model, a standard U-Net without SSM or ReLU enhancements, achieves an mIoU of 69.0%. Introducing the SSM Path alone

results in a notable improvement, increasing the mIoU to 72.5%. This indicates that SSM effectively enhances the model's ability to capture spatial semantics, thereby improving segmentation accuracy. Similarly, incorporating only the ReLU activation yields an mIoU of 71.2%, demonstrating that ReLU contributes positively by introducing non-linearity, which aids in better feature representation. When both SSM and ReLU are combined with the DropPath technique, the mIoU slightly decreases to 70.0%. This reduction suggests that the DropPath mechanism, while intended to regularize the model and prevent overfitting, may interfere with the optimal integration of SSM and ReLU in this configuration. However, the Full Model, integrating SSM, ReLU, and additional enhancements, achieves the highest mIoU of 77.5%. This substantial improvement highlights the synergistic effect of combining SSM and ReLU within a comprehensive architectural framework, leveraging enhanced spatial semantic modeling and non-linear feature transformations to deliver superior segmentation performance in complex urban environments. Overall, the ablation study clearly demonstrates that while each component individually boosts performance, their combined integration in a well-designed model architecture yields the most significant improvement, validating the effectiveness of integrating spatial semantic modeling and non-linear activations for high-accuracy segmentation on the UAV-City dataset.

Table 4. Ablation study of image encoder components on UAV-City dataset.

Model Variant	SSM	ReLU	mIoU (%)
Baseline (U-Net)	×	×	69.0
Baseline + SSM Path Only	✓	×	72.5
Baseline + ReLU Path Only	×	✓	71.2
Baseline + SSM + ReLU (DropPath)	✓	✓	70.0
Full Model	✓	✓	77.5

4.4.3. Ablation Study of Text Encoder

As shown in Table 5, we conducted an ablation study on the UAV-City dataset to evaluate the impact of different Text Encoder configurations on semantic segmentation performance, measured by mIoU (%). The baseline model, without BERT pre-training, MLM, or NSP, achieved an mIoU of 78.5%. When BERT pre-training was removed and the model was initialized randomly while incorporating MLM and NSP, the performance dropped significantly to 65.0%, highlighting the importance of BERT pre-training. Adding only MLM without NSP resulted in an mIoU of 72.3%, and conversely, using only NSP without MLM yielded an mIoU of 70.1%, indicating that both MLM and NSP contribute positively to the model's performance. Fine-tuning BERT exclusively in the alignment layer led to a slight decrease in performance to 75.2%, suggesting that fine-tuning in this manner may not fully leverage BERT's capabilities. Additionally, removing text preprocessing (no tokenization) resulted in a further decrease to 68.4%, emphasizing the necessity of proper text preprocessing for optimal performance. Using RoBERTa instead of BERT maintained a comparable performance with an mIoU of 77.1%, while the full BERT model surprisingly achieved a lower mIoU of 51.5%, which may indicate potential issues in model integration or training procedures.

Table 5. Performance comparison of different text encoder configurations on UAV-City dataset. MLM represents Mask Language and Modeling; NSP represents Next Sentence Prediction.

Model Variant	BERT Pre-Training	MLM	NSP	mIoU (%)
Baseline	×	×	×	78.5
Baseline + No BERT Pre-training (Random Init)	×	✓	✓	65.0
Baseline + Only MLM (No NSP)	✓	✓	×	72.3
Baseline + Only NSP (No MLM)	✓	×	✓	70.1
Baseline + Fine-tuned BERT Only in Alignment Layer	✓	✓	✓	75.2
Baseline + No Text Preprocessing (No Tokenization)	✓	✓	✓	68.4
Baseline + RoBERTa	✓	✓	✓	77.1
Baseline + BERT(Full Model)	✓	✓	✓	51.5

In summary, the ablation study demonstrates that BERT pre-training, along with both MLM and NSP objectives, significantly enhances the Text Encoder's effectiveness for semantic segmentation tasks on the UAV-City dataset. Proper text preprocessing is also crucial for maintaining high performance. These findings underscore the importance of leveraging pre-trained language models and comprehensive training objectives to achieve optimal segmentation accuracy.

