

Supplementary Materials: Selection and Reconstruction of Key Locals: A Novel Specific Domain Image-Text Retrieval Method

Anonymous Authors

1 VISUAL ANALYSIS OF RETRIEVAL RESULTS

We conducted visualizations for Remote Sensing Image-Text Retrieval (RSITR) as shown in Figure 1, and for Text-Image Person Re-identification (TIReID) as shown in Figure 2.

1.1 Visual Analysis of RSITR Retrieval Results

Some RSITR retrieval results are illustrated in Figure 1, where part (a) displays two examples of image retrieval (text-to-image retrieval) and part (b) shows two examples of text retrieval (image-to-text retrieval).

Text-to-image retrieval results: As shown in Figure 1(a), for the query involving four planes in an open field, the top five retrieved images depict airport scenes. The highest-ranked image accurately matches the query with four planes. In the case of retrieving red industrial zones, the correct result is ranked second. This discrepancy may stem from small, deep red industrial areas in the top-ranked image causing interference.

Image-to-text retrieval results: As shown in Figure 1(b). When retrieving sentences for an image of the plane, nearly all retrieved sentences mention "plane", confirming the model's ability to recognize the plane in the image. The first two sentences are matching texts, both mentioning "a blue house" and "plane". For the second image query, four matched texts were also accurately retrieved, demonstrating the excellent text retrieval performance of the EKLSR model.

1.2 Visual Analysis of TIReID Retrieval Results

Figure 2 presents the top-10 retrieval results from our proposed EKLSR model. As the figure shows, the images in the top-10 results are not only highly relevant to the query but also exhibit similarities among themselves. Our EKLSR model accurately ranks the images that match the query at the top, even among similar images. This is mainly due to our designed key local selection and reconstruction (KLSF and KLR) we designed, which effectively extract discriminative local cues to distinguish different pedestrians.

2 IMPORTANCE FACTORS DISTRIBUTION

To validate the universality of the interpretable importance factor distribution pattern proposed in Section 3.2, we have listed results on other datasets. In terms of images, as shown in Figure 3(a), multiple visualized images indicate that regions with higher importance factors correspond to the subject regions of the images, a pattern similar to that observed in Figure 3 of the paper. Additionally, in the context of the text, as illustrated in Figure 3(b) and similar to Figure 3 in the paper, the importance factors of content-rich words (adjectives, nouns, verbs, and adverbs) generally exceed those of function words (prepositions, conjunctions, etc.).

The results above demonstrate that the importance factors extracted by leveraging the robust prior knowledge of CLIP are interpretable and universally applicable across multiple datasets.

