

# RemoteCLIP: A Vision Language Foundation Model for Remote Sensing

Fan Liu, *Member, IEEE*, Delong Chen, Zhangqinyun Guan, Xiaocong Zhou, Jiale Zhu,  
Jun Zhou *Senior Member, IEEE*,

**Abstract**—General-purpose foundation models have become increasingly important in the field of artificial intelligence. While self-supervised learning (SSL) and Masked Image Modeling (MIM) have led to promising results in building such foundation models for remote sensing, these models primarily learn low-level features, require annotated data for fine-tuning, and not applicable for retrieval and zero-shot applications due to the lack of language understanding. In response to these limitations, we propose RemoteCLIP, the first vision-language foundation model for remote sensing that aims to learn robust visual features with rich semantics, as well as aligned text embeddings for seamless downstream application. To address the scarcity of pre-training data, we leverage data scaling, converting heterogeneous annotations based on Box-to-Caption (B2C) and Mask-to-Box (M2B) conversion, and further incorporating UAV imagery, resulting a  $12 \times$  larger pretraining dataset. RemoteCLIP can be applied to a variety of downstream tasks, including zero-shot image classification, linear probing,  $k$ -NN classification, few-shot classification, image-text retrieval, and object counting. Evaluations on 16 datasets, including a newly introduced RemoteCount benchmark to test the object counting ability, show that RemoteCLIP consistently outperforms baseline foundation models across different model scales. Impressively, RemoteCLIP outperform previous SoTA by 9.14% mean recall on RSICD dataset and by 8.92% on RSICD dataset. For zero-shot classification, our RemoteCLIP outperform CLIP baseline by up to 6.39% average accuracy on 12 downstream datasets.

**Index Terms**—Remote Sensing, Foundation Model, CLIP, Vision-language, Multi-modality

## I. INTRODUCTION

Foundation models [1] are becoming increasingly important in AI. Compared to small, specialized models tailored for specific tasks or domains, “one-for-all”-style general-purpose foundation models typically exhibit superior capabilities and generalization abilities in a wide range of downstream tasks. Numerous foundation models have emerged in recent years, such as SimCLR [2], MAE [3], Florence [4] SAM [5] for computer vision, the BERT model [6] and GPT [7], [8] series for natural language processing, and CLIP [9] and Flamingo [10] for vision-language learning, etc. Meanwhile, the remote sensing community is also progressing towards developing foundation models for satellite imagery analysis. To date, the prevailing approaches of building such foundation

Fan Liu and Delong Chen contributed equally. Corresponding author: Fan Liu.

Fan Liu, Delong Chen, Zhangqinyun Guan, Xiaocong Zhou, and Jiale Zhu are with the College of Computer and Information, Hohai University, Nanjing, 210098, China.

Jun Zhou is with the School of Information and Communication Technology, Griffith University, Nathan, Queensland 4111, Australia.

models are primarily inspired by the success of self-supervised learning (SSL) in computer vision, particularly the Masked Image Modeling (MIM) [3], [11], [12] method. Several recent works, including SatMAE [13], Scale-MAE [14], ViTAE [15], Billion-scale MAE [16], RingMo [17], GFM [18], have employed MIM on large Vision Transformers (ViT) and large-scale satellite imagery datasets, yielding encouraging results.

Nevertheless, recent studies have reveal that the MIM method, the basis of most current remote sensing foundation models, primarily learns low-level visual features instead of high-level semantics. For example, Kong et al. [19] proved that MIM pretraining is equivalent to learning occlusion-invariant visual features. Park et al. [20] showed that MIM methods prefer to learn high-frequency texture features instead of capturing longer-range global patterns. Although such low-level features are proved to be beneficial to visual recognition tasks in general domains (*e.g.*, natural images in ImageNet) [21], it is not clear whether they are optimal for the satellite imagery domain. Furthermore, all of these foundation models require annotated data and an additional fine-tuning stage, so that they can be adapted into downstream tasks (*e.g.*, scenes classification). They are unable to perform *zero-shot* inference like the CLIP model [9] due to the lack of joint modelling of vision and language. As noted in recent studies [22], [23], multi-modality should play a crucial role in building foundation models for geospatial artificial intelligence (GeoAI). A vision-language foundation model for remote sensing could pave the way for numerous CLIP-based vision-language applications in remote sensing scenarios, such as open-vocabulary object detection, zero-shot image segmentation, text-to-image generation and editing, multimodal large language models, etc.

Therefore, we step towards developing a *vision-language* foundation model for remote sensing. Our goal is to learn robust visual features with rich semantics of satellite imagery visual concepts while simultaneously learning text embeddings that aligned well with the visual features, which enables the learned aligned vision-language representations to be seamlessly applied into different downstream tasks and domains. To achieve this objective, the primary challenge that we faced is the scarcity of pre-training data. Although some recent works introduced high-quality human-annotated satellite imagery captioning datasets [24]–[26], their scale still remains much insufficient – all existing datasets contain fewer than 10k samples, we found that training a large vision language foundation model on such dataset results in a severe overfitting phenomenon.

To tackle this issue, in this paper, we perform data scaling

based on a ensemble of a wide range of remote sensing datasets, expanding the pre-training data to  $12\times$  than the combination of all existing available data [24]–[26]. We convert heterogeneous annotations, including object detection bounding boxes and semantic segmentation maps, into a unified image-caption data format based on proposed mask-to-box (M2C) and box-to-caption (B2C) generation strategies. We also incorporate Unmanned aerial vehicle (UAV) imaginary to further enhance the diversity of the pre-training data. We train the model to optimize the InfoNCE loss, a lower bound of mutual information between paired image and text samples, to align the vision language representations. After pre-training, we applied the resulting foundation model, which we named as RemoteCLIP, to a diverse set of downstream applications, including zero-shot image classification, linear probing,  $k$ -NN classification, few-shot classification, image-text retrieval. We also develop a novel benchmark, “RemoteCount”, based on automatically-created counterfactual examples to test the object counting ability. Our comprehensive evaluation on a total of 16 datasets demonstrates that our RemoteCLIP yields superior performance compared to various baseline foundation models, and the advantages of RemoteCLIP over baselines are consistent across different model scales, from ResNet-50 with 38 million parameters to ViT-Large-14 with 304 million parameters.

Here we briefly summarize our empirical results. We first evaluate our RemoteCLIP on cross-modal retrieval (image-to-text retrieval and text-to-image retrieval), and established new state-of-the-art (SoTA) on all of the three remote sensing retrieval benchmarks. Specifically, RemoteCLIP outperform previous SoTA [27] by 9.14% mean recall on RSITMD [24] dataset, and by 8.92% mean recall on RSICD [25] dataset. For zero-shot classification, our RemoteCLIP outperform CLIP baseline by up to 6.39% average accuracy on 12 downstream datasets. For object counting evaluation on our newly introduced RemoteCount dataset, RemoteCLIP outperform CLIP by 21.02% top-1 accuracy.

The contributions of this paper are summarized as follows:

- **A large-scale dataset for the remote sensing domain:** This paper introduces a comprehensive dataset that combines a wide range of remote sensing datasets. This dataset is 12 times larger than the combination of all existing available data, addressing the scarcity of pre-training data in remote sensing.
- **The first vision-language foundation model for remote sensing:** We propose a novel vision-language foundation model called RemoteCLIP. With the large-scale pretraining dataset, this model is trained to align vision-language representations and learns robust visual features with rich semantics of satellite imagery visual concepts.
- **Diverse downstream applications for remote sensing:** The effectiveness of RemoteCLIP is evaluated on various downstream tasks, including cross-modal retrieval, zero-/few-/full-shot satellite imagery classification, and object counting. We also introduce a new benchmark, called RemoteCount, for object counting evaluation in the remote sensing domain.

The remains of this paper are structured as follows. In Section II, we review related literature on vision language models and existing foundation models in remote sensing. Section III introduce our methodology of building RemoteCLIP – we first prove that the vision-language representation of large CLIP models are very powerful for remote sensing task, but the data for continual pretraining is a major bottleneck for further improve CLIP’s performance (Section III-A). Then, in Section III-B, we describe the details of how we perform data scaling to address this issue, and we present a comprehensive analysis of our new dataset in Section III-C. Section IV presents our empirical evaluation of RemoteCLIP, and Section V presents our discussion of the advantages and limitations of RemoteCLIP, and conclude this paper.

## II. RELATED WORK

### A. Vision-Language Models

1) *CLIP Models*: The joint understanding of image and text content has long been a challenging task and a research hotspot in the field of AI [28]–[32]. In the seminal work of CLIP [9], [33], a two-tower model was trained to contrastively align the representations of massive image-text pairs sourced from the Internet. The visual tower of the CLIP model could learn rich semantic representations, while the text tower could provide robust label embeddings for zero-shot inference. Beyond classification, CLIP models also exhibit the flexibility to be integrated into a variety of vision-language applications, including detection [34], [35], segmentation [36], [37], captioning [38], [39], VQA [40], [41], and conditional image generation [42]. Recent works on CLIP models have primarily focused on scaling model size and data size [43], incorporating self-supervisions [44]–[46], enhancing pre-training efficiency [47], [48], few-shot adaptation [49], and adapting CLIP to video [50], etc.

2) *Domain Specialized CLIP Models*: Meanwhile, a notable research direction is to investigate the domain specialization of CLIP models. Although CLIP’s overall performance across various domains is impressive, developing domain-specialized CLIP models offers more advantages for certain applications due to their significantly improved in-domain performance. For instance, in the medical domain, ConVIRT [51] PubMedCLIP [52], MedCLIP [53], and the recent BioMedCLIP [54] have yielded promising results. As another example, based on large-scale E-commerce image-text datasets, specialized CLIP models outperform naive CLIP baseline by a significant margin [55]–[57].

3) *Sigle-tower Vision Language Models*: Notably, alongside CLIP-style two-tower models, another line of vision-language models has adopted single-tower architectures and combined multiple losses (e.g., binary image-text matching, masked language modeling) for better multimodal fusion [58]. However, these multimodal transformers have recently been surpassed by models from an alternative technical route – combining CLIP models with Large Language Models (LLMs), such as Flamingo [10], BLIP-2 [59], KOSMOS-1 [60], GPT-4 [8], LLaVA [61], etc. Therefore, to build a vision-language foundation model for remote sensing, we opt to construct a

CLIP model rather than pretraining single-tower transformers, which enable us to connect our model to powerful large language models to enable text generation capabilities in the future.

### B. Foundation Models for Remote Sensing

Foundation visual models for natural images (*e.g.*, Florence [4]) attracts much attention in recent years, as they learn powerful semantic representations during large-scale pretraining, and yield impressive performance across a variety of downstream tasks. The remote sensing community is also attempting to develop such foundation models, and currently the main methodology follows the self-supervised learning (SSL) approaches in general computer vision, such as augmentation-based models (*e.g.*, SimCLR [2]) and Masked Image Modelling (MIM)-based models (*e.g.*, MAE [3]).

1) *Augmentation-based SSL Models*: Augmentation-based SSL models produce two augmentations (views) of the same image input, and align their representations by optimizing a contrastive loss. Swope et al. [62], Jung et al. [63], and Jean et al. [64], follows such methodology and perform SSL pretraining on remote sensing imagery. Improvements upon these methods includes introducing momentum contrast [65], rotation (0°, 90°, 180°, 270°) invariance [66], knowledge distillation [67], etc.

2) *MIM-based SSL Models*: MIM methods mask a proportion of the patches input, and train the vision transformer to reconstruct these masked patches. Some remote sensing models based such MIM method mainly focus on adding new properties to the model, including scale-invariance (Scale-MAE [14]), temporal information (SatMAE [13]), temporal invariance (SeCo [68]), etc. Another line of works attempted to scale up the MIM model, such as RingMo [17], billion-scale MAE [16], VITAE [15], etc.