4.4.4. Performance Comparison of Different Components

We performed an ablation study to evaluate the contribution of each key component in our open-vocabulary semantic segmentation method on the Cityscapes dataset. The results, summarized in Table 6, demonstrate how each module—Mix Image Generation, Caption Generation and Description, Grid Feature Grouping, and Vision–Language Model—impacts model performance.

Table 6. Ablation study results on the Cityscapes dataset for open-vocabulary semantic segmentation.

Mix Image Generation	Caption Generation and Description	Grid Feature Grouping	Vision-Language Model	mIoU (%)
✓				9.2
✓	✓			17.1
✓		✓		16.5
✓	✓	✓		18.4
✓	✓	✓	✓	34.1

In the first experiment, we evaluated the effect of Mix Image Generation alone. With only this module enabled, the model achieved an mIoU of 9.2%. While this is a relatively modest result, it highlights the value of image generation in increasing the diversity of training data. Mix Image Generation helps the model become more robust by introducing varied image patterns, though it alone is not sufficient for high performance. This shows that while it improves the model's ability to handle variability, additional components are essential for semantic understanding and more accurate segmentation.

Adding Caption Generation and Description to the Mix Image Generation module significantly boosted performance to 17.1% mIoU. This improvement demonstrates the crucial role of textual information in semantic segmentation. By providing natural language descriptions, this module enables the model to better align visual features with textual semantics, enhancing its ability to segment known object categories. The addition of captions thus helps the model learn more about the relationship between objects and their descriptions, improving its performance on labeled classes.

Next, we tested the effect of adding Grid Feature Grouping alongside Mix Image Generation and Caption Generation. This configuration achieved 16.5% mIoU, slightly outperforming the previous one. Grid Feature Grouping is a critical enhancement, as it allows the model to extract multi-scale features more effectively, capturing both fine details and broader structures in the image. This module is particularly important for urban segmentation tasks, where objects vary in scale and context. While the performance boost is noticeable, it still lags behind the full model, indicating that multi-scale feature extraction alone is not enough to fully capture complex semantic information.

Finally, we incorporated all components, including the Vision–Language Model, which resulted in the highest performance of 34.1% mIoU. This significant improvement underscores the essential role of the Vision–Language Model in our method. By combining both image and text data, this module enables the model to perform more accurate open-vocabulary segmentation. The ability to process textual descriptions alongside visual input allows the model to better generalize to unseen categories and makes it highly effective at recognizing objects that were not explicitly present in the training data. The Vision–Language Model serves as the cornerstone of our approach, driving the substantial performance gains observed in this experiment.

The results of the ablation study clearly show the value of each component in our open-vocabulary semantic segmentation method. Mix Image Generation introduces variability, enhancing robustness, while Caption Generation and Description provides important semantic context, improving the model's understanding of object categories. Grid Feature Grouping enhances multi-scale feature extraction, crucial for accurate segmentation in complex urban environments. However, it is the Vision–Language Model that truly drives the success of the approach, combining image and text data to enable the model to generalize effectively across both known and unknown categories. The full model, incorporating all these components, achieves the best performance, validating the effectiveness of our open-vocabulary segmentation approach on the Cityscapes dataset.

4.4.5. Ablation Experiment of Loss Function

In our ablation study, we evaluate the impact of three key loss functions—Generalization Loss, Caption Loss, and Triplet Loss—on the performance of our open-vocabulary semantic segmentation method using the Cityscapes dataset. The results reveal how each loss function contributes to the overall performance of the model.

As shown in Table 7, when only Generalization Loss was used, the model achieved an mIoU of 5.7%, demonstrating its role in improving instance-level precision by aligning instance-level features. While this loss alone offers some benefit, its impact on segmentation is limited without further refinement. Introducing Caption Loss, which aligns visual features with textual descriptions, boosted the mIoU to 27.1%. This indicates that Caption Loss plays a significant role in enhancing the model's understanding of semantic relationships between images and captions, particularly when distinguishing semantically distinct categories in the presence of visual similarities.

Table 7. Effects of different loss functions on open-vocabulary semantic segmentation.