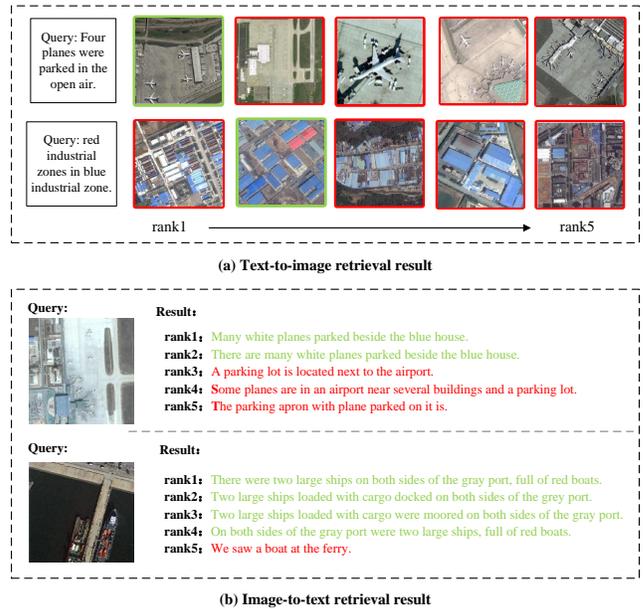


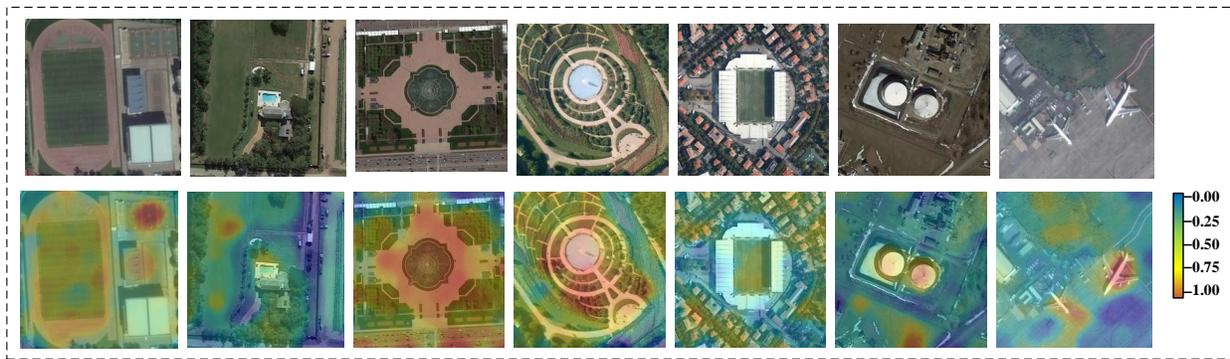
Figure 1: Visualization of RSITR retrieval results. The green result in the figure is the ground truth.



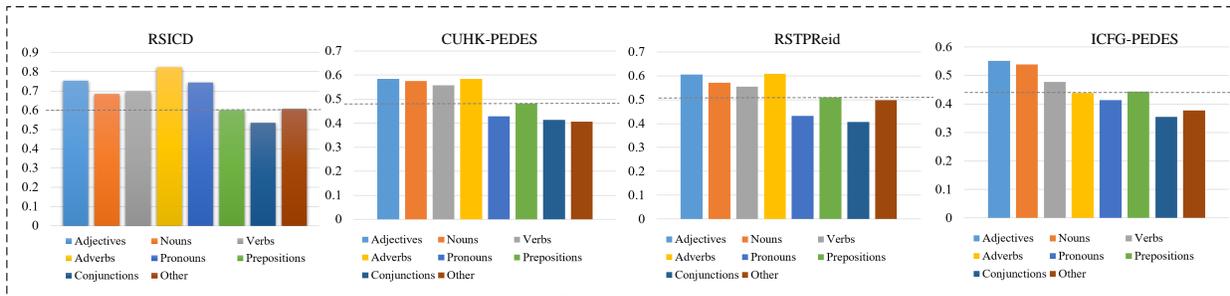
Figure 2: Visualization of TIReID retrieval results. The green result in the figure is the matched images.

3 FEATURE DISTRIBUTION ACROSS IMAGE REGIONS

Figure 1(d) in the paper illustrates that the key local features and non-key local features extracted by CLIP are intermixed, reflecting the lack of discriminability in CLIP's key local features. To address



(a) Image importance factors distribution



(b) Text importance factors distribution

Figure 3: Importance factors distribution.

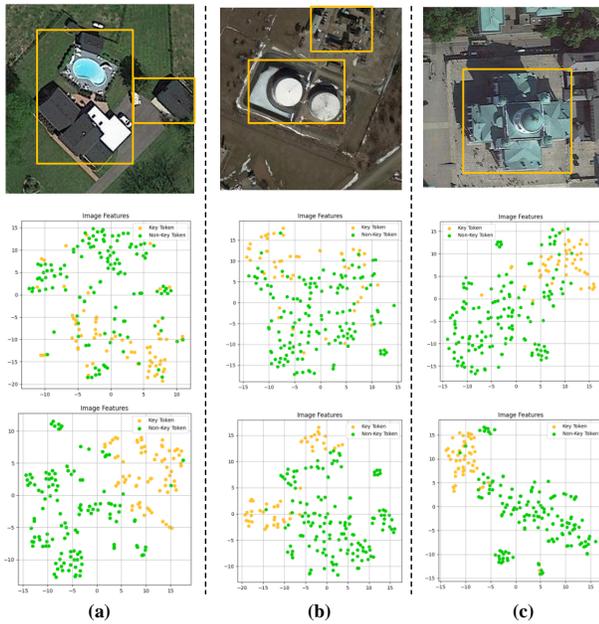


Figure 4: Visualization of image region features from CLIP and EKLSR.

this issue, we introduce the Key Local Segment Reconstruction

(KLR) based on multimodal interaction, which reconstructs the key local image regions to enhance their discriminative information. To demonstrate the effectiveness of the KLR strategy, we projected the local image features extracted by CLIP and EKLSR onto a 2D space, to visualize their feature distribution. As shown in Figure 4, (a)–(c) display three results, and each result from top to bottom is the source image, local image features from CLIP, and local image features from EKLSR. The yellow box highlights the key regions of the source image. Yellow dots and green dots respectively represent key and non-key local region features.