3) *Vision Language Models*: As detailed in Section I, SSL-based remote sensing foundation models mainly learn low-level features and lack an understanding of high-level semantics. Also, the lack of language understanding results in that the model can not carry out the zero-shot applications, existing models require an additional fine-tuning stage for downstream adaptation. To address such limitation, image-text retrieval models, which learn to align image and text representations, may have great potential. Early works of remote sensing retrieval includes Abdullah et al. [69] and Rahhal et al. [70] which used CNN for encode the image and used LSTM to encode the text caption. To equip the model with the capability of global and local understanding of satellite images, Yuan et al. [71] proposed a novel framework using a dynamic fusion module. Rahhal et al. [27] proposed a multi-language framework composed of a language encoder to adapt to the remote sensing semantics of different languages. However, the capability of these vision language models on downstream applications beyond retrieval have not been validated.

## III. REMOTECLIP

### A. Contrastive Language Image Pretraining

Vision language models trained with the Contrastive Language Image Pretraining (CLIP) [9] strategy have demon-

strated impressive generalization ability in various vision-language learning tasks. These models, usually referred to as CLIP models, learn to group and align the representations of semantically similar samples together under cross-modal supervision mined from billion-scale image-text pairs. The CLIP model optimizes a simple InfoNCE loss function, which encourages the alignment of paired image-text samples and pushes apart mismatched samples.

Formally, CLIP is trained with a large-scale image-text dataset  $\mathcal{D} = \{(x_i^I, x_i^T)\}_{i=1}^M$  that consists of a total of  $M$  training samples. The goal is to learn an image encoder  $f^I$  and a text encoder  $f^T$  that respectively encode image sample  $x_i^I$  and text sample  $x_i^T$  to their latent representations, *i.e.*,  $f^I(x_i^I) = z_i^I \in \mathbb{R}^{d_z \times 1}$  and  $f^T(x_i^T) = z_i^T \in \mathbb{R}^{d_z \times 1}$ . During pretraining, CLIP creates an instance discrimination task within each batch, and optimizes the following bi-directional InfoNCE objective, where  $N$  is the batch size and  $\tau$  is a learnable temperature parameter:

$$\begin{aligned} \mathcal{L}_{\text{InfoNCE}} = & - \left( \frac{1}{N} \sum_{i=1}^N \log \underbrace{\frac{\exp(z_i^I \cdot z_i^T / \tau_{\text{CLIP}})}{\sum_{j=1}^N \exp(z_i^I \cdot z_j^T / \tau_{\text{CLIP}})}}_{\text{image to text}} \right. \\ & + \left. \frac{1}{N} \sum_{i=1}^N \log \underbrace{\frac{\exp(z_i^T \cdot z_i^I / \tau_{\text{CLIP}})}{\sum_{j=1}^N \exp(z_i^T \cdot z_j^I / \tau_{\text{CLIP}})}}_{\text{text to image}} \right) / 2, \end{aligned} \quad (1)$$

According to Chen et al. [48], optimizing  $\mathcal{L}_{\text{InfoNCE}}$  brings the following two important properties to the CLIP model:

- Representation alignment: it produces high similarity  $z_i^I \cdot z_i^T$  of paired image and text samples  $x_i^I, x_i^T$ , and low similarity  $z_i^I \cdot z_j^T (i \neq j)$  between the unpaired samples  $x_i^I, x_j^T$ . Generally, perfect representation alignment yields strong downstream performance on cross-modal retrieval tasks.
- Representation grouping: it means that (uni-modal) representations of semantically similar samples are grouped together, while those of dissimilar samples should be pulled apart. Perfect representation grouping yields strong uni-modal recognition (*e.g.*, linear classification) performance.

While fulfilling perfect representation alignment and representation grouping simultaneously, coupled with a large dataset containing sufficient open-set concepts, the model can achieve strong zero-shot classification performance.

1) *Large CLIP is also a strong model for remote sensing tasks*: Although CLIP models do not have any special designs to optimize their performance in the remote sensing domain, they have shown strong zero-shot performance on several remote sensing benchmarks. In the original CLIP paper, OpenAI researchers evaluated CLIP's scene recognition performance on EuroSAT and RESISC45. The zero-shot performance of OpenAI's largest CLIP (ViT-Large-14-336) is only 59.6% and 71.7% respectively. The recent study on Satellite ImageNet (SATIN) dataset [72] also confirmed that the zero-shot performance of the CLIP family is unsatisfactory. However, the linear probing accuracy of CLIP reaches 98.1% and 94.9% on

EuroSAT and RESISC45 in OpenAI’s evaluation, outperforming all of the other 11 foundation visual models compared, including both fully-supervised ones and self-supervised ones. It shows that large-scale contrastive image text pretraining produces high-quality visual representations that are suitable for the remote sensing domain, but at the same time, the cross-modal alignment property of such representations is unsatisfactory.

To have a more thorough understanding of the potential of CLIP models for remote sensing vision language tasks, we performed a comprehensive evaluation of CLIP’s zero-shot retrieval on three commonly used remote sensing retrieval datasets: RSITMD, RSICD, and UCM (which we denote as RET-3). We use the pretrained weights provided by both OpenAI and OpenCLIP, covering models from ResNet-50 (38M parameters) to ViT-G-14 (1.8B parameters). We also compared representative single-tower vision language models including ALBEF and BLIP.

We report the evaluation results in Fig. 1, and we find that model size is an important factor. The larger model consistently yields better performance, the largest CLIP model ViT-G-14 even surpassed all the previous retrieval methods that are specially designed for the remote sensing domain, except only one (Rahhal *et al.* [27]) which is based on fine-tuning of CLIP model. In addition, large CLIP models outperform single-tower models (ALBEF and BLIP) by a large margin, demonstrating the power of simplicity - combining large-scale model and large-scale pretraining data clearly outperforms complex network structures or employing multiple loss functions.

2) *Continual pretraining on small CLIP further improves the performance:* Given the strong results of large CLIP models on remote sensing tasks, a natural question is whether we can further improve their performance using in-domain aerial imagery data. Continual pretraining is a popular methodology to achieve this goal, which has already shown its advantages for adapting CLIP models into the medical domain [54]. As an initial experiment, we performed continuous pretraining of the CLIP model (ResNet-50 and ViT-Base-32) on the union of three existing retrieval datasets RSITMD, RSICD, and UCM (RET-3)<sup>1</sup>. The resulting models, which we denote as CLIP-CP, yield extremely powerful performance. As shown in Fig. 1, it not only outperforms the zero-shot results of the largest CLIP model (ViT-G-14) with 2% (38M vs. 1.8B) parameters, but also established a new state-of-the-art on these three retrieval benchmarks.

It is also clear that tuning the foundation model on a collection of datasets is beneficial. Compared to the model of Rahhal *et al.* [27] using the same ViT-Base-32 architecture, our approach – continual pretraining on the RET-3 collection – improved the performance by a clear margin (4.6%). This multi-dataset tuning shares a similar spirit with recent studies

<sup>1</sup>Similar works have been done in the previous SoTA on remote sensing image-text retrieval method [27], where the authors fine-tuned CLIP models (ViT-Base-32) separately on RSITMD, RSICD, and UCM. However, our goal is different – we aim to build foundation vision language models based on the powerful pretrained CLIP models.

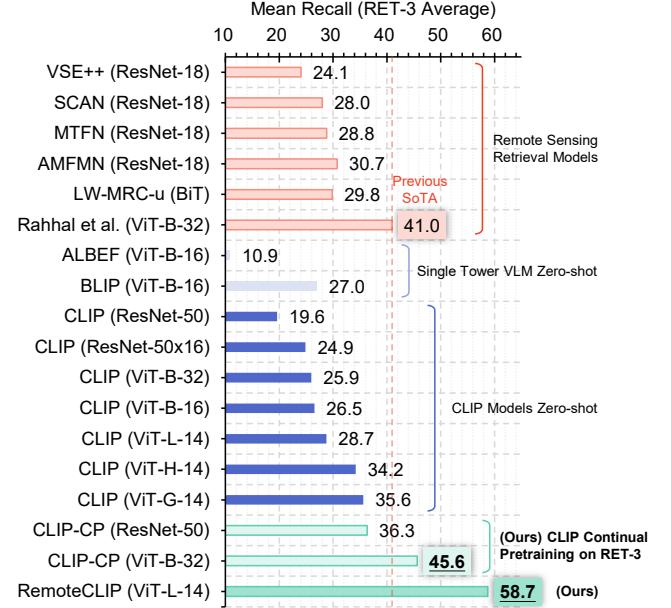


Fig. 1: Averaged mean recall on three remote sensing image-text retrieval benchmarks: RSITMD, RSICD, and UCM (RET-3). Key findings: (1) Zero-shot retrieval of large CLIP models (*e.g.*, ViT-G-14) outperforms all previous models specifically designed for remote sensing retrieval, except Rahhal *et al.* [27] that fine-tune a CLIP model. (2) Simply performing continual pretraining (CLIP-CP) significantly boosts the performance of CLIP models, and established a new SoTA.

on vision language learning, such as InstructBLIP [73] and PaLI-X [74].

Such a simple continual pretraining strategy yields encouraging results, but it is still far from perfect: when we try to scale up the model size (*e.g.*, to ViT-Large-14), a severe overfitting phenomenon appears. The reason is quite clear – the dataset we used for continual retraining is too small for a large CLIP model. The combination of all existing image-text data (RET-3) only has 13k samples, while the pretraining data for CLIP models usually range from several hundred million to several billion samples. This motivated us to perform data scaling to match the model capacity and complexity of large CLIP models. As shown in the last row in Fig. 1, such data scaling yields impressive results (+17.7% compared to the previous SoTA). We will detail our methodology of data scaling in the following section.

### B. Data Scaling via Annotation Unification

We have already shown that the vanilla CLIP model and its continual-pretrained version is promising for vision language tasks in the remote sensing domain, and we have also identified that data scale is the major bottleneck limiting the performance. These observations motivate us to scale up the dataset for continual pretraining, beyond the currently available image-text pairs (RET-3 with only 13k samples). A straightforward methodology is to annotate more captions based on crowd-sourcing, but despite this being very expensive

