

Pushing the Limits of Vision-Language Models in Remote Sensing without Human Annotations

Keumgang Cha¹, Donggeun Yu¹, Junghoon Seo², Hyunguk Choi¹, Taegyun Jeon¹

¹ SI Analytics, Daejeon 34051, South Korea - (chagmgang, donggeun, hyunguk, tgjeon)@si-analytics.ai

² KAIST, Daejeon 34141, South Korea - (jhseo)@kaist.ac.kr

Keywords: Remote Sensing, Foundation Model, Multi Modality, Vision-Language.

Abstract

The prominence of generalized foundation models in vision-language integration has witnessed a surge, given their multifarious applications. Within the natural domain, the procurement of vision-language datasets to construct these foundation models is facilitated by their abundant availability and the ease of web crawling. Conversely, in the remote sensing domain, although vision-language datasets exist, their volume is suboptimal for constructing robust foundation models. This study introduces an approach to curate vision-language datasets by employing an image decoding machine learning model, negating the need for human-annotated labels. Utilizing this methodology, we amassed approximately 9.6 million vision-language paired datasets in VHR imagery. The resultant model outperformed counterparts that did not leverage publicly available vision-language datasets, particularly in downstream tasks such as zero-shot classification, semantic localization, and image-text retrieval. Moreover, in tasks exclusively employing vision encoders, such as linear probing and k-NN classification, our model demonstrated superior efficacy compared to those relying on domain-specific vision-language datasets.

1. Introduction

Foundation models are at the forefront of breakthrough in the deep learning community. Unlike specialized models that demand new labeling and training for different target tasks, foundation models boast of a flexible architecture that can efficiently span diverse tasks. This includes zero-shot classification, semantic localization, and even cross-modal retrieval. In the world of computer vision, seminal contributions like DINO(Caron et al., 2021) and SAM(Kirillov et al., 2023) have carved a niche. Concurrently, the natural language processing domain has been revolutionized by models such as BERT(Devlin et al., 2019), GPT3(Brown et al., 2020), and PaLM(Chowdhery et al., 2023). Further amalgamating vision and language has led to transformative works such as Flamingo(Alayrac et al., 2022), Instruct-BLIP(Dai et al., 2023), and BEiT-3(Wang et al., 2023).

The remote sensing community, recognizing the potential of these models, has increasingly incorporated foundation models into its fold. Several works, prominently involving the Masked Image Modeling (MIM) approach, have made significant strides in tasks specific to this domain(Wang et al., 2022). However, these models often encounter hurdles. A persistent challenge lies in their reliance on supervised fine-tuning, especially when deployed for core computer vision tasks.

Addressing these challenges has led to an intensified focus on vision-language foundation models within the remote sensing community. Specifically, the principles of contrastive learning between vision and language, exemplified by models like CLIP(Radford et al., 2021), have gained traction. The allure of these models is their ability to adeptly manage a gamut of applications, often bypassing the tedious fine-tuning phase.

The bedrock of successful foundation models invariably remains quality datasets. Within the remote sensing context, although datasets like RSICD(Lu et al., 2017) and UCM(Qu et al., 2016) exist, they often pale in comparison to voluminous datasets from

more natural domains, such as LAION-5B(Schuhmann et al., 2022). Methods to bridge this gap have been devised. For instance, RS5M(Zhang et al., 2023) employed the BLIP-2(Li et al., 2023) model to curate vision-language pairs, while Remote-CLIP(Liu et al., 2023) aimed to convert traditional datasets into the vision-language format. By encoding location information instead of language and using it in clips, SatCLIP(Klemmer et al., 2023) achieved significant performance in various tasks.

In this context, contribution of this paper is twofold: Firstly, we delineate a methodology to create a robust vision-language dataset tailored specifically for the remote sensing domain. By leveraging the potential of InstructBLIP(Dai et al., 2023), we strive to ensure linguistic diversity and quality, sourcing images exclusively from esteemed remote sensing repositories. Secondly, building upon our crafted dataset, we introduce RSCLIP. Trained within the well-established CLIP framework(Radford et al., 2021), RSCLIP promises to bridge the performance gap, outdoing models trained on synthetic labels and standing toe-to-toe with those reliant on human-annotated labels.

