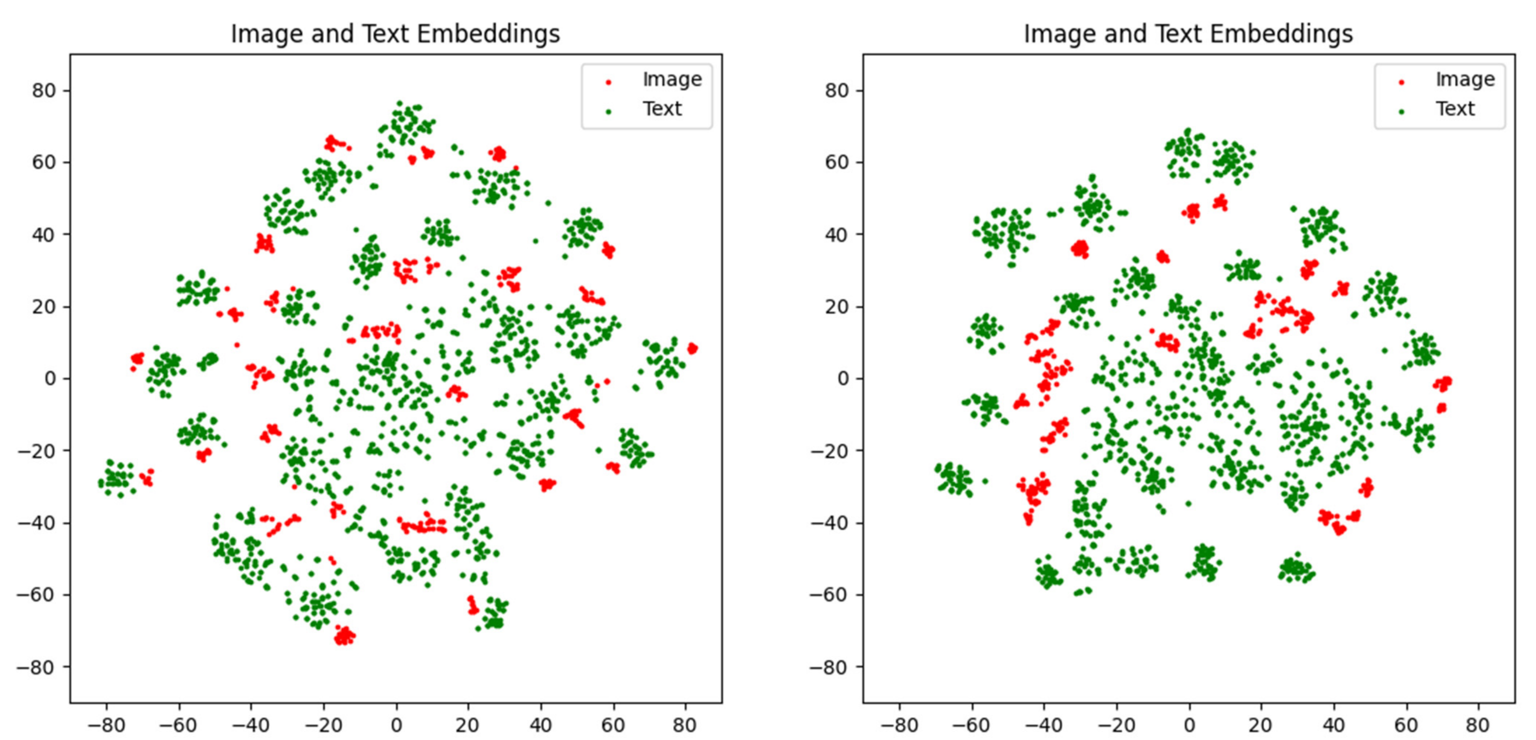
Efficient Remote Sensing with Harmonized Transfer Learning and Modality Alignment