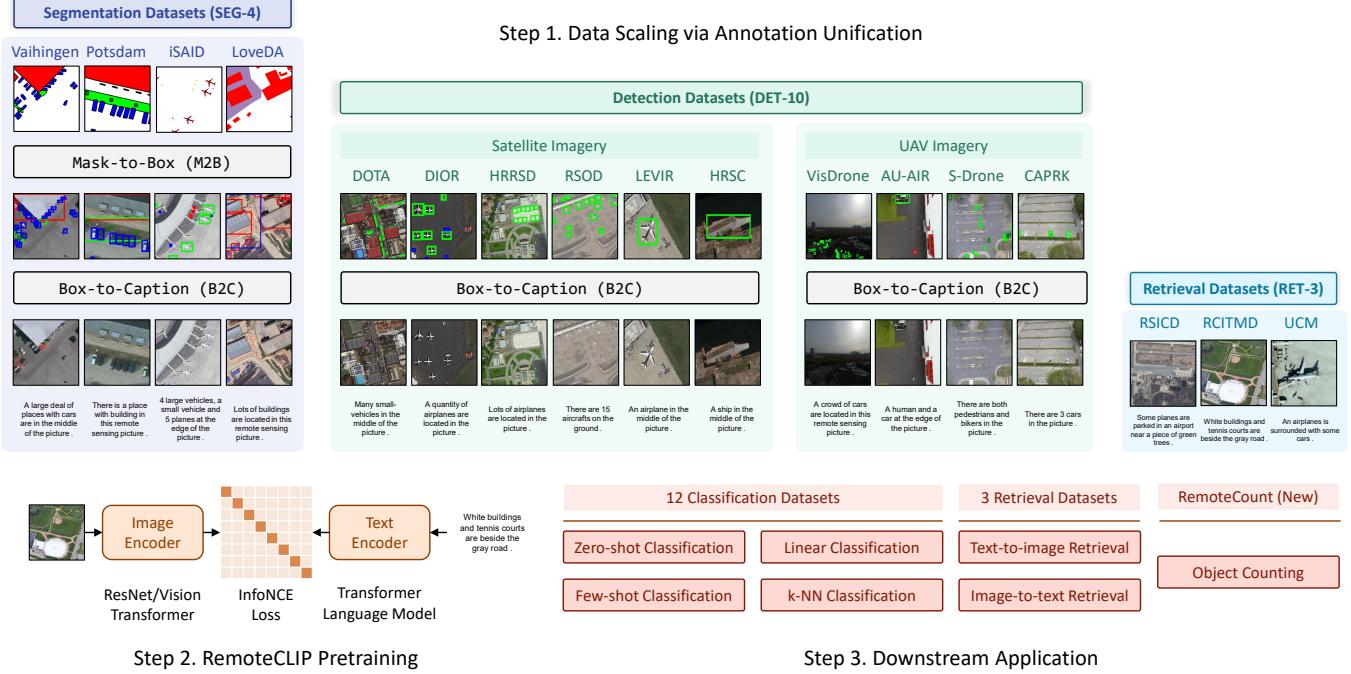


Fig. 2: Overview of the RemoteCLIP pipeline. **Step 1:** RemoteCLIP is trained on a diverse collection of remote sensing datasets, covering 10 object detection datasets (DET-10, 6 of them are satellite imagery datasets and 4 of them are UAV datasets), 4 remote sensing semantic segmentation datasets (SEG-4), and three remote sensing image-text datasets. We propose Box-to-Caption (B2C) generation and Mask-to-Box (M2B) conversion to fully utilize heterogeneous annotations, and scale up the training data to 12 $\times$  of the combination of all existing image-text data. **Step 2:** We perform continual pretraining based on the CLIP model, specializing it into the remote sensing domain. **Step 3:** we perform a comprehensive evaluation on 7 tasks using 16 downstream datasets, including a newly created RemoteCount dataset, to demonstrate the strong capability and generalization ability of RemoteCLIP.

therefore significantly lowering the scalability, the annotation quality and diversity are also hard to guarantee.

To solve this issue and thereby unleash the full potential of CLIP models, here we propose to scale up the dataset via annotation unification. We find that existing datasets annotated with object bounding boxes and class names, which were originally constructed for training object detectors, provide valuable information about the semantics within each satellite image. However, such object bounding box annotation can not be directly understood by the text encoder of CLIP, as it has only been trained on natural language captions. Therefore, we propose a Box-to-Caption (B2C) generation approach to transfer the bounding box annotations into a set of natural language captions to mitigate this gap. Further, to utilize the annotation in segmentation datasets, we proposed to use a Mask-to-Box (M2B) conversion method to unify segmentation datasets into bounding box annotations, and subsequently transfer them into captioning datasets. An overview of this process is shown in Fig. 2.

**1) Box-to-Caption (B2C) Generation:** The Box-to-Caption (B2C) generation enables the generation of textual descriptions for object detection datasets based on bounding box annotations and labels. This method employs a rule-based approach

to generate five<sup>2</sup> distinct captions that describe the objects in the image, and Algorithm 1 presents an overview of the B2C approach.

Specifically, the first two captions are generated according to the target location (the center point of the bounding box): the first caption describes the objects in the center of the image, while the second one describes the objects that are not located in the center. This differentiation provides additional context and information about the spatial distribution of objects within the image.

The remaining three captions are generated by considering the number of different object categories present in the image. Random objects from the list of bounding box annotations are selected, and a caption is generated accordingly. In cases where the number of appearances of an object exceeds ten, a more general term (*e.g.*, “many”, “a lot of”) is used instead of the exact number to enhance the readability and variability of the captions.

**2) Mask-to-Box (M2B) Conversion:** The conversion of segmentation annotations to bounding box annotations is a crucial step for the seamless integration of segmentation datasets into the B2C generation pipeline. To perform such conversion,

<sup>2</sup>Most image captioning/retrieval datasets such as MS-COCO, Flickr-30k, as well as three datasets in the RET-3 collection contain five captions for each image. We choose to generate five captions to be in line with existing datasets.

**Algorithm 1** Pseudo Code of the Box-to-Caption (B2C) Generation Method.

```

1 # Generate five different captions for one image
2 def box_to_caption(all_bbox_list)
3     all_captions = []
4
5     # Caption 1: Describe objects in the center of the image
6     caption_in_center = caption_generation(
7         objects=find_objects_in_center(all_bbox_list),
8         caption_suffix='in_the_middle_of_the_picture.'
9     )
10    all_captions.append(caption_in_center)
11
12    # Caption 2: Describe objects not in the center
13    caption_non_center = caption_generation(
14        objects=find_objects_non_center(all_bbox_list),
15        caption_suffix='in_the_edge_of_the_picture.'
16    )
17    all_captions.append(caption_non_center)
18
19    # Caption 3-5: Describe randomly sampled objects for 3
20    # times
21    for i in range(3):
22        caption_random_objects = caption_generation(
23            objects=random.sample(all_bbox_list),
24            caption_suffix='in_this_image.'
25        )
26        all_captions.append(caption_random_objects)
27
28    return all_captions
29
30 # Generate one caption from given object numbers and
31 # classnames
32 def caption_generation(objects, caption_suffix):
33     caption = f"There are"
34     for class_name, number in objects.items():
35         # If the number of object appearance is more than ten,
36         # replace the exact number with a probability of
37         # 90%
38         if number > 10 and random.random() > 0.9:
39             number = random.choice(["many", "a_lot_of"])
40             caption += f'{number}_{class_name}_'
41     return caption + caption_suffix
42
43 # Example usage
44 all_captions = box_to_caption(all_bbox_list = [
45     {'bbox': {'x': 100, 'y': 100}, 'class_name': 'airplane'},
46     {'bbox': {'x': 500, 'y': 500}, 'class_name': 'airplane'},
47     {'bbox': {'x': 800, 'y': 200}, 'class_name': 'car'},
48     ...
49 ])

```

the segmentation mask is processed by category, encoding each pixel label corresponding to the target class. Next, the contour points of the connected regions for each class in the mask image are identified. These contour points provide the necessary information to determine the bounding box coordinates. By sorting the horizontal and vertical coordinates of the contour points, we can extract the minimum and maximum values, denoted as  $(x_{min}, y_{min})$  and  $(x_{max}, y_{max})$ , respectively. These coordinate positions define the bounding box. As shown in Fig. 2, all the semantic segmentation annotations are converted to bounding box annotations via M2B, then B2C is performed to obtain the corresponding captions.

3) *Sample De-duplication:* RemoteCLIP is trained on a combination of datasets from different sources, and tested on a variety of downstream benchmarks, so it is essential to avoid possible test-set contamination. We generate p-Hash values for all images and used these values to detect duplicate images. If the number of different digits between two images is less than threshold 2, they are considered duplicates. Finally, the number of removed duplicated samples ranges from 40 to 3k in different datasets.

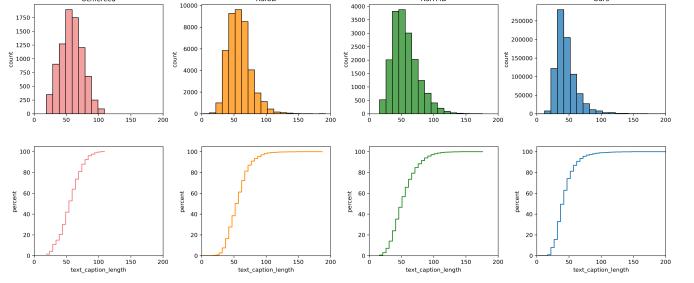


Fig. 3: Distribution of caption length of existing image-text datasets UCM (pink), RSICD (yellow), RSITMD (green) and our final dataset (blue).

### C. Data Analysis

Based on the proposed B2C generation and M2B conversion method, we can efficiently translate heterogeneous annotations in various detection or segmentation datasets into image-text samples based on the pipeline shown in Fig. 2. For a deeper understanding of the dataset produced by such scaling pipeline, in this section, we present a detailed yet comprehensive analysis for this dataset. Firstly, in Table I, we provide the details of each source dataset that we used to expand the data, which can be divided into the following three groups:

- 1) **Retrieval Data (RET-3).** All of the currently available image-text datasets for remote sensing, RSICD [25], RSITMD [24], UCM [26], are directly adopted. Captions of these datasets are annotated by human, it results in high caption quality but small dataset size.
- 2) **Detection Data (DET-10).** Detection dataset is the major source for dataset expansion. We collect six remote sensing dataset with object detection annotation, including DOTA [75], DIOR [76], HRRSD [77], RSOD [78], LEVIR [79] and HRSC [80]. As shown in Table I, these datasets have a significantly higher resolution than RET-3 datasets (at least  $800 \times 600$  vs.  $224 \times 224$ ). This group of datasets also exhibit high diversity as it consists of both satellite imagery and UAV imagery. The average number of objects appears in each image ranges from 1 (HRSC) to 70 (DOTA).
- 3) **Segmentation Data (SEG-4).** Four popular remote sensing semantic segmentation datasets, including Vaihingen [81], Postdam [82], iSAID [83], and LoveDA [84], are adopted and translated via M2B then B2C. These datasets also have high image resolution and domain diversity. Average number of objects ranges from 2 (Vaihingen) to 33 (iSAID).

In Fig. 3, we visualize the distribution of caption length of RET-3 data and our final data. It can be seen that the B2C and M2B approach produce a similar caption distribution of RET-3 data. Further, according to the word cloud visualization shown in Fig. 4, our final data covers more diverse semantics compared to the original RET-3 data.

Finally, we produce a T-SNE visualization of our final data (DET-10 + SEG-4 + Ret-3). We select 2k samples from each subset in our final data for T-SNE visualizations. For the

TABLE I: dataset statics

	Dataset	Year	#Image	#Class	#Box	Avg. Res.	Description
RET-3	RSICD [25]	2017	8483	-	-	224x224	RSICD is a dataset for remote sensing image captioning task, which contains more than ten thousands remote sensing images.
	RSITMD [24]	2021	3603	-	-	256x256	RSITMD dataset contains multi-source remote sensing images and textual descriptions.
	UCMerced [26]	2018	1676	-	-	256x256	UCMerced dataset covers 21 different scene classes, with 100 images per class.
DET-10	AUAIR [85]	2020	32,823	8	132,031	1920x1080	AU-AIR dataset features multi-modal sensor data, including visual, temporal, location, altitude, IMU, velocity, and more.
	CARPK [86]	2017	1,568	1	106,690	1280x720	CARPK dataset contains nearly 90,000 cars collected from four different parking lots by drones.
	DIOR [76]	2019	23,463	20	192,472	800x800	DIOR dataset consists of 190,288 instances of 20 different object classes, with approximately 1,200 images per class.
	DOTA [75]	2017	1,409	15	98,990	1504x1395	DOTA dataset consists of 188,282 instances of 15 different object classes, including airplanes, ships, and others.
	HRRSD [77]	2019	21,761	13	57,137	1406x1264	HRRSD dataset is used for studying object detection in high-resolution remote sensing images.
	HRSC [80]	2017	1,055	1	1,055	1105x791	HRSC dataset includes high-resolution satellite images along with corresponding ship positions and class labels.
Visdrone [88]	LEVIR [79]	2020	37,91	3	11,028	800x600	LEVIR dataset covers most types of ground features in human residential environments, such as urban, rural, mountainous, and marine areas.
	RSOD [78]	2021	936	4	7,400	1051x900	RSOD dataset includes objects such as airplanes, oil tanks, sports fields, and overpasses.
	Stanford [87]	2016	17,351	6	355,443	1424x1088	Stanford Drone dataset contains trajectory and interaction information of 20,000 objects across 8 different scenes on a campus in drones perspective.
	Visdrone [88]	2018	6,471	11	77,547	1509x849	Visdrone dataset consists of high-quality images and videos captured by UAV, along with rich object annotation information.
SEG-4	iSAID [83]	2019	30,821	15	987,239	896x896	iSAID dataset consists of a large number of high spatial resolution images and includes fifteen important and common categories.
	LoveDA [84]	2021	4,187	6	97,989	1024x1024	LoveDA dataset consists of high-resolution images with a spatial resolution of 0.3 meters and 166,768 annotated semantic objects from 3 cities.
	Potsdam [82]	2012	5,421	4	92,161	512x512	Potsdam dataset is a semantic segmentation urban remote sensing dataset and involves five foreground classes .
	Vaihingen [81]	2012	742	4	16,875	512x512	Vaihingen dataset is a semantic segmentation urban remote sensing dataset and involves the same category information as the Potsdam dataset