2. Proposed Method

2.1 Generation of Large-Scale Vision-Language Datasets

The InstructBLIP(Dai et al., 2023) is utilized to extract vision-language pairs from individual images. Since InstructBLIP is tailored to echo the user's intent in generating captions, two distinct captions are produced for each image in this study. To guide the description of each image, the prompts "Write a short description for the image." and "Describe the image in detail" are provided, aiming to yield both concise and extended captions, respectively.

The source datasets employed to generate the vision-language pairs include fMoW(Christie et al., 2018), Million-AID(Long et al., 2021), DFC2019(Lian et al., 2020), DFC2021(Malkin et al.,

Model	Params	RSICD							RSITMD						
		Image-to-Text			Text-to-Image				Image-to-Text			Text-to-Image			
		$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	mR	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	mR
CLIP(ViT-B-32)(Zhang et al., 2023)	$\approx 151M$	5.4	15	24.06	6.44	19.82	30.28	16.83	9.51	23.01	32.74	8.81	27.92	43.23	24.20
CLIP(ViT-L-14)(Zhang et al., 2023)	$\approx 427M$	-	-	-	-	-	-	-	12.61	29.87	42.48	15.17	39.2	52.92	32.04
CLIP(ViT-H-14)(Zhang et al., 2023)	$\approx 986M$	-	-	-	-	-	-	-	12.61	33.41	44.69	14.2	39.47	55.27	33.28
CLIP(ViT-bigG-14)(Zhang et al., 2023)	$\approx 2500M$	-	-	-	-	-	-	-	13.94	34.51	45.13	13.98	41.59	56.59	34.29
VSE++(Faghri et al., 2018)	-	3.38	9.51	17.46	2.82	11.32	18.1	10.43	10.38	27.65	39.6	7.79	24.87	38.67	24.83
AFMFN(Yuan et al., 2022a)	-	5.39	15.08	23.4	4.9	18.28	31.44	16.42	11.06	29.2	38.72	9.96	34.03	52.96	29.32
KCR(Mi et al., 2022)	-	5.84	22.31	36.12	4.76	18.59	27.2	19.14	-	-	-	-	-	-	-
GalR(Yuan et al., 2022c)	-	6.59	19.85	31.04	4.69	19.48	32.13	18.96	14.82	31.64	42.48	11.15	36.68	51.68	31.41
Pfeiffer(Zhang et al., 2023)	$\approx 152M$	7.87	18.21	27.26	5.84	20.57	33.14	18.82	11.5	25	36.28	9.65	31.59	46.9	26.82
PrefixTuning(Zhang et al., 2023)	$\approx 152M$	9.61	22.05	32.11	6.99	22.09	33.06	20.99	13.72	30.97	43.14	6.25	30.04	47.26	28.56
LoRA(Zhang et al., 2023)	$\approx 152M$	7.14	18.48	27.17	6.18	19.05	29.66	17.95	13.5	28.98	39.38	6.86	26.55	40.53	25.97
UniPELT(Zhang et al., 2023)	$\approx 152M$	8.87	21.04	31.29	6.81	24.01	35.75	21.30	13.27	29.2	41.37	9.69	32.57	48.36	29.08
RSCLIP	$\approx 197M$	10.43	25.34	39.34	9.9	30.52	45.03	26.76	19.25	36.06	46.68	12.92	42.04	63.14	36.68

Table 1. Image-text retrieval in both RSICD and RSITMD dataset. As main experiment results, the models included RS5M is used as comparison. The RSCLIP shows the highest performance in all data sets and all top-k except for R@1 in RSITMD.

Model	Params	Zero-shot Classification				Semantic Localization			
		AID	RESISC45	Avg		AIR-SLT			
		Top-1 Accuracy			$R_{eu} \uparrow$	$R_{as} \downarrow$	$R_{du} \uparrow$	$R_{mi} \uparrow$	
CLIP(ViT-B-32)(Zhang et al., 2023)	151M	60.84	58.97	59.91	0.7220	0.2848	0.6880	0.7111	
SeLov1(Yuan et al., 2022b)	-	-	-	-	0.6920	0.3323	0.6667	0.6772	
SeLov2(Yu et al., 2023)	-	-	-	-	0.7199	0.2925	0.6658	0.7021	
Pfeiffer(Zhang et al., 2023)	152M	68.37	67.79	68.08	0.7180	0.3116	0.6589	0.6912	
PrefixTuning(Zhang et al., 2023)	152M	69.83	66.74	68.29	0.7241	0.3132	0.6867	0.7017	
LoRA(Zhang et al., 2023)	152M	67.38	65.53	66.46	0.7176	0.2857	0.6911	0.7098	
UniPELT(Zhang et al., 2023)	152M	70.92	66.61	68.77	0.7292	0.3463	0.6461	0.6820	
RSCLIP	192M	75.82	68.59	72.20	0.7349	0.2877	0.7070	0.7200	