From the second row of Figure 4, it is visible that both key and non-key local features extracted by CLIP are mixed. Furthermore, as shown in the second row of Figure 4(a), some key local features in the top-left are isolated from the majority of key local features. These observations indicate that CLIP has poor capabilities in representing local features. To address this issue, our EKLSR model incorporates KLR. As shown in the third row of Figure 4, the key and non-key local features extracted by our EKLSR are completely separated. Moreover, the distribution of our key local features is relatively concentrated. These demonstrate the effectiveness of our KLR strategy, which can enhance the discriminability of local features.

4 SEMANTIC LOCALIZATION

In this section, we validate the feasibility of the EKLSR method for the Semantic Localization (SeLo) task. The semantic localization

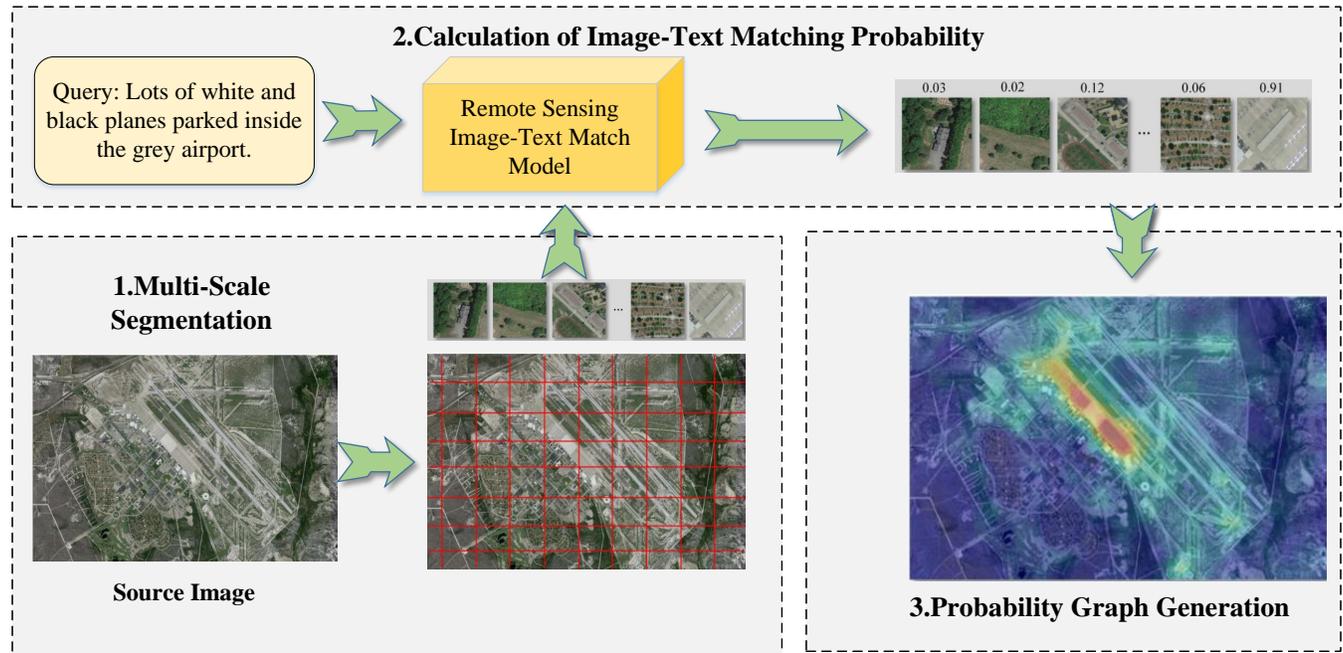


Figure 5: Framework of Semantic Localization.

task was introduced by Yuan et al. [6] in 2022. It involves a text-image matching model that identifies the region that best matches the text within a large remote sensing scene. The complex and varied characteristics of large-scale remote sensing pose a significant challenge to the model’s accurate representation of text and image features. Insufficient richness and discriminative power of features can lead to overdetections and missed detections, making it a more advanced and difficult task than remote sensing text-image retrieval. The semantic localization task is a challenging practical application that serves as an excellent test of the robustness and practical utility of the model discussed in this paper.

4.1 SeLo Method and Implementation Details

Figure 5 illustrates the semantic localization, which is divided into three steps:

1) Multi-Scale Segmentation: The large-scale remote sensing images are segmented using a multi-scale sliding window approach. The scales used for cropping in our experiment are 256×256 , 512×512 , and 768×768 .