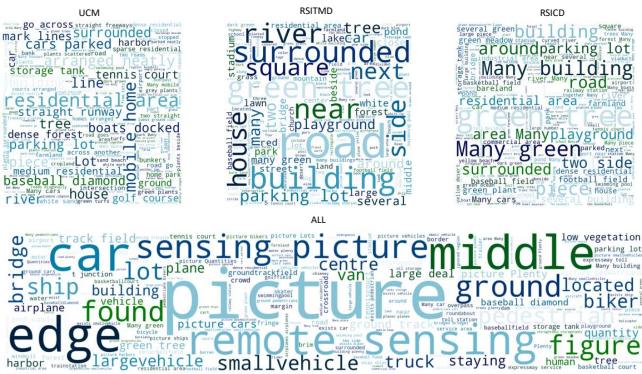


Fig. 4: Word clouds of captions in existing image-text datasets UCM, RSITMD, and RSICD (upper row) and our final dataset produced by B2C and M2B from DET-10, SEG-4, and RET-3 (bottom row).

text T-SNE visualization, we employ paraphrase-distilroberta-base-v2 from Sentence-Transformer to extract features from textual descriptions. For the image T-SNE visualization, we simply choose ViT-Base-32 from OpenCLIP to extract visual features. From Fig. 5, it can be seen that our data scaling approach provides much more enriched samples. Learning multimodal representations from such a diverse sample distribution results in a strong RemoteCLIP model that handles downstream tasks in various domains.

#### IV. EXPERIMENTS

#### A. Implementation Details

*1) Model:* We select three type of visual backbone architecture for the RemoteCLIP model, ranging from small scale model ResNet-50 (38M parameters), medium scale model ViT-Base-32 (87M parameters) and large scale model ViT-Large-14 (304M parameters), to prove that our data scaling approach benefit different size of model. The ResNet-50 structure is the OpenAI modified version, it replaces the original three  $3 \times 3$  convolutions with a single  $7 \times 7$  convolution and replaces average pooling with max pooling. Additionally, a santiialiased rect-2 blur pooling layer is added on top of the ResNet-50 architecture, and the original average pooling layer is replaced with a multi-head self-attention-based pooling. ViT-B-32 partitions the input image into fixed-size image patches of  $32 \times 32$

pixels and consists of 12 layers and 12 attention heads. ViT-L-14 partitions the input image into patches of  $14 \times 14$  pixels and comprises 24 layers and 16 attention heads. The text encoder utilizes the Transformer architecture, consisting of 12 layers and 8 attention heads. The maximum token sequence length is set to 77, same as the original OpenAI CLIP. The InfoNCE loss operate on the [CLS] token produced by image and text backbone.

2) *Data*: Our final training data consists of a total of 165,745 images, and each image have 5 corresponding captions, resulting in 828,725 training image-text pairs. In terms of data augmentation, in addition to standard operations such as random crops and horizontal flips, we also apply random rotations of 0°, 90°, 180°, and 270° degrees to the images to learn rotation invariance.

3) *Optimization*: The implementation of RemoteCLIP is based on the ITRA codebase<sup>3</sup> developed from OpenCLIP. We utilize automatic mixed-precision (AMP) to maintain model accuracy while reducing memory usage. Similar to CLIP, The training process is accelerated by employing the Adam optimizer. We adopt linear warm-up and cosine learning rate scheduler. The learning rate is set to 7e-5, 4e-5, and 1e-4 respectively for ResNet-50, ViT-Base-32, and ViT-Large-14 model, and the corresponding batch size is set to 256, 256, and 28. We train all models for a total step of 108,215. Using a single-node 4×NVIDIA 3090Ti machine, training our largest RemoteCLIP model takes 233.4 hours.

### B. Benchmarking RemoteCLIP

*1) Cross-modal Retrieval:* We first present the performance on RemoteCLIP on three remote sensing image-text retrieval benchmarks (RSITMD, RSICD, UCM) and compare it with previous results. To perform cross-modal retrieval with RemoteCLIP, we extract image and text representations on the test split, perform L-2 normalization, and retrieval most similar samples based on the dot-product similarity measure. We report the retrieval recall of top-1 (R@1), top-5 (R@5), top-10 (R@10), and the mean recall of these values. We do not perform any dataset-specific fine-tuning or re-ranking to improve the results.

Table II summarize the results. We also provide the model details of each model, including training data, backbone

<sup>3</sup><https://itra.readthedocs.io/>

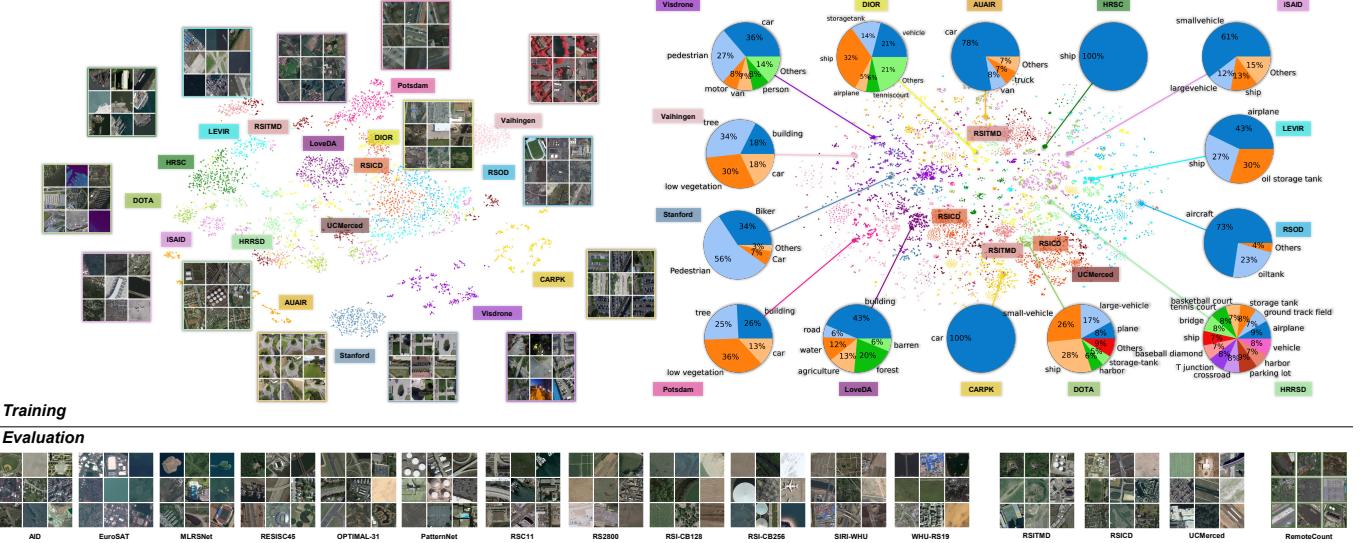


Fig. 5: T-SNE visualization of image (upper left) and caption samples (upper right) in our final dataset. We provide random image samples of each datasets and label distribution of the caption samples. In the bottom row, we visualize random samples from downstream datasets which we used for evaluation, including 12 classification datasets, 3 retrieval datasets and a novel object counting dataset.

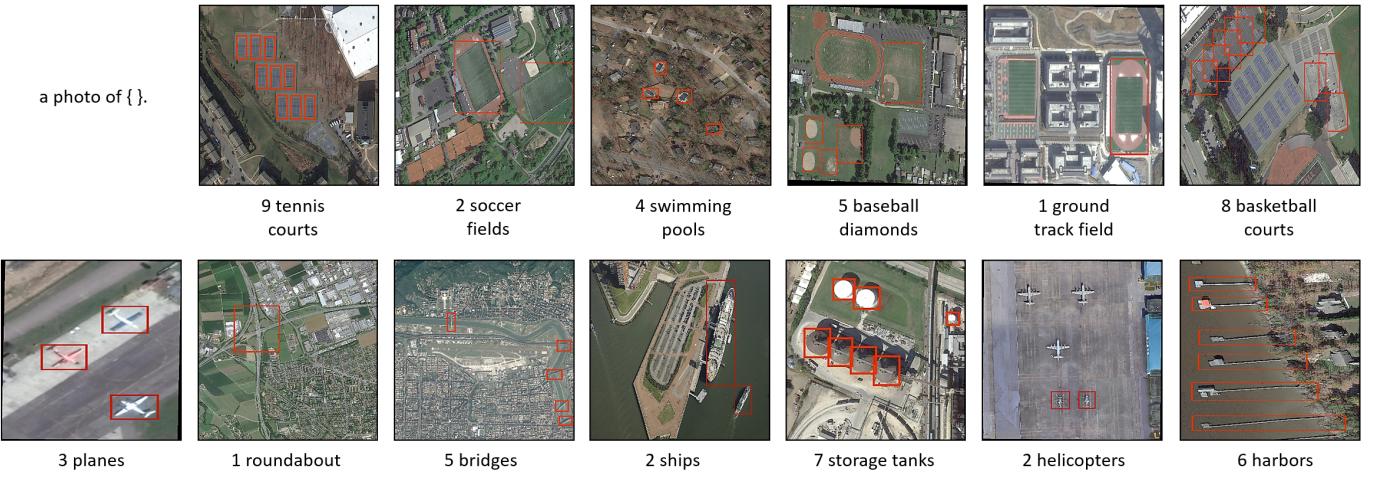


Fig. 6: Visualization of the RemoteCount dataset samples. Objects of interest are annotated by red bounding boxes.

architecture, number of parameters (in million), for better comparision. Our RemoteCLIP model achieves SoTA on all of the three retrieval benchmarks. On the challenging RSITMD and RSICD datasets, our model outperforms previous SOTA (Rahhal et al. [27] ) by a large margin (9.14% and 8.92% respectively). Such results are achieved with large-scale model ViT-Large-14 with 304 parameters, but when come to smaller models, RemoteCLIP is still competitive – the ResNet-50-based RemoteCLIP can also exceed previous SoTA on RSITMD and RSICD datasets. In addition, on all of the three retrieval benchmarks, RemoteCLIP outperforms the CLIP-CL baseline, which only uses the existing RET-3 data, showing the effectiveness of RemoteCLIP data scaling.

2) *Object Counting*: Recent study shows that large-scale pretraining empowers CLIP model to do zero-shot object counting [93]. Here we are interested in whether RemoteCLIP

have such fine-grained language understanding capability. To verify this question, we introduce a new remote sensing counting benchmark, which we named as “RemoteCount” to evaluate the ability of accurately counting objects from 1 to 10. This dataset consists of 947 image-text pairs, which are mainly selected from the validation set of the DOTA dataset. It covers 13 categories, including planes, helicopters, roundabouts, bridges, baseball diamonds, ground track fields, basketball courts, tennis courts, harbors, soccer fields, swimming pools, ships, and storage tanks. The dataset is annotated by five graduate students, careful manual verification were conducted to ensure its quality. Fig. 6 visualize random samples within RemoteCount.

We focus on comparing the zero-shot counting accuracy of CLIP and RemoteCLIP. For each image, we augment the existing caption with nine other possible captions by replacing

TABLE II: Cross-modal retrieval performance on RSITMD, RSICD and UCM benchmarks. CLIP-CL corresponds to the continual pretraining of CLIP on existing RET-3 data only (details presented in Section III-A2). Previous SoTA results are marked by **RED**, while the RemoteCLIP models are marked by **BLUE**. Our RemoteCLIP achieves SoTA performance on all of the three retrieval benchmarks by a large margin.