Table 2. The zero-shot classification and semantic localization results. In zero-shot classification, the RSCLIP has the best performance as shown in table. In semantic localization, the RSCLIP records the best performance except for R_{as} .

metrics follow the original research's hyperparameters(Yuan et al., 2022b). Although the RSCLIP shows the superior performance except for R_{as} , the RSCLIP's R_{mi} , which is their comprehensive indicator, records the best performance.

3.2 Additional Experiment Results

Distinct from the approach in this paper, we also contrasted RSCLIP with models like S-CLIP and RemoteCLIP, which directly utilize vision-language pairs. S-CLIP employs a semi-supervised technique, capitalizing on only 10% of existing vision-language pairs. However, because its text encoder was informed directly by the vision-language pair, it's classified as an additional experiment. Similarly, RemoteCLIP, which learned all vision-language pairs directly, was also placed in this category.

Generally, RSCLIP doesn't top the charts in downstream tasks. This is expected as other models benefit from text encoders directly trained on downstream language distributions. Yet, RSCLIP remains competitive even without this advantage. Impressively, in tasks like few-shot, linear probing, and k-NN Classification, RSCLIP reigns supreme using only a vision encoder. For clarity, in the Additional Experiment Results section, models directly leveraging vision-language pairs are marked with \diamond , while those that didn't utilize them at all bear the \blacklozenge symbol. Detailed results follow below.

3.2.1 Image-Text Retrieval For evaluation metric in retrieval, the retrieval recall of top-1 ($R@1$), and top-5 ($R@5$) are reported. Table 3 displays image-text retrieval results. Expectedly, RemoteCLIP, trained on the most direct vision-language pairs, outshines the rest. Still, when compared to S-CLIP, RSCLIP displays superior performance even without the direct 10% vision-language advantage. This indicates the potential of our vision-language pair generation method.

3.2.2 Zero Shot Classification Table 4 presents the top-1 accuracy for zero-shot classification across multiple datasets.

For this evaluation, we utilized ten downstream datasets, including RSICD-CLS, UCMerced Land Use (UCM-CLS)(Yang and Newsam, 2010), WHU-RS19(Xia et al., 2009), AID(Xia et al., 2017), RESISC45(Cheng et al., 2017), EuroSAT(Helber et al., 2019), RSI-CB128(Li et al., 2020a), RSI-CB256(Li et al., 2020a), MLRSNet(Qi et al., 2020), and PatternNet(Zhou et al., 2018). Within the table, "Avg 1" represents the average performance across RSICD-CLS, UCM-CLS, WHU-RS19, and AID datasets and serves as a comparison with S-CLIP. "Avg 2" calculates the average for datasets WHU-RS19, AID, RESISC45, EuroSAT, RSI-CB128, RSI-CB256, MLRSNet, and PatternNet, intended for comparison with RemoteCLIP.

Regarding "Avg 1", RSCLIP, despite not immediately employing language, displays accuracy surpassing the ResNet-50 variant of S-CLIP, yet falling short of its ViT-Base counterpart. In the "Avg 2" category, RSCLIP doesn't top the charts for WHU-RS19, AID, and RESISC45. However, it excels in RSI-CB128, RSI-CB256, MLRSNet, and PatternNet. Moreover, in terms of average performance, RSCLIP achieves the highest score. Collectively, while it seems optimal to directly incorporate vision-language from the downstream task, our method of constructing a vision-language pair yields comparable results.