Tengjun Huang

Shandong University

Motivation

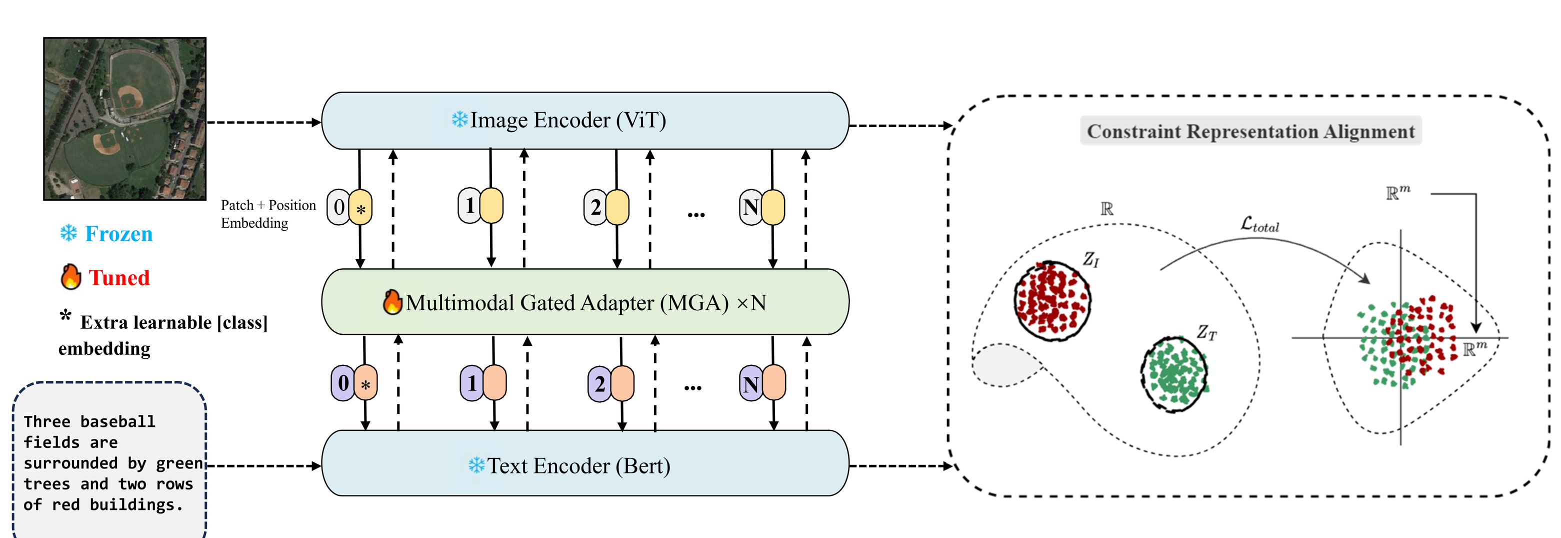


(A) Retrieving a better performing model (B) Retrieving a worse performing model

We have observed that poorly performing models sometimes exhibit a clustering phenomenon within the same modality embedding. Figure 1 illustrates the visualization of the last layer embeddings for two models with differing performance in the filed of remote sensing image-text retrieval; the clustering phenomenon is noticeably more pronounced in the right image than in the left. We hypothesize that this may be attributed to the high intra-class and inter-class similarity of remote sensing images, leading to semantic confusion when modeling a low-rank visual-language joint space. This raises a critical question: ***"How can we model a highly aligned visual-language joint space while ensuring efficient transfer learning?"***