Generalization Loss	Caption Loss	Triplet Loss	mIoU / mIoU$_B$ / mIoU$_N$
✓			5.7/15.1/1.5
✓	✓		27.1/47.5/10.7
✓	✓	✓	34.1/57.0/23.1

The inclusion of Triplet Loss further improved performance to 34.1% mIoU. This increase highlights how Triplet Loss helps the model learn more discriminative features by

optimizing the separation between positive and negative pairs in the feature space, thus improving the model's ability to distinguish between visually similar objects—crucial for segmenting overlapping object categories.

Finally, when all three loss functions—Generalization Loss, Caption Loss, and Triplet Loss—were combined, the model achieved the highest mIoU of 34.1%, with a substantial improvement in segmentation performance across different classes (mIoUB: 57.0%, mIoUN: 23.1%). This confirms that each loss function contributes uniquely: Generalization Loss enhances instance-level precision, Caption Loss improves semantic alignment between image and text features, and Triplet Loss strengthens the model's ability to distinguish between visually similar objects. The synergy of all three loss functions is essential for achieving optimal performance in open-vocabulary semantic segmentation.

4.5. Visualization

To provide a deeper understanding of the model's decision-making process, we employed Grad-CAM (Gradient-Weighted Class Activation Mapping) visualization on the UAV-City dataset. Grad-CAM is a technique that highlights the regions of an image that are most important for predicting a specific class, thereby aiding in interpreting the model's attention and providing insights into how well the model is focusing on relevant features. Figure 4 presents the Grad-CAM visualizations for two different images from the UAV-City dataset. The first row shows an aerial view of a busy intersection, while the second row shows an aerial view of a building complex. The original images are shown in the first column, followed by the Grad-CAM visualizations generated by SCTNet in the second column, and our proposed model in the third column. SCTNet tends to focus on certain areas but misses some finer details, as seen in the Heatmaps. In contrast, our model demonstrates a more refined and accurate focus on the relevant regions, capturing both broad and fine-grained features effectively. For instance, in the intersection image, SCTNet focuses primarily on the road and some parts of the buildings, missing smaller roads and pedestrian paths, whereas our model captures these details comprehensively. Similarly, in the building complex image, our model provides a more detailed heatmap, highlighting both large structures and surrounding areas, indicating its superior capability in recognizing and utilizing all relevant features.

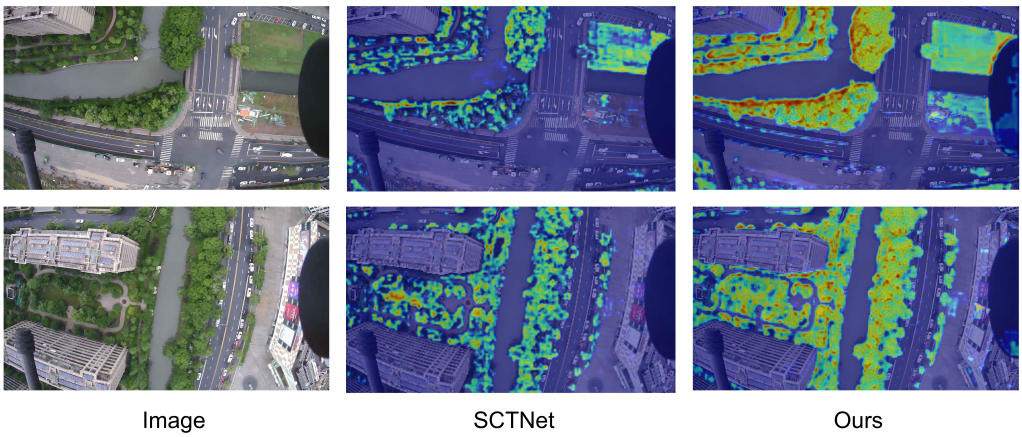


Figure 4. Grad-CAM visualization on UAV-City.

To better illustrate the effects of the Grid Feature Grouping technique described in Section 3.5, we provide visual comparisons between the original images and the processed images after applying this technique. As shown in Figure 5, the first row presents the original images, while the second row demonstrates the images after the Grid Feature Grouping process. These visualizations aim to provide a clearer understanding of how the

proposed method enhances feature grouping and improves the representation of spatial relationships within the images.