2) Calculation of Image-Text Matching Probability: The probability distribution between the text and each image slice is calculated using a remote sensing image-text match model. For our experiment, the image-text matching model employed is the EKLSR model, which was trained on the RSITMD remote sensing image-text dataset.

3) Probability Graph Generation: The obtained probability distributions are merged and median filtering is applied to remove impulse noise from the probability graph, resulting in the final probability graph.

4.2 The Test Dataset for Semantic Localization

The images in the test set are sourced from Google Earth and comprise 22 large-scale remote sensing images with three channels. The dimensions of these images range from 3000×2000 to 10000×10000 pixels. Each query sentence corresponds to one or more image regions, totaling 59 annotated image regions.

4.3 Metrics

Following the [6], we use four metrics (Rsu , Ras , Rda , and Rmi) to measure the performance of the model in semantic localization. Rsu : Represents the ratio of the model’s attention within the labeled regions to its attention outside these regions. Ras : Quantifies the distance between the center of the labeled region and the k nearest model attention regions. Rda : Measures the concentration of the model’s attention regions. Rmi : It is used to assess the semantic localization task comprehensively. The calculation is as follows:

$$Rmi = wsu * Rsu + was * (1 - Ras) + wda * Rda \quad (1)$$

where wsu , was , and wda are the weight parameters, with values of 0.4, 0.35, and 0.25, respectively.

4.4 Quantitative Results

We employ our EKLSR model alongside several other methods (VSE++ [1], LW-MCR [7], SCAN [2], CAMP [4], AMFMN [5], and CLIP [3]) as a Remote Sensing Image-Text Match Model to conduct semantic localization experiments. Notably, all models were fine-tuned on the RSITMD dataset before being tested on the semantic localization dataset. The experimental results are shown in Table 1. As can be seen from the table, our EKLSR model achieves the best semantic localization result, achieving 73.23% in the average score