Overall Method

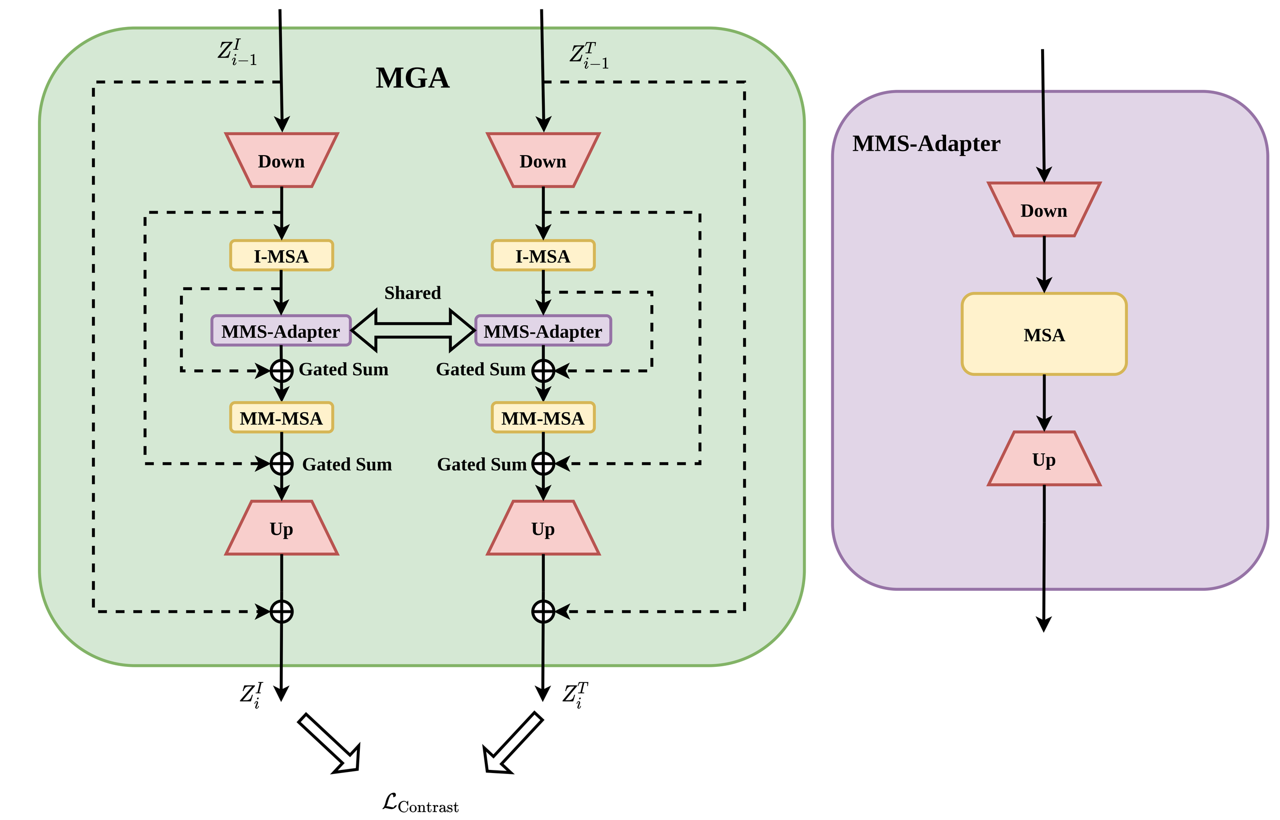
In the brains of congenitally blind individuals, parts of the visual cortex can take on the function of language processing. Concurrently, in the typical human cortex, several small regions—such as the Angular Gyrus and the Visual Word Form Area (VWFA)—serve as hubs for integrated visual-language processing. These areas hierarchically manage both low-level and high-level stimuli information. Inspired by this natural phenomenon, we propose "Efficient Remote Sensing with **Har**monized Transfer Learning and **M**odality **A**lignment (HarMA)".



Specifically, similar to the information processing methods of the human brain, we designed a hierarchical multimodal adapter with mini-adapters. This framework emulates the human brain's strategy of utilizing shared mini-regions to process neural impulses originating from both visual and linguistic stimuli. It models the visual-language semantic space from low to high levels by hierarchically sharing multiple mini-adapters. Finally, we introduced a new objective function to alleviate the severe clustering of features within the same modality. Thanks to its simplicity, the method can be easily integrated into almost all existing multimodal frameworks.