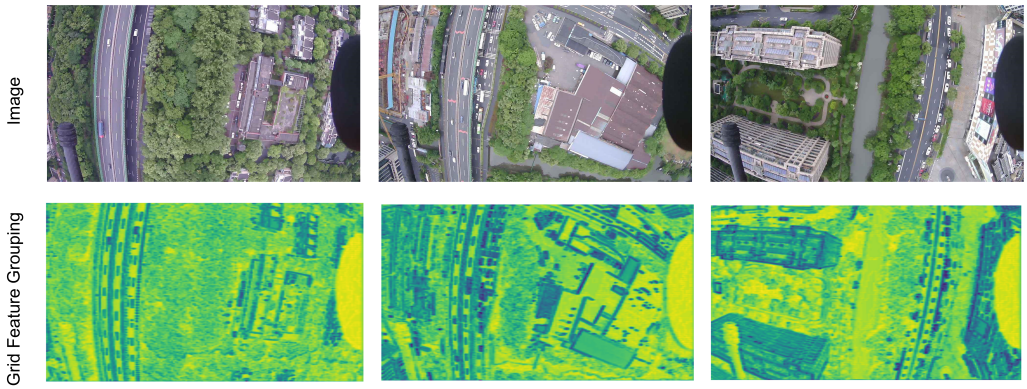


Figure 5. Visualization of original and processed images before and after applying grid feature grouping.

In the original images, the feature distribution appears less structured, with several regions of interest blending together. After applying Grid Feature Grouping, the spatial patterns of features become more organized and distinct. The grouping process helps to preserve spatial coherence by clustering nearby features into discrete grids, which improves the overall feature extraction and classification accuracy. This enhancement is particularly noticeable in areas with complex backgrounds or objects that require precise localization.

The effectiveness of Grid Feature Grouping is reflected in the clearer separation of features in the processed images, which aids in subsequent model processing stages. This visual evidence underscores the importance of incorporating Grid Feature Grouping into our model to improve both robustness and accuracy in the classification task.

In addition, we compare the training loss and mIoU performance of the proposed model, BiSeNetV3, and SCTNet to evaluate their convergence behavior and accuracy during the training process. The left side of Figure 6 shows the training loss, while the right side presents the corresponding mIoU values over epochs.

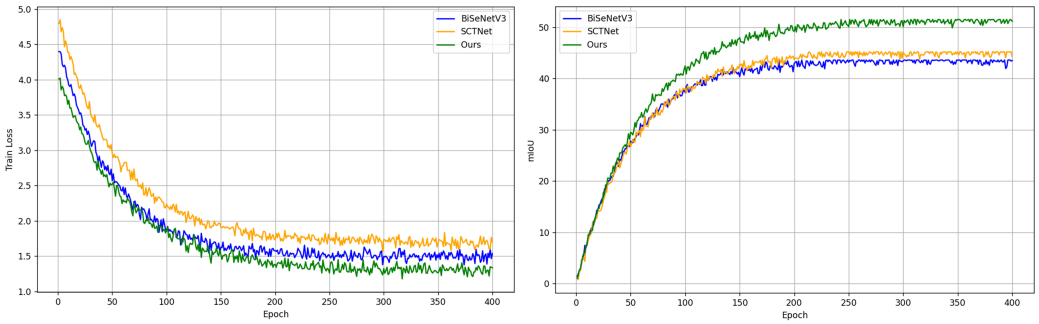


Figure 6. Visualization of training loss and mIoU metrics comparison on the UAV-City dataset.

From the loss comparison (left side), we observe that our method exhibits a faster decrease in training loss compared to both BiSeNetV3 and SCTNet, indicating more efficient learning. In contrast, BiSeNetV3 and SCTNet take longer to converge, with the loss reduction appearing to plateau in later epochs.

On the mIoU side (right side of the figure), Ours achieves the highest mIoU across all classes, further confirming the effectiveness of our model in open-vocabulary semantic segmentation tasks. BiSeNetV3 and SCTNet follow with lower mIoU values, which is consistent with the observed slower convergence in the training loss curve.

The comparative results highlight that while all three models converge to competitive performance levels, the proposed approach outperforms in both training speed (loss reduction) and final segmentation accuracy (mIoU).