Testing Dataset	Training Dataset	Training Samples	Year	Method	Image Backbone		Text Backbone		Total Params	Image to Text			Text to Image			Mean Recall
					Name	Params	Name	Params		R@1	R@5	R@10	R@1	R@5	R@10	
RSITMD	4,743	2017	VSE++ [89]	VGG19	-	GRU	-	-	10.38	27.65	39.60	7.79	24.87	38.67	24.83	
		2018	SCAN [90]	ResNet-101	-	GRU	-	-	11.06	25.88	39.38	9.82	29.38	42.12	26.28	
		2019	MTFN [91]	ResNet	-	GRU	-	-	10.40	27.65	36.28	9.96	31.37	45.84	26.92	
		2020	AMFMN [92]	ResNet-50	-	GloVe fasText	-	-	10.63	24.78	41.81	11.51	34.69	54.87	29.72	
		2020	LW-MRC-u [70]	Big Transfer	-	Bi-LSTM	-	-	9.73	26.77	37.61	9.25	34.07	54.03	28.58	
		2022	GaLR [71]	ResNet-18	-	GRU	-	-	14.82	31.64	42.48	11.15	36.68	51.68	31.41	
		2022	Rahhal et al. [27]	ViT-B-32	87	Transformer	63	151	19.69	40.26	54.42	17.61	49.73	66.59	41.38	
		RET-3	13,713	CLIP-CL	ResNet-50	38	Transformer	64	102	19.25	39.82	51.33	15.09	41.46	56.64	37.27
		2023	CLIP-CL	ViT-B-32	88	Transformer	64	151	24.78	50.00	63.27	22.61	55.27	69.87	47.63	
RSICD	10,921	2023	RemoteCLIP	ResNet-50	38	Transformer	64	102	23.67	47.57	64.60	19.29	51.55	70.58	46.21	
		2023	RemoteCLIP	ViT-B-32	87	Transformer	63	151	27.88	50.66	65.71	22.17	56.46	73.41	49.38	
		2023	RemoteCLIP	ViT-L-14	304	Transformer	124	428	28.76	52.43	63.94	23.76	59.51	74.73	50.52	
		2017	VSE++ [89]	VGG19	-	GRU	-	-	3.38	9.51	17.46	2.82	11.32	18.10	10.43	
		2018	SCAN [90]	ResNet-101	-	GRU	-	-	5.85	12.89	19.84	3.71	16.40	26.73	14.23	
		2019	MTFN [91]	ResNet	-	GRU	-	-	5.02	12.52	19.74	4.90	17.17	29.49	14.81	
		2020	AMFMN [92]	ResNet-50	-	GloVe fasText	-	-	5.39	15.08	23.40	4.90	18.28	31.44	16.42	
		2020	LW-MRC-u [70]	Big Transfer	-	Bi-LSTM	-	-	4.39	13.35	20.29	4.30	18.85	32.34	15.59	
		2022	GaLR [71]	ResNet-18	-	GRU	-	-	6.59	19.85	31.04	4.69	19.48	32.13	18.96	
		2022	Rahhal et al. [27]	ViT-B-32	87	Transformer	63	151	10.70	29.64	41.53	9.14	28.96	44.59	27.43	
UCM	2,100	2023	CLIP-CL	ResNet-50	38	Transformer	64	102	12.99	26.35	36.32	8.56	25.60	39.16	24.83	
		2023	CLIP-CL	ViT-B-32	88	Transformer	64	151	17.84	35.96	50.14	13.89	35.15	50.08	33.96	
		2023	RemoteCLIP	ResNet-50	38	Transformer	64	102	13.36	32.94	44.83	10.76	32.83	48.75	30.58	
		2023	RemoteCLIP	ViT-B-32	87	Transformer	63	151	17.02	37.97	51.51	13.71	37.11	54.25	35.26	
		2023	RemoteCLIP	ViT-L-14	304	Transformer	124	428	18.39	37.42	51.05	14.73	39.93	56.58	36.35	
		2017	VSE++ [89]	VGG19	-	GRU	-	-	12.38	44.76	65.71	10.10	31.80	56.85	36.93	
		2018	SCAN [90]	ResNet-101	-	GRU	-	-	12.85	47.14	69.52	12.48	46.86	71.71	43.43	
		2019	MTFN [91]	ResNet	-	GRU	-	-	10.47	47.62	64.29	14.19	52.38	78.95	44.65	
		2020	AMFMN [92]	ResNet-50	-	GloVe fasText	-	-	16.67	45.71	68.57	12.86	53.24	79.43	46.08	
		2020	LW-MRC-u [70]	Big Transfer	-	Bi-LSTM	-	-	18.10	47.14	63.81	13.14	50.38	79.52	45.35	
		2022	Rahhal et al. [27]	ViT-B-32	87	Transformer	63	151	19.04	53.33	77.61	19.33	64.00	91.42	54.12	
RET-3	13,713	2023	CLIP-CL	ResNet-50	38	Transformer	64	102	14.29	44.76	72.86	13.52	53.71	82.10	46.87	
		2023	CLIP-CL	ViT-B-32	88	Transformer	64	151	20.00	55.24	80.95	18.19	64.86	92.38	55.27	
		2023	RemoteCLIP	ResNet-50	38	Transformer	64	102	13.33	50.48	74.76	15.24	57.14	84.57	49.25	
		2023	RemoteCLIP	ViT-B-32	87	Transformer	63	151	20.48	59.85	83.33	18.67	61.52	94.29	56.36	
		2023	RemoteCLIP	ViT-L-14	304	Transformer	124	428	19.05	54.29	80.95	17.71	62.19	93.90	54.68	

TABLE III: Zero-shot accuracies on 12 remote sensing image classification datasets.

Method	Backbone	RSI-CB128	RSI-CB256	WHU-earth	EuroSAT	MLRSNet	PatternNet	RESISC45	AID	RS2800	OPTIMAL-31	RSC11	WHU-RS19	Average
CLIP	ResNet-50	<b>26.56</b>	<b>34.24</b>	40.42	<b>42.02</b>	<b>45.54</b>	<b>46.96</b>	<b>53.57</b>	57.35	62.14	64.52	64.54	69.42	50.61
		13.95	33.03	<b>56.25</b>	17.19	40.68	45.51	53.24	<b>86.55</b>	<b>62.86</b>	<b>70.16</b>	<b>66.93</b>	<b>95.15</b>	<b>53.46</b>
		-12.61	-1.21	+15.83	-24.83	-4.86	-1.45	-0.33	+29.20	+0.72	+5.64	+2.39	+25.73	+2.85
CLIP	ViT-B-32	<b>28.88</b>	37.35	51.18	<b>47.11</b>	55.29	<b>58.95</b>	60.92	65.65	59.31	68.62	58.35	80.61	56.02
		24.18	<b>39.5</b>	<b>63.12</b>	35.96	<b>59.28</b>	57.71	<b>70.33</b>	<b>91.3</b>	<b>68.57</b>	<b>77.96</b>	<b>64.94</b>	<b>96.12</b>	<b>62.41</b>
		-4.70	+2.15	+11.94	-11.15	+3.99	-1.24	+9.41	+25.65	+9.26	+9.34	+6.59	+15.51	+6.40
CLIP	ViT-L-14	<b>40.23</b>	47.94	58.33	<b>60.21</b>	64.89	<b>73.78</b>	69.23	69.88	72.15	76.83	67.11	87.5	65.67
		37.22	<b>52.82</b>	<b>70.83</b>	59.94	<b>66.32</b>	68.75	<b>79.84</b>	<b>87.90</b>	<b>72.32</b>	<b>90.05</b>	<b>74.9</b>	<b>94.66</b>	<b>71.30</b>
		-3.01	+4.88	+12.50	-0.27	+1.43	-5.03	+10.61	+18.02	+0.17	+13.22	+7.79	+7.16	+5.62

the number in its caption with all the numbers from 1 to 10, and calculate the similarity score between the image and each of the ten captions. The number in the caption that obtains the highest similarity score with the image is considered the predicted number.

The evaluation results are provided in Fig. 7. The top row shows the confusion matrix of CLIP and RemoteCLIP. We normalized the final output confusion matrix. CLIP has shown poor results in this task, but RemoteCLIP has a clear diagonal representing its higher accuracy. We also present the top 1 to top 10 accuracy of CLIP and RemoteCLIP. It is clear that at the top 6 accuracy level, RemoteCLIP still significantly outperforms CLIP. In addition, we also test a variation of replacing the number with “1”–“10” instead of “one”–“ten”, which we denote as “(digit)” in Fig. 7. As it can be seen, RemoteCLIP is much more robust to such variation.

3) *Zero-shot Image Classification*: This section presents the zero-shot image classification evaluation of our RemoteCLIP model. We used a total of 12 downstream

datasets, including PatternNet [94], EuroSAT [95], OPTIMAL-31 [96], RSC11 [97], AID [98], MLRSNet [99], RSI-CB128 [100], RSI-CB256 [100], RESISC45 [101], WHU-earth [102], WHU-RS19 [103], RS2800 [104]. We use the standard template-based prompting method, *e.g.*, using “a satellite photo of {class name}.” as the text input to obtain the zero-shot classifier.

The details of evaluation results are shown in table III. It can be seen from the experimental results that it has an overall improvement compared with CLIP baseline. Overall, RemoteCLIP improves the averaged zero-shot accuracy improvement of +2.85%, +6.39% and +5.63% on 12 downstream datasets. Our largest RemoteCLIP, the ViT-Large-14-based model, outperformed the CLIP counterpart on 9 of 12 (75%) datasets.

However, the zero-shot performance of RemoteCLIP is consistently inferior to CLIP in some datasets. We suspect it is caused by the domain gap of image distribution. Our RemoteCLIP models are trained on a collection of high-

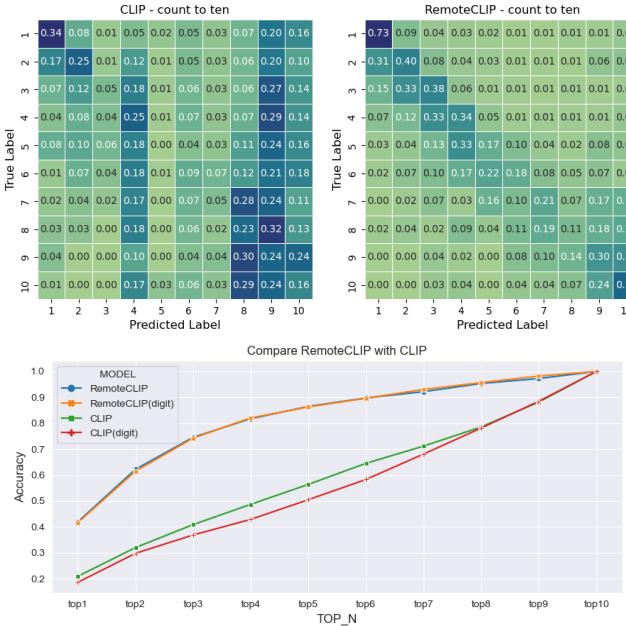


Fig. 7: The object counting experiment of CLIP and RemoteCLIP on RemoteCount dataset. Upper row: The confusion matrix of CLIP and RemoteCLIP. Bottom row: Top-1 accuracy to top-10 accuracy of CLIP and RemoteCLIP.

resolution images (see Table I for detailed statics), but several downstream datasets, such as the EuroSAT dataset, have a much lower image resolution (*e.g.*,  $64 \times 64$ ). In addition, samples used to train RemoteCLIP usually covers rich semantics and variations, while several land cover classification dataset have a much different distribution (see visualizations in Fig. 5).