MultiModal Gated Adapter (MGA)

Previous methods use simple shared-weight modal interaction, causing potential semantic matching confusion. We designed a cross-modal adapter with adaptive gating:



The MMS-Adapter aligns multimodal representations via shared-weight self-attention. However, directly outputting these can hurt retrieval performance due to off-diagonal semantic matches in the low-dimensional space, contradicting contrastive objectives. To mitigate this, aligned representations undergo further shared-weight processing in I-MSA, leveraging prior modality knowledge. Early image-text matching supervision is added in the MGA output for finer-grained semantic matching between modalities. Features are finally projected back to original dimensions with a residual connection. The final layer is initialized to zero to protect pre-trained performance early on.

Learning Objectives

In multimodal transfer learning for downstream tasks, we define the objective:

$$\min_{\theta^*}\left(\sum_i \mathbb{E}_{x_i\sim\mathcal{D}^i}[L^i_{\text{task}}(f(x_i;\theta^*))] + \sum_{j\neq k}\mathbb{E}_{(x_j,x_k)\sim\mathcal{D}^j\times\mathcal{D}^k}[L^{jk}_{\text{align}}(f(x_j;\theta^*),f(x_k;\theta^*))]\right).$$

Where L^i_{task} is task loss, L^{jk}_{align} is alignment loss between modalities, and θ^* are target parameters. However, same-modality embeddings may cluster excessively, limiting transferability. To ensure uniform alignment, we propose:

$$\min_{\theta^*}\left(L_{\text{ini}} + \lambda_1\sum_i \mathbb{E}_{x_i\sim\mathcal{D}^i}[L^i_{\text{uniform}}(f(x_i;\theta^*))]\right) \quad \text{s.t. } D(\theta,\theta^*)\leq\delta.$$

Where L^i_{uniform} is single-modality uniformity loss, and $D(\theta,\theta^*)$ constrains parameter updates. For image-text retrieval, we propose Adaptive Triplet Loss to align modalities while preventing over-clustering:

$$\mathcal{L}_{\text{ada-triplet}} = \sum_{i=1}^{N} w_i[m+s_{ij}-s_{ii}]_+ + \sum_{j=1}^{N} w_j[m+s_{ji}-s_{ii}]_+.$$

Where w_i and w_j are the weights of sample i and j, determined by the loss size of different samples:

$$w_i = (1-\exp(-[m+s_{ij}-s_{ii}]_+))^\gamma, w_j = (1-\exp(-[m+s_{ji}-s_{ii}]_+))^\gamma.$$

This adaptively focuses on hard samples, enhancing discrimination within/between classes.