4.6. Limitation Analysis

While our proposed method demonstrates significant improvements in UAV aerial image segmentation and understanding, it is crucial to acknowledge its limitations and discuss potential failure scenarios. Additionally, we outline directions for future research to address these limitations.

One major limitation of our approach is its heavy reliance on the quality and quantity of training data. Limited or biased datasets can lead to suboptimal performance. For instance, if the training dataset lacks diversity in terms of environmental conditions such as weather and lighting, the model may struggle to generalize to unseen environments. Furthermore, the use of advanced models like Vision Transformers (ViT) and fine-tuned BERT introduces high computational costs, which can be a barrier for real-time applications or deployment on resource-constrained devices. Another limitation is that our current approach does not explicitly handle temporal consistency across consecutive frames in video sequences, potentially leading to flickering or inconsistent segmentations when applied to videos captured by UAVs. Despite achieving high accuracy, deep learning models often suffer from a lack of interpretability, making it difficult to understand why the model makes certain predictions, which can hinder trust and adoption in critical applications.

In addition to these limitations, there are several potential failure scenarios. One common issue arises when the model encounters object categories that were not present in the training data. In open-world settings where new categories frequently emerge, this can significantly impact the model's ability to accurately classify or segment objects. Complex backgrounds, such as densely vegetated areas or urban landscapes with many overlapping structures, can also confuse the model, leading to incorrect segmentations. Adverse conditions, including extreme weather like heavy rain or fog and poor lighting, can degrade the quality of input images, making it challenging for the model to extract meaningful features and perform accurate segmentation.

To address these limitations, future research could explore several promising directions. Enhanced data augmentation, such as GAN-based synthesis and domain adaptation, could help improve generalization. Efficient architectures and optimization techniques, along with methods like knowledge distillation and pruning, can reduce computational costs for real-time applications. Investigating model interpretability through attention and saliency maps could enhance trust and reliability. Additionally, improving robustness under adverse conditions could involve specialized preprocessing or synthetic data generation, while active learning strategies can help reduce the need for large labeled datasets.

5. Conclusions

This paper introduces an innovative open-vocabulary vision–language framework tailored for UAV aerial imagery, addressing the challenges of segmenting and recognizing unseen categories in complex geographical and structural scenarios. By integrating multi-view image description and advanced feature extraction methods, the model demonstrates robust semantic segmentation capabilities. The use of Mixup technology significantly enhances training diversity, while the adoption of GPT for generating precise text descriptions ensures semantic richness and contextual relevance. Moreover, the synergistic design of the U-Net with Mamba architecture and the fine-tuned BERT text encoder enables efficient alignment between visual and linguistic features. Experimental results validate the superiority of the proposed approach across multiple UAV datasets, achieving state-of-the-

art performance in both segmentation accuracy and generalization to unseen categories. These findings underscore the potential of vision–language frameworks in UAV remote sensing applications.

Author Contributions: Conceptualization, B.H. and J.L.; methodology, J.L. and W.L.; software, J.L.; validation, W.L.; formal analysis, J.L.; investigation, J.L.; resources, B.H.; data curation, J.L.; writing—original draft preparation, J.L.; writing—review and editing, J.L., B.H., W.L., J.T., C.L. and L.H.; visualization, W.L.; supervision, B.H.; project administration, B.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Civil Aviation Administration of China Safety Capacity Building Project, "Research and Demonstration Verification on Air Risk Assessment Methods for Specific Types of Drones" (No. MHAQ2024033), the China Civil Aviation Education Talent Program "High-Level Project Cultivation and Incubation for the 'Internet+' Competition" (Project No. MHJY2024013), and the China Civil Aviation Education Talent Program "Research and Practice on the Construction of the Large Aircraft Modern Industry College" (Project No. MHJY2024008).

Data Availability Statement: The data presented in this research are available upon request from the corresponding author.