3.2.3 Few-shot Classification Few-shot classification evaluates the standalone vision encoder. Datasets are split into training and testing sets at a ratio of 0.8 to 0.2. Depending on the settings, images from the training set are extracted per class based on the designated number of shots. These extracted images provide representations for shots, serving as training features for the linear probing model. Upon training this model, test images are transformed into representations via the vision encoder, then input into the trained model to predict image classes. For this experiment, datasets RSI-CB128, RSI-CB256, EuroSAT, MLRSNet, PatternNet, RESISC45, AID, and WHU-RS19 were employed. Shot numbers for few-shot classification were set at 1, 4, 8, 16, and 32, with the logistic regression model from scikit-learn functioning as the linear probing model. Table 5 indicates that, despite RSCLIP not using direct vision-language pairs from the downstream task dataset, it surpasses RemoteCLIP across all few-shot settings, even in average accuracy only except for 1-shot classification in RESISC45. Two potential reasons underpin this outcome. Firstly, only the vision encoder is deployed in few-shot classification. Secondly, RSCLIP's pretraining phase utilized a significantly larger image corpus than RemoteCLIP.

3.2.4 Full-shot Linear Probing and k-NN Classification Full-shot linear probing can be viewed as an extension of the few-shot classification. Unlike its few-shot counterpart where a limited number of images serve as input features for the linear

Model	Params	RSICD				RSITMD				UCM				Sydney			
		Image-to-Text	Text-to-Image	Image-to-Text	Text-to-Image	Image-to-Text	Text-to-Image	Image-to-Text	Text-to-Image	Image-to-Text	Text-to-Image	Image-to-Text	Text-to-Image	Image-to-Text	Text-to-Image	Image-to-Text	Text-to-Image
S-CLIP($L = U$) (Mo et al., 2023)	$\approx 102M$	4.2	18.4	4.2	16.8	-	-	-	-	11.6	45.7	11.1	43.5	14.9	50	17.8	55.1
S-CLIP($L \neq U$) (Mo et al., 2023)	$\approx 102M$	4.2	17.1	3.9	15.8	-	-	-	-	9.8	43.5	10.8	42.5	13.8	48.9	17.8	52.3
RemoteCLIP (Liu et al., 2023)	$\approx 102M$	13.36	32.94	10.76	32.83	23.67	47.57	19.29	51.55	13.33	50.48	15.24	57.14	-	-	-	-
RemoteCLIP (Liu et al., 2023)	$\approx 151M$	17.02	37.97	13.71	37.11	27.88	50.66	22.17	56.46	20.48	59.85	18.67	61.52	-	-	-	-
RemoteCLIP (Liu et al., 2023)	$\approx 428M$	18.39	37.42	14.73	39.93	28.76	52.43	23.76	59.51	19.05	54.29	17.71	62.19	-	-	-	-
RSCLIP	$\approx 197M$	10.43	25.34	9.9	30.52	19.25	36.06	12.92	42.04	19.05	56.19	16.38	62.29	29.31	58.62	22.07	57.93

Table 3. The additional evaluation results of image-text retrieval in RSICD, RSITMD, UCM and Sydney dataset. In this experiment, although the RSCLIP is not trained with vision-language pairs presented in the downstream tasks, it can be seen in table that the RSCLIP shows the performance that is just as good as the model using it.

RSICD-CLS UCM-CLS WHU-RS19 AID RESISC45 EuroSAT RSI-CB128 RSI-CB256 MLRSNet PatternNet													
Method	Params	Top-1 Accuracy										Avg 1	Avg 2
S-CLIP(ResNet-50)◊(Mo et al., 2023)	≈ 102M	66.90	66.70	86.90	73.00	-	-	-	-	-	-	73.38	
S-CLIP(ViT-Base)◊(Mo et al., 2023)	≈ 151M	87.40	88.90	97.30	93.10	-	-	-	-	-	-	91.67	
RemoteCLIP(ResNet-50)◊(Liu et al., 2023)	≈ 102M	-	-	95.15	86.55	53.24	17.19	13.95	33.03	40.68	45.51	- 48.16	
RemoteCLIP(ViT-Base)◊(Liu et al., 2023)	≈ 151M	-	-	96.12	91.30	70.33	35.96	24.18	39.50	59.28	57.71	- 59.30	
RSCLIP♦	≈ 197M	69.33	68.33	86.67	75.82	68.59	48.44	30.59	47.19	65.12	66.74	75.04 61.14	

Table 4. The zero-shot classification with text prompt, which is "the satellite image of class name". The RSCLIP shows the competitive performance without using the vision-language pairs of the downstream tasks.

probing model, full-shot classification utilizes all training-split images for this purpose. For k-NN classification, the k parameter for nearest neighbors is consistently set to 20, aligning with RemoteCLIP's approach (Liu et al., 2023). FAISS (Johnson et al., 2019) underpins the k-NN algorithm. Datasets RSI-CB128, RSI-CB256, EuroSAT, MLRSNet, PatternNet, RESISC45, AID, and WHU-RS19 were harnessed as benchmark datasets for these evaluations. Except for four cases, the table 6 reveals that the RSCLIP consistently outperforms other models in both full-shot linear probing and k-NN classification across all datasets. The four cases includes both the linear probing of EuroSAT, RESISC45 and k-NN classification in RSI-CB128, RSI-CB256. However, its average performance also stands unmatched. These outcomes might stem from reasons similar to those discussed in the few-shot classification section.