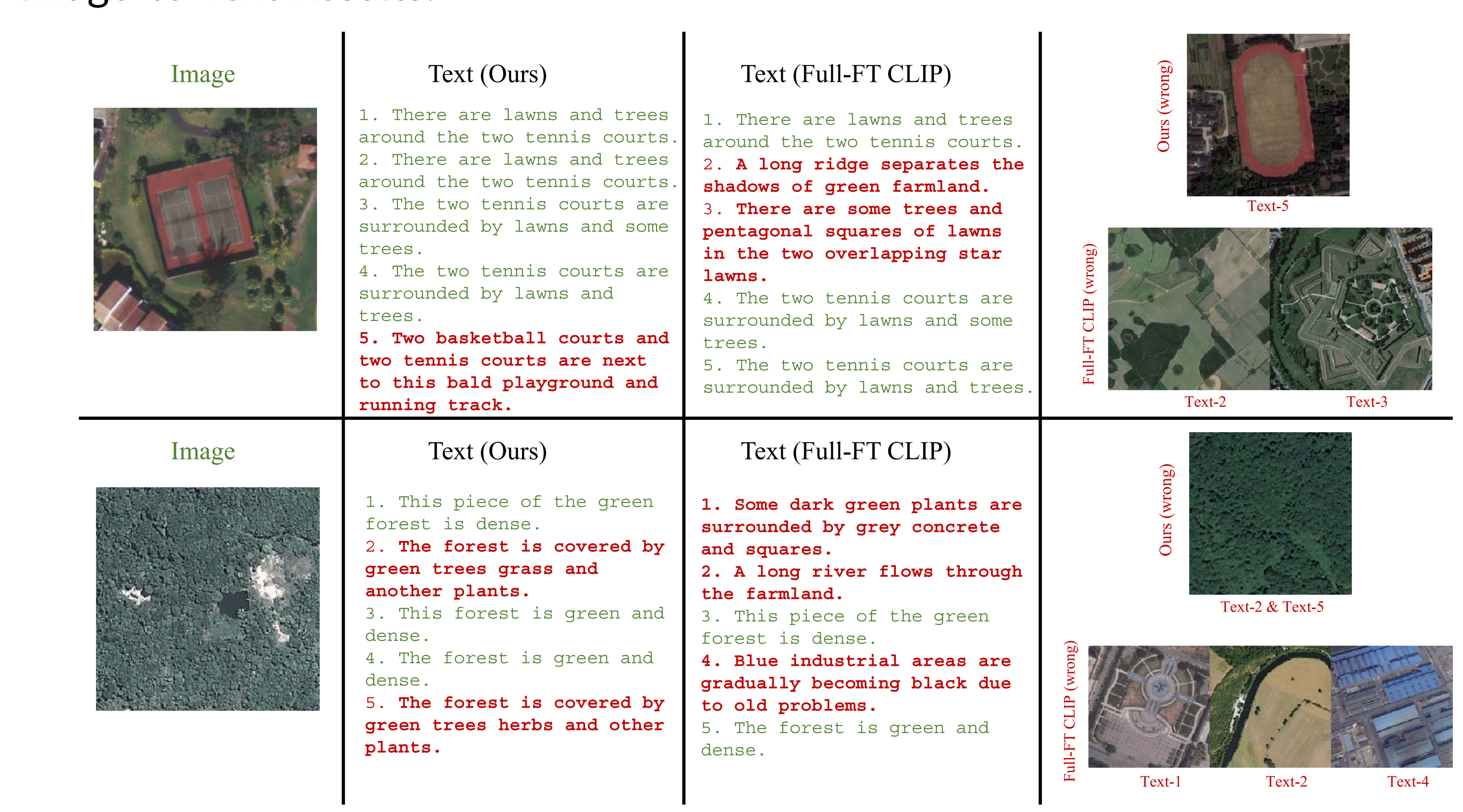
Quantitative Experiment

Table 1. Retrieval Performance Summary on RSICD and RSITMD Test Sets. †: The parameter amount of a single adapter module. Red: Our method; Blue: Full fine-tuned CLIP.