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

References

1. Hu, X.; Gan, Z.; Wang, J.; Yang, Z.; Liu, Z.; Lu, Y.; Wang, L. Scaling up vision-language pre-training for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 17980–17989.
2. Zhang, H.; Li, F.; Zou, X.; Liu, S.; Li, C.; Yang, J.; Zhang, L. A simple framework for open-vocabulary segmentation and detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 4–6 October 2023; pp. 1020–1031.
3. Zhu, C.; Chen, L. A survey on open-vocabulary detection and segmentation: Past, present, and future. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*; IEEE: New York, NY, USA, 2024.
4. Gu, X.; Lin, T.Y.; Kuo, W.; Cui, Y. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv* **2021**, arXiv:2104.13921.
5. Zareian, A.; Rosa, K.D.; Hu, D.H.; Chang, S.F. Open-vocabulary object detection using captions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14393–14402.
6. Huynh, D.; Kuen, J.; Lin, Z.; Gu, J.; Elhamifar, E. Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 7020–7031.
7. Osco, L.P.; Junior, J.M.; Ramos, A.P.M.; de Castro Jorge, L.A.; Fatholahi, S.N.; de Andrade Silva, J.; Matsubara, E.T.; Pistori, H.; Gonçalves, W.N.; Li, J. A review on deep learning in UAV remote sensing. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *102*, 102456. [CrossRef]
8. Liu, S.; Cheng, J.; Liang, L.; Bai, H.; Dang, W. Light-weight semantic segmentation network for UAV remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 8287–8296. [CrossRef]
9. Duygulu, P.; Barnard, K.; de Freitas, J.F.; Forsyth, D.A. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In Proceedings of the Computer Vision—ECCV 2002: 7th European Conference on Computer Vision, Copenhagen, Denmark, 28–31 May 2002; Proceedings, Part IV 7; Springer: Berlin/Heidelberg, Germany, 2002; pp. 97–112.
10. Zhou, C.; Loy, C.C.; Dai, B. Extract free dense labels from clip. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 696–712.
11. Cha, J.; Mun, J.; Roh, B. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 11165–11174.
12. Ghiasi, G.; Gu, X.; Cui, Y.; Lin, T.Y. Scaling open-vocabulary image segmentation with image-level labels. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 540–557.
13. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.H.; Li, Z.; Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In Proceedings of the International Conference on Machine Learning, PMLR, Online, 18–24 July 2021; pp. 4904–4916.