**4) Few-shot Classification:** Although the zero-shot classification performance of RemoteCLIP outperforms CLIP by a significant margin, in some dataset the accuracy is still far from satisfactory. In this section, we validate that whether RemoteCLIP can be adapted to certain dataset with a few available training samples. We randomly sample few-shot training set with 1, 4, 8, 16, and 32 shot samples, and use them to train an additional linear layer on top of the image representation via logistic regression. For logistic regression, the learning rate is set to 0.8, with SGD as the optimizer, and the CosineAnnealingLR scheduling strategy is used to automatically update the learning rate. And we use CrossEntropyLoss as the criterion. We set the weight decay at  $4e-5$ , the total number of epochs at 1000, and fix the batch size at 10,000. To choose suitable parameters for few-shot classification, we train the model through 5 iterations of a random search for optimal hyperparameters. Each iteration involves the use of distinct learning rates and weight decay coefficients, while recording the accuracy achieved at each stage. Finally, the parameters associated with the best accuracy are used for the few-shot classification.

Fig. 8 shows the few-shot evaluation on 12 remote sensing classification datasets. We compare RemoteCLIP with a variety of baselines, including the vanilla CLIP model (ViT-Base-32 and ResNet-50), Self-supervised Learning (SSL-based) foundation visual models (SwAV, Barlow Twins, VICReg),

ImageNet pretrained models (ViT-Base-32 and ResNet-50), and existing remote sensing foundation models (ViTAE and SatMAE). Visualization of experimental results shows that few-shot training set could significantly boost the performance of RemoteCLIP models in all dataset. Using 32-shot samples, RemoteCLIP model outperform all compared baselines in all of the 12 dataset.

#### 5) Full-shot Linear Probing and $k$ -NN Classification:

Finally, we turn to benchmark RemoteCLIP for conventional linear probing (linear classification) and  $k$ -NN classification. We use the same 12 classification dataset that was previously used for zero-shot and few-shot evaluation. For linear classification, the hyperparameter settings are the same as those for the few-shot classification experiment. For  $k$ -NN classification, the number of nearest neighbors  $k$  is set to 20 and the temperature parameter  $T$  is set to 0.07. We take the accuracy of the top 1 category as the output of  $k$ -NN classification.

The results are shown in table IV. We found that the classification performance of RemoteCLIP was stronger than CLIP and other self-supervised models. This is not surprising that RemoteCLIP produces such strong visual representations. As mentioned in Section III-A, the vanilla CLIP model can already outperform a variety of foundation visual models in linear probing on remote sensing datasets. RemoteCLIP further enhanced such representation.

### C. Ablation Study

**Backbone ablation:** In Table V, we investigate the effects of the image and text backbones through ablation experiments conducted on the Ret-3 + Det-10 + Seg-4 dataset using RemoteCLIP. The results indicate that the optimal outcome is achieved when both the image and text backbones are pre-trained. Furthermore, the experiment highlights the greater significance of pre-training the image backbone compared to pre-training the text backbone.

**Preprocessing ablation:** To ensure controlled conditions, we conducted ablation experiments on the RET-3 dataset. As shown in Table V, the retrieval results exhibit higher performance with the application of rotation.

**Pre-training model ablation:** The experimental findings can be observed in Table V. When comparing RemoteCLIP to prior pre-training techniques, notable advancements were observed in both the Retrieval task and Zero-shot tasks. Specifically, RemoteCLIP exhibited substantial improvements of approximately 10% and 15% in the Retrieval task and Zero-shot tasks, respectively.

**Dataset ablation:** To explore the validity of sentence-making rules, we conducted an ablation experiment on the dataset using RemoteCLIP. We applied different sentence-making rules and evaluated their impact on the same test set. The experimental results, presented in Table V, revealed superior performance with the Ret-3 + Det-10 + Seg-4 dataset. These findings indicate the task's demand for richer textual information and affirm the effectiveness of our sentence-making strategy.

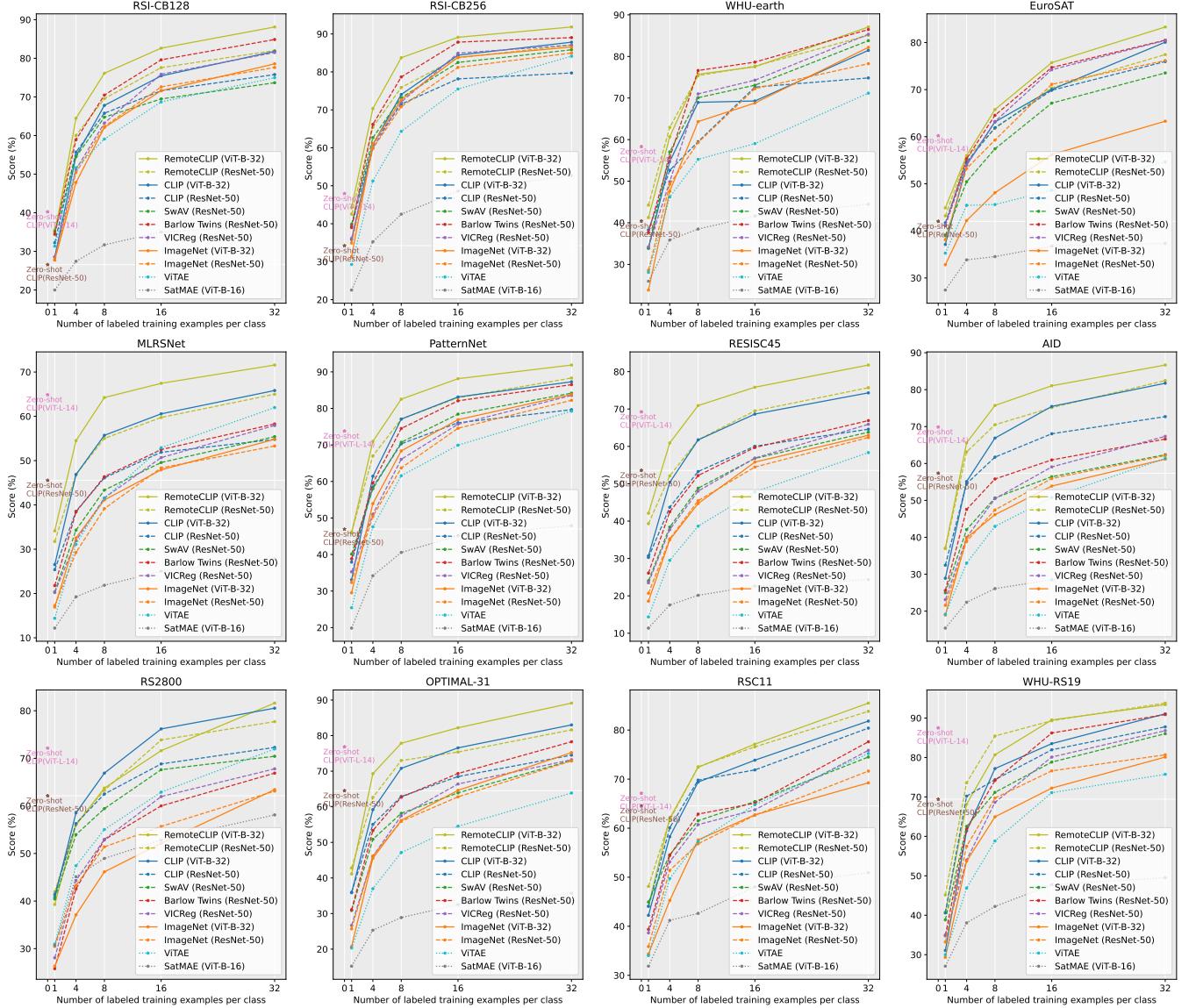


Fig. 8: Few-shot classification results on 12 remote sensing datasets. Zero-shot results of CLIP models are marked by brown for ResNet-50 and by pink for ViT-L-14. On the 32-shot setting, RemoteCLIP outperform all of the compared models on all 12 datasets.

## V. CONCLUSION

In this paper, we present RemoteCLIP, the first general-purpose vision-language foundation model for remote sensing. During developing such foundation model, our key insights are two-fold. First, CLIP models, which are pretrained on massive image-text pairs collected from the internet, is a surprisingly powerful model for remote sensing tasks. Secondly, although in-domain fine-tuning (*i.e.*, continual pretraining) significantly improves that performance, data quantity becomes a major bottleneck in this process, especially when we attempts to specialize large CLIP models into the remote sensing domain.

Based on the above observations, we develop a pipeline for data scaling and subsequently tune the CLIP models on the expanded dataset. The resulting RemoteCLIP model showed very powerful results on downstream tasks. Importantly, the complexity of RemoteCLIP model architecture design, loss

function design are extremely simple – RemoteCLIP established a series of SoTA on various benchmarks, without any bells and whistles. It highlights the importance of data-centric methodology in developing foundation models, such observation is also in line with other efforts of building in-domain foundation models, such as BioMedCLIP [54] in the medical domain.

Although, RemoteCLIP still has several known limitations, and we would like to address these issues in our future works:

- First, our largest RemoteCLIP is initialized from OpenAI's ViT-Large-14 CLIP model, which as 304M parameters in its visual backbone, and was trained on 400M data. Although it is significantly larger than previous remote sensing retrieval models, there are still much room for further scale-up. For example, Billion-scale MAE [16] proved that ViT with 2B scale can be successfully applied

TABLE IV: Linear probing and and  $k$ -NN classification results on 12 remote sensing datasets.

Method	Backbone	RSI-CB128		RSI-CB256		WHU-earth		EuroSAT		MLRSNet		PatternNet		RESISC45		AID		RS2800		OPTIMAL-31		RSC11		WHU-RS19		Average	
		Linear	$k$ -NN	Linear	$k$ -NN	Linear	$k$ -NN	Linear	$k$ -NN	Linear	$k$ -NN	Linear	$k$ -NN	Linear	$k$ -NN	Linear	$k$ -NN	Linear	$k$ -NN	Linear	$k$ -NN	Linear	$k$ -NN	Linear	$k$ -NN		
ImageNet	ResNet-50	95.69	93.24	97.92	97.40	92.92	93.69	91.48	88.41	78.99	74.78	96.18	93.45	86.16	83.60	83.03	79.45	75.89	79.29	87.10	86.29	79.68	78.09	95.63	90.21	87.32	87.08
SwAV		95.27	95.61	98.59	98.17	95.20	93.96	91.17	91.37	79.04	76.12	96.94	94.18	88.60	85.59	86.00	80.80	81.07	86.07	88.44	84.14	84.88	78.89	96.12	92.23	88.97	88.83
Barlow Twins		98.07	95.90	99.03	98.13	95.83	95.42	94.78	91.57	82.41	77.55	97.73	93.83	91.10	86.10	88.22	81.75	77.32	86.07	91.94	86.83	85.26	78.09	97.09	91.75	90.00	89.73
VICReg		97.47	96.03	99.07	97.44	95.46	95.06	91.66	82.49	78.02	96.36	94.94	90.03	86.30	88.10	81.77	77.32	86.09	96.32	84.73	87.74	96.00	90.48	93.85	89.56	89.56	
CLIP		94.89	94.05	97.30	97.24	93.12	91.88	91.67	88.54	80.08	77.14	95.61	92.86	85.73	85.65	90.95	86.90	83.75	81.43	88.99	87.63	87.65	87.25	97.57	93.69	89.47	89.42
CLIP-CL		95.99	94.92	98.41	98.09	96.25	94.79	89.80	87.65	79.32	76.99	97.30	95.15	89.10	92.67	94.27	92.67	95.95	92.55	86.96	87.86	95.97	94.35	91.63	84.86	98.06	97.57
RemoteCLIP		96.06	94.78	98.39	97.62	95.42	95.63	92.56	90.20	83.32	81.21	97.37	95.95	90.94	90.05	94.35	90.10	85.00	89.46	92.74	90.86	91.63	85.66	98.06	95.63	92.06	
ImageNet	ViT-Base	96.45	91.29	98.11	97.00	93.75	91.67	85.57	76.56	78.61	74.05	96.81	92.98	86.89	81.63	83.55	76.45	78.93	78.04	89.51	81.18	81.67	80.88	94.17	89.81	86.34	86.05
VITAE		93.10	95.65	98.41	94.05	93.33	78.96	61.41	82.27	91.15	80.37	98.50	90.82	87.94	65.33	88.30	64.05	92.86	78.93	86.29	54.84	92.83	71.31	91.74	70.39	84.02	83.03
CLIP		97.36	94.17	98.55	97.40	95.00	92.08	95.15	90.28	85.43	82.26	97.58	94.36	92.60	89.73	94.95	90.35	88.57	88.21	93.55	90.86	90.84	86.85	97.09	93.69	92.31	92.15
RemoteCLIP		98.02	95.82	99.01	98.51	95.42	97.08	96.19	93.50	87.00	85.11	98.47	97.32	94.27	92.67	95.95	92.55	86.96	87.86	95.97	94.35	91.63	89.24	97.57	94.17	93.93	93.77

TABLE V: Ablation study results. From left to right: pertrianing ablation, backbone ablation, dataset ablation, and preprocessing ablation. Settings adopted by RemoteCLIP is marked by blue.