4. Conclusion

This paper demonstrates the potential of leveraging large language models for image decoding to construct vision-language models without the need for human-annotated labels. We introduced a vision-language foundational model, RSCLIP, built using a straightforward image-text contrastive learning approach with our proposed dataset. To assess the efficacy of this foundational model, we conducted primary downstream tasks including zero-shot classification, image-text retrieval, and semantic localization. When comparing RSCLIP to models not trained on the distribution of direct language descriptions, RSCLIP consistently outperformed its counterparts. Even though RSCLIP might not always surpass models trained directly with language descriptions, its performance remains highly competitive. Looking ahead, our future endeavors will explore the integration of various modalities present in remote sensing imagery, expressed in the form of language.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M. et al., 2022. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 35, 23716–23736.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A.

et al., 2020. Language models are few-shot learners. *NeurIPS*, 33, 1877–1901.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers. *ICCV*, 9650–9660.

Cheng, G., Han, J., Lu, X., 2017. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10), 1865–1883.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S. et al., 2023. Palm: Scaling language modeling with pathways. *JMLR*.

Christie, G., Fendley, N., Wilson, J., Mukherjee, R., 2018. Functional map of the world. *CVPR*, 6172–6180.

Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S. C. H., 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *arXiv preprint arXiv:2305.06500*.

Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., Raskar, R., 2018. Deepglobe 2018: A challenge to parse the earth through satellite images. *CVPR Workshops*.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*.

Faghri, F., Fleet, D. J., Kiros, J. R., Fidler, S., 2018. Vse++: Improving visual-semantic embeddings with hard negatives. *BMVC*.

Helber, P., Bischke, B., Dengel, A., Borth, D., 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE J-STARS*, 12(7), 2217–2226.

Johnson, J., Douze, M., Jégou, H., 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3), 535–547.

Method	Backbone	Shot	RSI-CB128	RSI-CB256	EuroSAT	MLRSNet	PatternNet	RESISC45	AID	WHU-RS19	Avg
RemoteCLIP◇	ResNet-50	1	35.59	42.52	43.20	31.75	46.10	39.33	36.95	45.15	40.07
RemoteCLIP◇	ViT-Base		34.31	44.28	44.89	34.14	45.98	42.10	37.04	40.78	40.44
RSCLIP◆	ViT-Base		60.65	83.28	54.13	78.44	82.38	37.37	97.62	100.00	74.23
RemoteCLIP◇	ResNet-50	4	60.04	65.44	55.53	46.90	66.99	52.11	63.13	73.59	60.47
RemoteCLIP◇	ViT-Base		64.49	70.33	55.99	54.52	70.98	60.91	65.59	68.16	63.87
RSCLIP◆	ViT-Base		80.65	88.66	73.53	96.00	98.25	78.00	99.52	100.00	89.33
RemoteCLIP◇	ResNet-50	8	69.55	75.89	61.75	55.02	77.07	61.75	70.50	85.44	69.62
RemoteCLIP◇	ViT-Base		76.13	83.73	65.76	64.24	82.53	70.92	75.72	80.68	74.96
RSCLIP◆	ViT-Base		89.35	96.72	75.50	95.51	99.00	88.71	98.57	100.00	92.92
RemoteCLIP◇	ResNet-50	16	77.58	83.72	70.36	59.74	82.93	69.51	75.12	89.32	76.04
RemoteCLIP◇	ViT-Base		82.63	89.12	75.73	67.45	88.13	75.83	81.05	89.51	81.18
RSCLIP◆	ViT-Base		94.84	98.21	94.10	96.34	99.00	88.29	99.05	100.00	96.23
RemoteCLIP◇	ResNet-50	32	82.02	87.04	77.44	64.99	88.32	75.71	82.46	93.79	81.47
RemoteCLIP◇	ViT-Base		88.11	91.83	83.30	71.58	91.87	81.77	86.67	93.40	86.07
RSCLIP◆	ViT-Base		96.77	99.10	95.60	97.12	99.63	88.43	98.81	100.00	96.93

Table 5. The few-shot classification results in additional experiment. The RSCLIP is compared with the RemoteCLIP in various scene classification dataset. In all datasets and all k-shot settings, the RSCLIP is the best performance with the same reason of full linear probing and k-NN classification.