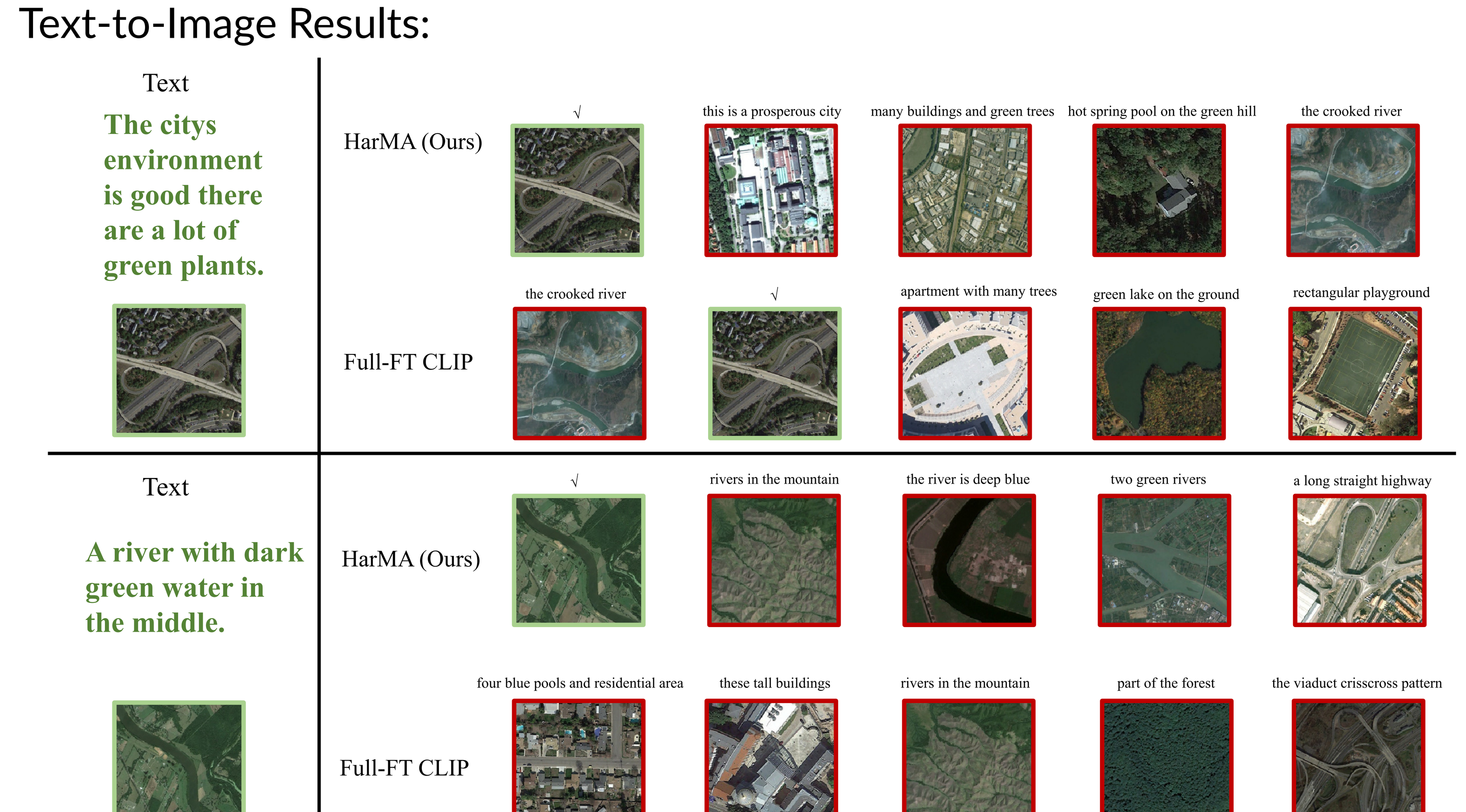
Methods	Backbone	Trainable Params	mR RSICD	mR RSITMD
PIR	Swin Transformer, Bert	-	24.46	38.24
Full-FT CLIP	CLIP(ViT-B-32)	151M	30.39	46.13
Adapter	CLIP(ViT-B-32)	0.17M †	24.84	32.37
CLIP-Adapter	CLIP(ViT-B-32)	0.52M †	21.65	32.38
UniAdapter	CLIP(ViT-B-32)	0.55M †	28.84	39.23
PE-RSITR	CLIP(ViT-B-32)	0.16M †	31.12	44.47
Ours (HarMA w/o Extra Data)	CLIP(ViT-B-32)	0.50M †	32.49	46.53
Ours (HarMA w/o Extra Data)	GeoRSCLIP(ViT-B-32-RET-2)	0.50M †	38.95	52.27

Qualitative Analysis

Image-to-Text Results:



Text-to-Image Results:



- HarMA outperforms fully fine-tuned CLIP in capturing overall semantics.
- HarMA identifies core elements (e.g., tennis courts) while CLIP focuses on irrelevant details (shadows, trees).
- For challenging cases, HarMA retrieves more relevant descriptions and exhibits less hallucination.

 htj@mail.sdu.edu.cn