Backbone	Method	Retrial	Average	Image	Pre-trained	Text	Pre-trained	Retrial	Average	Preprocessing	Retrial	Average
		Retrial	Average					Retrial	Average			
ResNet-50	ImageNet	37.07	44.36	ResNet-50	✓	✓	✓	42.01	53.46	Rotation Augmentation	38.90	47.98
	SwAV	34.6	44.59		✗	✓	✓	25.56	37.46		37.74	48.05
	VICReg	34.28	41.01		✓	✗	✗	35.72	46.03		37.98	47.18
	Barlow Twins	32.95	40.36		✗	✗	✗	24.44	36.97		37.99	48.07
ViT-Base	CLIP	42.01	55.06		✓	✓	✓	47.00	64.52		42.01	53.46
	VITAE	39.08	47.85	ViT-B-32	✗	✓	✓	21.56	42.60		37.24	46.94
	VITAE	38.75	48.5		✓	✗	✗	37.13	54.30		39.72	51.31
	DINOv2	38.14	50.24		✗	✗	✗	18.92	30.93		42.01	53.46
ImageNet	ImageNet	35.08	46.19		✓	✓	✓	✓	✓		✓	✓
	CLIP	47.00	64.52		✓	✓	✓	✓	✓		✓	✓

in the remote sensing imagery.

- Keeping scaling up the model size require the data scale to be expended simultaneously. RemoteCLIP data, although already become  $12\times$  larger than the combination of all existing image-text data, could be still far from sufficient to train a much larger model. In the future, we would like to expend the pretraining data by further incorporating weakly-labeled data (classification datasets) and unlabelled data (via pseudo labelling).
- Data quality and diversity is important. Although our B2C and M2B approach effectively translate heterogeneous annotations (*e.g.*, bounding box and segmentation maps) into homogeneous captions, our rule-based conversion methodology has limited diversity. In our future works, we will consider generating richer captions by introducing generative language models. In addition, the modality diversity of RemoteCLIP is limited, and exploring more sensory modality beyond RGB is a promising direction.

## REFERENCES

- [1] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. S. Chatterji, A. S. Chen, K. A. Creel, J. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. E. Gillespie, K. Goel, N. D. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. F. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. S. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladakh, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. F. Nyarko, G. Ogut, L. J. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelman, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. H. Roohani, C. Ruiz, J. Ryan, C. R'e, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. P. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. A. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang, “On the opportunities and risks of foundation models,” *ArXiv*, vol. abs/2108.07258, 2021.
- [2] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, “A simple framework for contrastive learning of visual representations,” *ArXiv*, vol. abs/2002.05709, 2020.
- [3] K. He, X. Chen, S. Xie, Y. Li, P. Doll’ar, and R. B. Girshick, “Masked autoencoders are scalable vision learners,” *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15 979–15 988, 2021.
- [4] L. Yuan, D. Chen, Y.-L. Chen, N. C. F. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li, C. Liu, M. Liu, Z. Liu, Y. Lu, Y. Shi, L. Wang, J. Wang, B. Xiao, Z. Xiao, J. Yang, M. Zeng, L. Zhou, and P. Zhang, “Florence: A new foundation model for computer vision,” *ArXiv*, vol. abs/2111.11432, 2021.
- [5] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. B. Girshick, “Segment anything,” *ArXiv*, vol. abs/2304.02643, 2023.
- [6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *ArXiv*, vol. abs/1810.04805, 2019.
- [7] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Sastry, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. J. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language models are few-shot learners,” *ArXiv*, vol. abs/2005.14165, 2020.
- [8] OpenAI, “Gpt-4 technical report,” *ArXiv*, vol. abs/2303.08774, 2023.
- [9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” in *International Conference on Machine Learning*, 2021.
- [10] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, “Flamingo: A visual language model for few-shot learning,” *ArXiv*, vol. abs/2204.14198, 2022.
- [11] H. Bao, L. Dong, and F. Wei, “Beit: Bert pre-training of image transformers,” *ArXiv*, vol. abs/2106.08254, 2021.
- [12] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, “Simmim: a simple framework for masked image modeling,” *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9643–9653, 2021.
- [13] Y. Cong, S. Khanna, C. Meng, P. Liu, E. Rozi, Y. He, M. Burke, D. Lobell, and S. Ermon, “Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery,” *ArXiv*, vol. abs/2207.08051, 2022.
- [14] C. Reed, R. Gupta, S. Li, S. Brockman, C. Funk, B. Clipp, S. Candido, M. Uyttendaele, and T. Darrell, “Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning,” *ArXiv*, vol. abs/2212.14532, 2022.
- [15] D. Wang, Q. Zhang, Y. Xu, J. Zhang, B. Du, D. Tao, and L. Zhang, “Advancing plain vision transformer toward remote sensing foundation model,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2022.

- [16] K. Cha, J. Seo, and T. Lee, “A billion-scale foundation model for remote sensing images,” *ArXiv*, vol. abs/2304.05215, 2023.
- [17] X. Sun, P. Wang, W. Lu, Z. Zhu, X. Lu, Q. He, J. Li, X. Rong, Z. Yang, H. Chang, Q. He, G. Yang, R. Wang, J. Lu, and K. Fu, “Ringmo: A remote sensing foundation model with masked image modeling,” *IEEE Transactions on Geoscience and Remote Sensing*, 2022.
- [18] M. Mendieta, B. Han, X. Shi, Y. Zhu, C. Chen, and M. Li, “Gfm: Building geospatial foundation models via continual pretraining,” *ArXiv*, vol. abs/2302.04476, 2023.
- [19] X. Kong and X. Zhang, “Understanding masked image modeling via learning occlusion invariant feature,” *ArXiv*, vol. abs/2208.04164, 2022.
- [20] N. Park, W. Kim, B. Heo, T. Kim, and S. Yun, “What do self-supervised vision transformers learn?” *ArXiv*, vol. abs/2305.00729, 2023.
- [21] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. Wichmann, and W. Brendel, “Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness,” *ArXiv*, vol. abs/1811.12231, 2018.
- [22] G. Mai, C. Cundy, K. Choi, Y. Hu, N. Lao, and S. Ermon, “Towards a foundation model for geospatial artificial intelligence (vision paper),” *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, 2022.
- [23] G. Mai, W. Huang, J. Sun, S. Song, D. Mishra, N. Liu, S. Gao, T. Liu, G. Cong, Y. Hu, C. Cundy, Z. Li, R. Zhu, and N. Lao, “On the opportunities and challenges of foundation models for geospatial artificial intelligence,” *ArXiv*, vol. abs/2304.06798, 2023.
- [24] Z. Yuan, W. Zhang, K. Fu, X. Li, C. Deng, H. Wang, and X. Sun, “Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–19, 2021.
- [25] X. Lu, B. Wang, X. Zheng, and X. Li, “Exploring models and data for remote sensing image caption generation,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, pp. 2183–2195, 2017.
- [26] Y. Yang and S. Newsam, “Bag-of-visual-words and spatial extensions for land-use classification,” in *ACM SIGSPATIAL International Workshop on Advances in Geographic Information Systems*, 2010.
- [27] M. M. A. Rahhal, Y. Bazi, N. A. Alsharif, L. Bashmal, N. A. Alajlan, and F. Melgani, “Multilanguage transformer for improved text to remote sensing image retrieval,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 9115–9126, 2022.
- [28] Y. Du, Z. Liu, J. Li, and W. X. Zhao, “A survey of vision-language pre-trained models,” in *International Joint Conference on Artificial Intelligence*, 2022.
- [29] S. Long, F. Cao, S. C. Han, and H. Yang, “Vision-and-language pretrained models: A survey,” *ArXiv*, vol. abs/2204.07356, 2022.
- [30] Z. Gan, L. Li, C. Li, L. Wang, Z. Liu, and J. Gao, “Vision-language pre-training: Basics, recent advances, and future trends,” *Found. Trends Comput. Graph. Vis.*, vol. 14, pp. 163–352, 2022.
- [31] F. Chen, D. Zhang, M. Han, X. Chen, J. Shi, S. Xu, and B. Xu, “Vlp: A survey on vision-language pre-training,” *Machine Intelligence Research*, vol. 20, pp. 38 – 56, 2022.
- [32] J. Zhang, J. Huang, S. Jin, and S. Lu, “Vision-language models for vision tasks: A survey,” *ArXiv*, vol. abs/2304.00685, 2023.
- [33] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y.-H. Sung, Z. Li, and T. Duerig, “Scaling up visual and vision-language representation learning with noisy text supervision,” in *International Conference on Machine Learning*, 2021.
- [34] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, “Open-vocabulary object detection via vision and language knowledge distillation,” in *International Conference on Learning Representations*, 2021.
- [35] Y. Zhong, J. Yang, P. Zhang, C. Li, N. C. F. Codella, L. H. Li, L. Zhou, X. Dai, L. Yuan, Y. Li, and J. Gao, “Regionclip: Region-based language-image pretraining,” 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16 772–16 782, 2021.
- [36] G. Ghiasi, X. Gu, Y. Cui, and T.-Y. Lin, “Open-vocabulary image segmentation,” *ArXiv*, vol. abs/2112.12143, 2021.
- [37] F. Liang, B. Wu, X. Dai, K. Li, Y. Zhao, H. Zhang, P. Zhang, P. Vajda, and D. Marculescu, “Open-vocabulary semantic segmentation with mask-adapted clip,” *ArXiv*, vol. abs/2210.04150, 2022.
- [38] R. Mokady, “Clipcap: Clip prefix for image captioning,” *ArXiv*, vol. abs/2111.09734, 2021.
- [39] Y. Tewel, Y. Shalev, I. Schwartz, and L. Wolf, “ZeroCap: Zero-shot image-to-text generation for visual-semantic arithmetic,” 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17 897–17 907, 2021.
- [40] S. Shen, L. H. Li, H. Tan, M. Bansal, A. Rohrbach, K.-W. Chang, Z. Yao, and K. Keutzer, “How much can clip benefit vision-and-language tasks?” *ArXiv*, vol. abs/2107.06383, 2021.
- [41] H. Song, L. Dong, W. Zhang, T. Liu, and F. Wei, “Clip models are few-shot learners: Empirical studies on vqa and visual entailment,” in *Annual Meeting of the Association for Computational Linguistics*, 2022.
- [42] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 674–10 685, 2021.
- [43] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev, “Reproducible scaling laws for contrastive language-image learning,” *ArXiv*, vol. abs/2212.07143, 2022.
- [44] N. Mu, A. Kirillov, D. A. Wagner, and S. Xie, “Slip: Self-supervision meets language-image pre-training,” *ArXiv*, vol. abs/2112.12750, 2021.
- [45] Y. Li, F. Liang, L. Zhao, Y. Cui, W. Ouyang, J. Shao, F. Yu, and J. Yan, “Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm,” *ArXiv*, vol. abs/2110.05208, 2021.
- [46] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, “Coca: Contrastive captioners are image-text foundation models,” *ArXiv*, vol. abs/2205.01917, 2022.
- [47] Y. Li, H. Fan, R. Hu, C. Feichtenhofer, and K. He, “Scaling language-image pre-training via masking,” *ArXiv*, vol. abs/2212.00794, 2022.
- [48] D. Chen, Z. Wu, F. Liu, Z. Yang, Y. Huang, Y. Bao, and E. Zhou, “Prototypical contrastive language image pretraining,” *ArXiv*, vol. abs/2206.10996, 2022.
- [49] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, “Learning to prompt for vision-language models,” *International Journal of Computer Vision*, vol. 130, pp. 2337 – 2348, 2021.
- [50] H. Xue, Y. Sun, B. Liu, J. Fu, R. Song, H. Li, and J. Luo, “Clip-vip: Adapting pre-trained image-text model to video-language representation alignment,” *ArXiv*, vol. abs/2209.06430, 2022.
- [51] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. Langlotz, “Contrastive learning of medical visual representations from paired images and text,” *ArXiv*, vol. abs/2010.00747, 2020.
- [52] S. Eslami, G. de Melo, and C. Meinel, “Does clip benefit visual question answering in the medical domain as much as it does in the general domain?” *ArXiv*, vol. abs/2112.13906, 2021.
- [53] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, “Medclip: Contrastive learning from unpaired medical images and text,” *ArXiv*, vol. abs/2210.10163, 2022.
- [54] S. Zhang, Y. Xu, N. Usuyama, J. K. Bagga, R. Tinn, S. Preston, R. N. Rao, M.-H. Wei, N. Valluri, C. Wong, M. P. Lungren, T. Naumann, and H. Poon, “Large-scale domain-specific pretraining for biomedical vision-language processing,” *ArXiv*, vol. abs/2303.00915, 2023.
- [55] X. Dong, X. Zhan, Y. Wu, Y. Wei, M. C. Kampffmeyer, X. Wei, M. Lu, Y. Wang, and X. Liang, “M5product: Self-harmonized contrastive learning for e-commercial multi-modal pretraining,” 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21 220–21 230, 2021.
- [56] F. Liu, D. Chen, X. Du, R. Gao, and F. Xu, “Mep-3m: A large-scale multi-modal e-commerce product dataset,” *Pattern Recognition*, 2023.
- [57] W. Shin, J. Park, T. Woo, Y. Cho, K. Oh, and H. Song, “e-clip: Large-scale vision-language representation learning in e-commerce,” *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022.
- [58] J. Li, R. R. Selvaraju, A. D. Gotmare, S. R. Joty, C. Xiong, and S. C. H. Hoi, “Align before fuse: Vision and language representation learning with momentum distillation,” *ArXiv*, vol. abs/2107.07651, 2021.
- [59] J. Li, D. Li, S. Savarese, and S. Hoi, “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” *ArXiv*, vol. abs/2301.12597, 2023.
- [60] S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O. K. Mohammed, Q. Liu, K. Aggarwal, Z. Chi, J. Bjork, V. Chaudhary, S. Som, X. Song, and F. Wei, “Language is not all you need: Aligning perception with language models,” *ArXiv*, vol. abs/2302.14045, 2023.
- [61] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” *ArXiv*, vol. abs/2304.08485, 2023.
- [62] A. M. Swope, X. Rudelis, and K. T. Story, “Representation learning for remote sensing: An unsupervised sensor fusion approach,” *ArXiv*, vol. abs/2108.05094, 2019.
- [63] H. Jung, Y. Oh, S. Jeong, C. Lee, and T. Jeon, “Contrastive self-supervised learning with smoothed representation for remote sensing,” *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