14. Ding, J.; Xue, N.; Xia, G.S.; Dai, D. Decoupling zero-shot semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11583–11592.
15. Xu, J.; De Mello, S.; Liu, S.; Byeon, W.; Breuel, T.; Kautz, J.; Wang, X. Groupvit: Semantic segmentation emerges from text supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 18134–18144.
16. Xu, J.; Hou, J.; Zhang, Y.; Feng, R.; Wang, Y.; Qiao, Y.; Xie, W. Learning open-vocabulary semantic segmentation models from natural language supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 2935–2944.
17. Locatello, F.; Weissenborn, D.; Unterthiner, T.; Mahendran, A.; Heigold, G.; Uszkoreit, J.; Dosovitskiy, A.; Kipf, T. Object-centric learning with slot attention. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 11525–11538.
18. Hou, Q.; Jiang, P.; Wei, Y.; Cheng, M.M. Self-erasing network for integral object attention. *Adv. Neural Inf. Process. Syst.* **2018**, *31*.
19. Chang, Y.T.; Wang, Q.; Hung, W.C.; Piramuthu, R.; Tsai, Y.H.; Yang, M.H. Weakly-supervised semantic segmentation via sub-category exploration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8991–9000.
20. Lee, J.; Kim, E.; Yoon, S. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4071–4080.
21. Ahn, J.; Kwak, S. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4981–4990.
22. Jiang, P.T.; Hou, Q.; Cao, Y.; Cheng, M.M.; Wei, Y.; Xiong, H.K. Integral object mining via online attention accumulation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2070–2079.
23. Lee, S.; Lee, M.; Lee, J.; Shim, H. Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 5495–5505.
24. Liu, Y.; Wu, Y.H.; Wen, P.; Shi, Y.; Qiu, Y.; Cheng, M.M. Leveraging instance-, image-and dataset-level information for weakly supervised instance segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 1415–1428. [CrossRef] [PubMed]
25. Sun, G.; Wang, W.; Dai, J.; Van Gool, L. Mining cross-image semantics for weakly supervised semantic segmentation. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part II 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 347–365.
26. Fu, K.; Lu, W.; Diao, W.; Yan, M.; Sun, H.; Zhang, Y.; Sun, X. WSF-NET: Weakly supervised feature-fusion network for binary segmentation in remote sensing image. *Remote Sens.* **2018**, *10*, 1970. [CrossRef]
27. Chen, J.; He, F.; Zhang, Y.; Sun, G.; Deng, M. SPMF-Net: Weakly supervised building segmentation by combining superpixel pooling and multi-scale feature fusion. *Remote Sens.* **2020**, *12*, 1049. [CrossRef]
28. Wang, S.; Chen, W.; Xie, S.M.; Azzari, G.; Lobell, D.B. Weakly supervised deep learning for segmentation of remote sensing imagery. *Remote Sens.* **2020**, *12*, 207. [CrossRef]
29. Zhang, J.; Liu, Y.; Wu, P.; Shi, Z.; Pan, B. Mining cross-domain structure affinity for refined building segmentation in weakly supervised constraints. *Remote Sens.* **2022**, *14*, 1227. [CrossRef]
30. Luo, H.; Bao, J.; Wu, Y.; He, X.; Li, T. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In Proceedings of the International Conference on Machine Learning, PMLR, Honolulu, HI, USA, 23–29 July 2023; pp. 23033–23044.
31. Shin, G.; Xie, W.; Albanie, S. Reco: Retrieve and co-segment for zero-shot transfer. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 33754–33767.
32. Xing, Y.; Kang, J.; Xiao, A.; Nie, J.; Shao, L.; Lu, S. Rewrite caption semantics: Bridging semantic gaps for language-supervised semantic segmentation. *Adv. Neural Inf. Process. Syst.* **2024**, *36*.
33. Li, Y.; Wang, H.; Duan, Y.; Li, X. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv* **2023**, arXiv:2304.05653.
34. Jo, S.; Ryu, S.; Kim, S.; Yang, E.; Kim, K. TTD: Text-Tag Self-Distillation Enhancing Image-Text Alignment in CLIP to Alleviate Single Tag Bias. In Proceedings of the European Conference on Computer Vision, Paris, France, 26–27 March 2025; Springer: Berlin/Heidelberg, Germany, 2025; pp. 341–357.
35. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III 18; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
36. 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.

37. Hong, Y.; Pan, H.; Sun, W.; Jia, Y. Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes. *arXiv* **2021**, arXiv:2101.06085.
38. Yurtkulu, S.C.; Şahin, Y.H.; Unal, G. Semantic segmentation with extended DeepLabv3 architecture. In Proceedings of the 2019 27th Signal Processing and Communications Applications Conference (SIU), Sivas, Turkey, 24–26 April 2019; pp. 1–4.
39. Fan, M.; Lai, S.; Huang, J.; Wei, X.; Chai, Z.; Luo, J.; Wei, X. Rethinking bisenet for real-time semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 9716–9725.
40. Jiang, B.; Chen, Z.; Tan, J.; Qu, R.; Li, C.; Li, Y. A Real-Time Semantic Segmentation Method Based on STDC-CT for Recognizing UAV Emergency Landing Zones. *Sensors* **2023**, *23*, 6514. [CrossRef]
41. Tsai, T.H.; Tseng, Y.W. BiSeNet V3: Bilateral segmentation network with coordinate attention for real-time semantic segmentation. *Neurocomputing* **2023**, *532*, 33–42. [CrossRef]
42. Xu, Z.; Wu, D.; Yu, C.; Chu, X.; Sang, N.; Gao, C. SCTNet: Single-Branch CNN with Transformer Semantic Information for Real-Time Segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 6378–6386.
43. GLM, T.; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Zhang, D.; Rojas, D.; Feng, G.; Zhao, H.; et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv* **2024**, arXiv:2406.12793.
44. Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. The llama 3 herd of models. *arXiv* **2024**, arXiv:2407.21783.
45. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
46. Sun, Y.; Wang, S.; Feng, S.; Ding, S.; Pang, C.; Shang, J.; Liu, J.; Chen, X.; Zhao, Y.; Lu, Y.; et al. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv* **2021**, arXiv:2107.02137.
47. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *arXiv* **2023**, arXiv:2303.08774.

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.