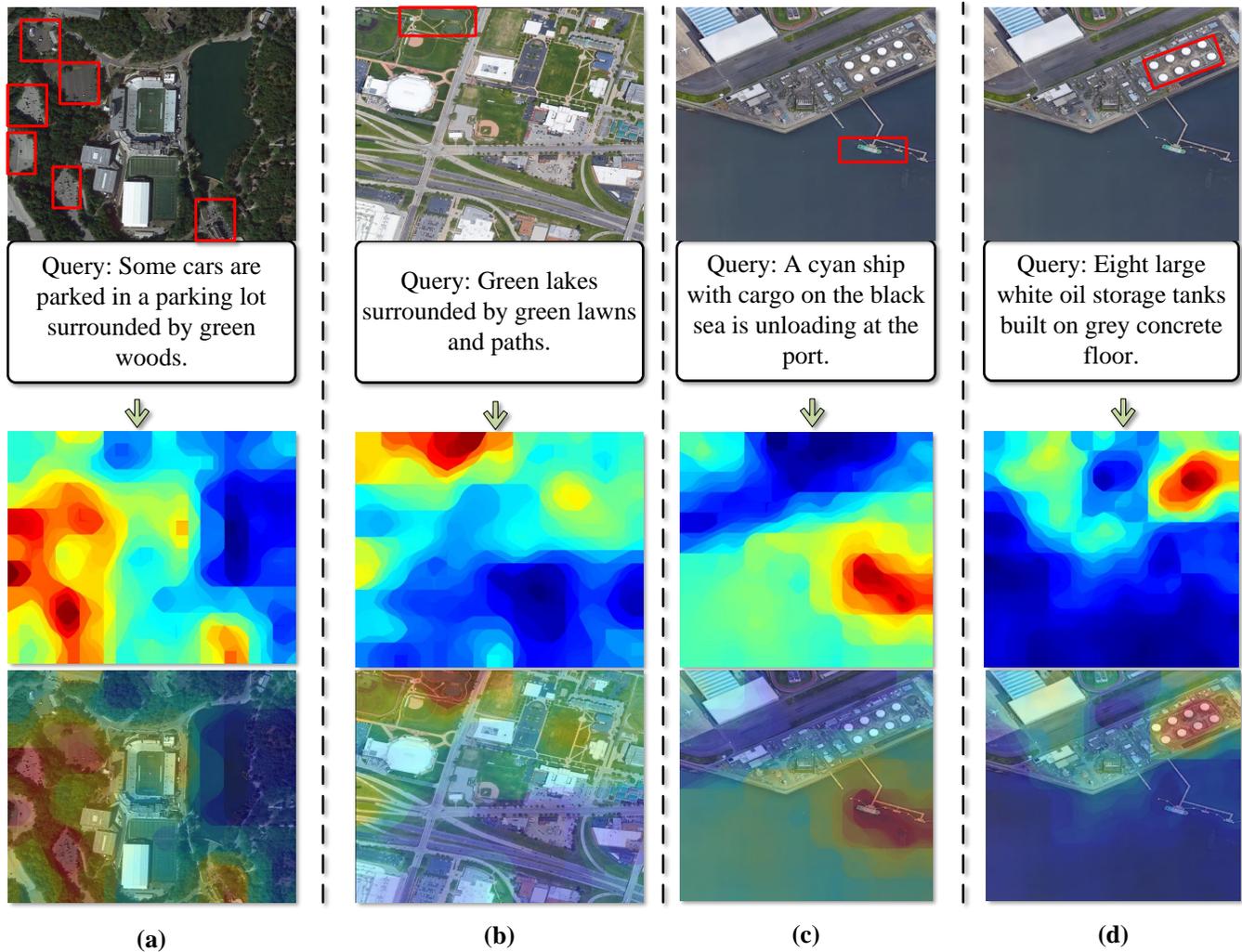


Figure 6: Visualization of semantic localization results. (a)–(d) Four semantic localization results. Each result from top to bottom is the source image with ground truth, query text, the corresponding probability map, and the located image.

Method	\uparrow Rsu	\uparrow Rda	\downarrow Ras	\uparrow Rmi
VSE++	63.64	58.29	41.66	60.45
LW-MCR	66.98	60.21	43.35	61.67
SCAN	64.21	61.32	38.71	62.47
CAMP	68.19	63.14	39.12	64.37
AMFMN	69.2	66.67	33.23	67.72
CLIP	75.36	66.23	26.89	72.29
EKLSR (ours)	75.28	70.10	26.63	73.23

Table 1: Experimental results on Semantic Localization task.

(Rmi), which is approximately 1% higher than that of the CLIP model. Moreover, our EKLSR model significantly outperforms other models. These indicate that EKLSR has a higher adaptability to specific domain image-text retrieval and is capable of excellent

image-text understanding even in complex, large-scale specific domain scenes.

4.5 Qualitative Results

We present several representative localization results, as shown in Figure 6. In Figure 6(a), we attempt to locate a "parking lot surrounded by green woods" within a large-scale scene image. Even if the generated probability map can locate the ground truth, part of the probability still falls elsewhere, indicating room for improvement in the model. In Figure 6(b), we successfully locate "green lakes" adjacent to green lawns and paths, with the model accurately pinpointing the lakes at the top end of the image. Figures 6(c) and 6(d) display the semantic localization results for two different queries in the same image, targeting objects of different scales such as "A cyan ship" and "Eight large white oil storage tanks". Our model achieves precise localization for both queries. These experiments

demonstrate that EKLSR can effectively localize remote sensing images.

Our EKLSR model achieved excellent semantic localization performance in both quantitative and qualitative experimental results. This demonstrates that the features extracted by our EKLSR model possess high informational richness and strong discriminative power, enabling outstanding remote sensing image-text understanding capabilities even in complex, large-scale remote sensing scenarios.

REFERENCES

- [1] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612* (2017).
- [2] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*. 201–216.
- [3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [4] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. 2019. Camp: Cross-modal adaptive message passing for text-image retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5764–5773.
- [5] Zhiqiang Yuan, Wenkai Zhang, Kun Fu, Xuan Li, Chubo Deng, Hongqi Wang, and Xian Sun. 2022. Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval. *arXiv preprint arXiv:2204.09868* (2022).
- [6] Zhiqiang Yuan, Wenkai Zhang, Chongyang Li, Zhaoying Pan, Yongqiang Mao, Jialiang Chen, Shuoke Li, Hongqi Wang, and Xian Sun. 2022. Learning to Evaluate Performance of Multi-modal Semantic Localization. *IEEE Transactions on Geoscience and Remote Sensing* (Jan 2022), 1–18. <https://doi.org/10.1109/TGRS.2022.3207171>
- [7] Zhiqiang Yuan, Wenkai Zhang, Xuee Rong, Xuan Li, Jialiang Chen, Hongqi Wang, Kun Fu, and Xian Sun. 2021. A lightweight multi-scale crossmodal text-image retrieval method in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021), 1–19.

465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522

523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580