- [64] N. Jean, S. Wang, A. Samar, G. Azzari, D. Lobell, and S. Ermon, “Tile2vec: Unsupervised representation learning for spatially distributed data,” *ArXiv*, vol. abs/1805.02855, 2018.
- [65] J. Kang, R. Fernández-Beltran, P. Duan, S. Liu, and A. J. Plaza, “Deep unsupervised embedding for remotely sensed images based on spatially augmented momentum contrast,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, pp. 2598–2610, 2020.
- [66] Z. Zhao, Z. Luo, J. Li, C. Chen, and Y. Piao, “When self-supervised learning meets scene classification: Remote sensing scene classification based on a multitask learning framework,” *Remote. Sens.*, vol. 12, p. 3276, 2020.
- [67] W. Li, K. Chen, H. Chen, and Z. Shi, “Geographical knowledge-driven representation learning for remote sensing images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. PP, pp. 1–16, 2021.
- [68] O. Mañas, A. Lacoste, X. G. i Nieto, D. Vázquez, and P. R. López, “Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data,” *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9394–9403, 2021.
- [69] T. A. M. Ali, Y. Bazi, M. M. A. Rahhal, M. L. Mekhalfi, L. Rangarajan, and M. A. A. Zuair, “Textrs: Deep bidirectional triplet network for matching text to remote sensing images,” *Remote. Sens.*, vol. 12, p. 405, 2020.
- [70] M. M. A. Rahhal, Y. Bazi, T. Abdullah, M. L. Mekhalfi, and M. A. A. Zuair, “Deep unsupervised embedding for remote sensing image retrieval using textual cues,” *Applied Sciences*, 2020.
- [71] Z. Yuan, W. Zhang, C. Tian, X. Rong, Z. Zhang, H. Wang, K. Fu, and X. Sun, “Remote sensing cross-modal text-image retrieval based on global and local information,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.
- [72] J. Roberts, K. Han, and S. Albanie, “Satin: A multi-task metadata set for classifying satellite imagery using vision-language models,” *ArXiv*, vol. abs/2304.11619, 2023.
- [73] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, “Instructclip: Towards general-purpose vision-language models with instruction tuning,” *ArXiv*, vol. abs/2305.06500, 2023.
- [74] X. Chen, J. Djolonga, P. Padlewski, B. Mustafa, S. Changpinyo, J. Wu, C. R. Ruiz, S. Goodman, X. Wang, Y. Tay, S. Shakeri, M. Dehghani, D. Salz, M. Lucic, M. Tschannen, A. Agrawal, H. Hu, M. Joshi, B. Pang, C. Montgomery, P. Pietrzak, M. Ritter, A. J. Piergiovanni, M. Minderer, F. Pavetic, A. Waters, G. Li, I. Alabdulmohsin, L. Beyer, J. Amelot, K. Lee, A. Steiner, Y. Li, D. Keysers, A. Arnab, Y. Xu, K. Rong, A. Kolesnikov, M. Seyedhosseini, A. Angelova, X. Zhai, N. Houlsby, and R. Soricut, “Pali-x: On scaling up a multilingual vision and language model,” 2023.
- [75] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. J. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, “Dota: A large-scale dataset for object detection in aerial images,” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3974–3983, 2017.
- [76] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, “Object detection in optical remote sensing images: A survey and a new benchmark,” *ArXiv*, vol. abs/1909.00133, 2019.
- [77] Y. Zhang, Y. Yuan, Y. Feng, and X. Lu, “Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, pp. 5535–5548, 2019.
- [78] W. Sun, L. Dai, X. Zhang, P. Chang, and X. He, “Rsod: Real-time small object detection algorithm in uav-based traffic monitoring,” *Applied Intelligence*, vol. 52, pp. 8448 – 8463, 2021.
- [79] H. Chen and Z. Shi, “A spatial-temporal attention-based method and a new dataset for remote sensing image change detection,” *Remote. Sens.*, vol. 12, p. 1662, 2020.
- [80] Z. Liu, L. Yuan, L. Weng, and Y. Yang, “A high resolution optical satellite image dataset for ship recognition and some new baselines,” in *International Conference on Pattern Recognition Applications and Methods*, 2017.
- [81] “Vaihingen dataset,” <https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx>, 2012.
- [82] “Potsdam dataset,” <https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx>, 2012.
- [83] S. W. Zamir, A. Arora, A. Gupta, S. H. Khan, G. Sun, F. S. Khan, F. Zhu, L. Shao, G. Xia, and X. Bai, “Isaid: A large-scale dataset for instance segmentation in aerial images,” in *CVPR Workshops*, 2019.
- [84] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, “Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation,” *ArXiv*, vol. abs/2110.08733, 2021.
- [85] S. Vujasinović, S. Becker, T. Breuer, S. Bullinger, N. Scherer-Negenborn, and M. Arens, “Integration of the 3d environment for uav onboard visual object tracking,” *Applied Sciences*, 2020.
- [86] M.-R. Hsieh, Y.-L. Lin, and W. H. Hsu, “Drone-based object counting by spatially regularized regional proposal network,” *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 4165–4173, 2017.
- [87] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, “Learning social etiquette: Human trajectory understanding in crowded scenes,” in *European Conference on Computer Vision*, 2016.
- [88] P. F. Zhu, L. Wen, X. Bian, H. Ling, and Q. Hu, “Vision meets drones: A challenge,” *ArXiv*, vol. abs/1804.07437, 2018.
- [89] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, “Vse++: Improving visual-semantic embeddings with hard negatives,” in *British Machine Vision Conference*, 2017.
- [90] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, “Stacked cross attention for image-text matching,” *ArXiv*, vol. abs/1803.08024, 2018.
- [91] T. Wang, X. Xu, Y. Yang, A. Hanjalic, H. T. Shen, and J. Song, “Matching images and text with multi-modal tensor fusion and re-ranking,” *Proceedings of the 27th ACM International Conference on Multimedia*, 2019.
- [92] G. Hoxha, F. Melgani, and B. Demir, “Toward remote sensing image retrieval under a deep image captioning perspective,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 4462–4475, 2020.
- [93] R. Paiss, A. Ephrat, O. Tov, S. Zada, I. Mosseri, M. Irani, and T. Dekel, “Teaching CLIP to count to ten,” *CoRR*, vol. abs/2302.12066, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2302.12066>
- [94] W. Zhou, S. Newsam, C. Li, and Z. Shao, “Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval,” *ArXiv*, vol. abs/1706.03424, 2017.
- [95] P. Helber, B. Bischke, A. R. Dengel, and D. Borth, “Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, pp. 2217–2226, 2017.
- [96] Q. Wang, S. Liu, J. Chanussot, and X. Li, “Scene classification with recurrent attention of vhr remote sensing images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, pp. 1155–1167, 2019.
- [97] L. Zhao, P. Tang, and L. Huo, “Feature significance-based multibag-of-visual-words model for remote sensing image scene classification,” *Journal of Applied Remote Sensing*, vol. 10, 2016.
- [98] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, “Aid: A benchmark data set for performance evaluation of aerial scene classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, pp. 3965–3981, 2016.
- [99] X. Qi, P. Zhu, Y. Wang, L. Zhang, J. Peng, M. Wu, J. Chen, X. Zhao, N. Zang, and P. T. Mathiopoulos, “Mirsnet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding,” *ArXiv*, vol. abs/2010.00243, 2020.
- [100] H. Li, C. Tao, Z. Wu, J. Chen, J. Gong, and M. Deng, “Rsi-cb: A large-scale remote sensing image classification benchmark using crowdsourced data,” *Sensors (Basel, Switzerland)*, vol. 20, 2017.
- [101] G. Cheng, J. Han, and X. Lu, “Remote sensing image scene classification: Benchmark and state of the art,” *Proceedings of the IEEE*, vol. 105, pp. 1865–1883, 2017.
- [102] B. Zhao, Y. Zhong, G.-S. Xia, and L. Zhang, “Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, pp. 2108–2123, 2016.
- [103] G.-S. Xia, W. Yang, J. Delon, Y. Gousseau, H. Sun, and H. Maître, “Structural high-resolution satellite image indexing,” 2010.
- [104] Q. Zou, L. Ni, T. Zhang, and Q. Wang, “Deep learning based feature selection for remote sensing scene classification,” *IEEE Geoscience and Remote Sensing Letters*, vol. 12, pp. 2321–2325, 2015.