		RSI-CB128		RSI-CB256		EuroSAT		MLRSNet		PatternNet		RESISC45		AID		WHU-RS19		Avg	
Method	Backbone	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	91.48	88.41	78.98	74.78	96.18	93.45	86.16	83.60	83.00	79.45	95.63	90.21	90.63	87.57
SwAV◆	ResNet-50	95.27	95.61	98.29	98.17	91.17	91.37	79.04	76.12	96.94	94.18	88.60	85.59	86.00	80.80	96.12	92.23	91.43	89.26
Barlow Twins◆	ResNet-50	98.07	95.91	99.03	98.13	94.78	91.57	82.41	77.55	97.73	93.83	91.10	86.10	88.25	81.75	97.09	91.75	93.56	89.57
VICReg◆	ResNet-50	97.47	96.03	98.67	98.21	95.06	91.44	82.59	78.02	98.83	94.03	91.03	86.75	88.10	81.50	96.60	90.78	93.54	89.60
CLIP◆	ResNet-50	94.89	97.05	97.30	97.24	91.67	88.54	80.08	77.14	95.61	92.86	85.73	85.65	90.95	86.90	97.57	93.69	91.73	89.88
CLIP-CL◇	ResNet-50	95.99	94.92	98.41	98.09	89.80	87.65	79.32	76.99	97.30	95.15	89.10	88.19	94.80	92.85	98.06	97.57	92.85	91.43
ImageNet◆	ViT-Base	96.45	91.29	98.11	97.00	85.57	76.56	78.61	74.05	96.81	92.98	86.89	81.63	83.55	76.45	94.17	89.81	90.02	84.97
ViTAE◆	ViT-Base	93.10	95.65	98.41	94.05	61.41	82.27	91.15	80.37	98.50	90.82	87.94	65.33	88.30	64.05	91.74	70.39	88.82	80.37
CLIP◆	ViT-Base	97.36	94.17	98.55	97.40	95.15	90.28	85.43	82.26	97.58	94.36	92.60	89.73	94.95	90.35	97.09	93.69	94.84	91.53
RemoteCLIP◇	ResNet-50	96.06	94.78	98.39	97.62	92.56	90.20	83.32	81.21	97.37	95.95	90.94	90.05	94.35	92.10	98.06	95.63	93.88	92.19
RemoteCLIP◇	ViT-Base	98.02	95.82	99.01	98.51	96.19	93.50	87.00	85.11	98.47	97.32	94.27	92.67	95.95	92.55	97.57	74.17	95.81	91.21
RSCLIP◆	ViT-Base	98.13	96.70	99.09	98.02	95.50	94.33	94.01	93.36	99.08	98.60	94.14	93.64	97.95	97.65	99.50	98.01	97.18	96.29

Table 6. The full linear probing and k-NN classification in additional experiment. As mentioned, the ◇ is the model trained with direct vision-language pair of downstream tasks and the ◆ is the model not using the direct language expression of downstream tasks. Although the RSCLIP is marked as ◆, the RSCLIP scores the best performance in all dataset because this downstream tasks require only vision encoder.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y. et al., 2023. Segment anything. *ICCV*.

Klemmer, K., Rolf, E., Robinson, C., Mackey, L., Rußwurm, M., 2023. Satclip: Global, general-purpose location embeddings with satellite imagery. *arXiv preprint arXiv:2311.17179*.

Li, H., Dou, X., Tao, C., Wu, Z., Chen, J., Peng, J., Deng, M., Zhao, L., 2020a. RSI-CB: A large-scale remote sensing image classification benchmark using crowdsourced data. *Sensors*, 20(6), 1594.

Li, J., Li, D., Savarese, S., Hoi, S., 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ICML*.

Li, K., Wan, G., Cheng, G., Meng, L., Han, J., 2020b. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159, 296–307.

Lian, Y., Feng, T., Zhou, J., Jia, M., Li, A., Wu, Z., Jiao, L., Brown, M., Hager, G., Yokoya, N. et al., 2020. Large-scale semantic 3-D reconstruction: Outcome of the 2019 IEEE GRSS data fusion contest—Part B. *IEEE J-STARS*, 14, 1158–1170.

Liu, F., Chen, D., Guan, Z., Zhou, X., Zhu, J., Zhou, J., 2023. RemoteCLIP: A Vision Language Foundation Model for Remote Sensing. *arXiv preprint arXiv:2306.11029*.

Liu, Z., Yuan, L., Weng, L., Yang, Y., 2017. A high resolution optical satellite image dataset for ship recognition and some new baselines. *ICPRAM*, 2, 324–331.

Long, Y., Xia, G.-S., Li, S., Yang, W., Yang, M. Y., Zhu, X. X., Zhang, L., Li, D., 2021. On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid. *IEEE J-STARS*, 14, 4205–4230.

Lu, X., Wang, B., Zheng, X., Li, X., 2017. Exploring models and data for remote sensing image caption generation. *IEEE TGRS*, 56(4), 2183–2195.

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. *IGARSS*, IEEE, 3226–3229.

Malkin, N., Robinson, C., Jojic, N., 2021. High-resolution land cover change from low-resolution labels: Simple baselines for the 2021 ieee grss data fusion contest. *arXiv preprint arXiv:2101.01154*.

Mi, L., Li, S., Chappuis, C., Tuia, D., 2022. Knowledge-aware cross-modal text-image retrieval for remote sensing images. *IJCAI-ECAI Workshop*.

Mo, S., Kim, M., Lee, K., Shin, J., 2023. S-clip: Semi-supervised vision-language pre-training using few specialist captions. *NeurIPS*.

Oord, A. v. d., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Qi, X., Zhu, P., Wang, Y., Zhang, L., Peng, J., Wu, M., Chen, J., Zhao, X., Zang, N., Mathiopoulos, P. T., 2020. MLRSNet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169, 337–350.

Qu, B., Li, X., Tao, D., Lu, X., 2016. Deep semantic understanding of high resolution remote sensing image. *IEEE CITS*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al., 2021. Learning transferable visual models from natural language supervision. *ICML*, 8748–8763.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M. et al., 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 35, 25278–25294.

Wang, D., Zhang, Q., Xu, Y., Zhang, J., Du, B., Tao, D., Zhang, L., 2022. Advancing plain vision transformer toward remote sensing foundation model. *IEEE TGRS*, 61, 1–15.

Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O. K., Singhal, S., Som, S. et al., 2023. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *CVPR*.

Xia, G.-S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., Lu, X., 2017. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE TGRS*, 55(7), 3965–3981.

Xia, G.-S., Yang, W., Delon, J., Gousseau, Y., Sun, H., Maître, H., 2009. Structural high-resolution satellite image indexing. *ISPRS TC VII Symposium*.

Yang, Y., Newsam, S., 2010. Bag-of-visual-words and spatial extensions for land-use classification. *SIGSPATIAL*, 270–279.

Yu, M., Yuan, H., Chen, J., Hao, C., Wang, Z., Yuan, Z., Lu, B., 2023. SeLo v2: Towards for Higher and Faster Semantic Localization. *IEEE GRSL*.

Yuan, Z., Zhang, W., Fu, K., Li, X., Deng, C., Wang, H., Sun, X., 2022a. Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval. *IEEE TGRS*.

Yuan, Z., Zhang, W., Li, C., Pan, Z., Mao, Y., Chen, J., Li, S., Wang, H., Sun, X., 2022b. Learning to evaluate performance of multi-modal semantic localization. *IEEE TGRS*.

Yuan, Z., Zhang, W., Tian, C., Rong, X., Zhang, Z., Wang, H., Fu, K., Sun, X., 2022c. Remote sensing cross-modal text-image retrieval based on global and local information. *IEEE TGRS*, 60, 1–16.

Zhang, Z., Zhao, T., Guo, Y., Yin, J., 2023. RS5M: A Large Scale Vision-Language Dataset for Remote Sensing Vision-Language Foundation Model. *arXiv preprint arXiv:2306.11300*.

Zhou, W., Newsam, S., Li, C., Shao, Z., 2018. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS journal of photogrammetry and remote sensing*, 145, 197–209.