SCIENCE CHINA
Information Sciences

• LETTER •

November 2026, Vol. 69, Iss. 11, 219101:1–219101:2
https://doi.org/10.1007/s11432-025-5038-y

Optimal transport adapter tuning for bridging modality gaps in few-shot remote sensing scene classification

Zhong JI[1], Ci LIU[1], Jingren LIU[1*], Chen TANG[1*], Yanwei PANG[1] & Xuelong LI[2]

[1]*School of Electrical and Information Engineering, Tianjin Key Laboratory of Brain-Inspired Intelligence Technology, Tianjin University, Tianjin 300072, China*
[2]*Institute of Artificial Intelligence (TeleAI), China Telecom Corp Ltd., Beijing 100033, China*

Received 12 February 2025/Revised 9 April 2026/Accepted 18 May 2026/Published online 29 October 2026

Citation Ji Z, Liu C, Liu J R, et al. Optimal transport adapter tuning for bridging modality gaps in few-shot remote sensing scene classification. Sci China Inf Sci, 2026, 69(11): 219101, https://doi.org/10.1007/s11432-025-5038-y

Few-shot remote sensing scene classification (FS-RSSC) [1] plays a pivotal role in land-use monitoring, environmental change detection, and disaster assessment, yet it faces significant challenges due to the scarcity of labeled training samples. Although traditional approaches relying on transfer learning, meta-learning, or metric learning with CNN backbones have achieved moderate performance in controlled settings, they often fail to capture the complex spatial patterns and diverse information inherent in remote sensing imagery, which leads to suboptimal generalization under extreme data constraints.

In recent years, vision-language models (VLMs), particularly contrastive language-image pre-training (CLIP) [2], have demonstrated exceptional generalization capabilities in data-scarce regimes by aligning image and text embeddings through large-scale vision-language pre-training. This cross-modal alignment capability has inspired numerous studies to adapt VLMs for FS-RSSC tasks via parameter-efficient fine-tuning (PEFT) techniques. While these PEFT methods significantly boost performance by leveraging semantic priors from text labels, existing approaches are predominantly designed and evaluated on natural image datasets, which causes them to face severe challenges when applied directly to remote sensing data. The unique characteristics of remote sensing, such as broad fields of view, varying spatial resolutions, and high information density, create a pronounced modality gap where visual embeddings capture rich spatial details and texture features, whereas textual cues remain semantically sparse and limited to simple category names. Consequently, standard optimization algorithms tend to skew the shared latent space toward the visually dominant modality, which hinders effective cross-modal alignment and degrades classification accuracy, especially in extremely low-shot scenarios containing only one or two examples per class.

To address these limitations, we propose the optimal transport adapter tuning (OTAT) framework, which leverages optimal transport (OT) theory to model cross-modal complementarity and construct an unbiased shared representation space. Specifically, our contributions are as follows.

(1) We introduce the OTAT framework, which utilizes OT theory to minimize modality disparities and reduce the distributional gap between visual and textual features, thereby enabling more effective cross-modal complementarity and synergy to significantly enhance FS-RSSC performance.

(2) We propose the optimal transport adapter (OTA), which employs a cross-modal attention mechanism to enrich sparse textual features with visual semantics and serves as the core carrier for information flow.

(3) Additionally, we introduce a lightweight entropy-aware weighted (EAW) regularization loss as an auxiliary optimization constraint to stabilize the OT optimization (OTO) process and enhance cross-modal alignment.

Method. The OTAT framework is illustrated in Figure 1. We begin by freezing the CLIP encoders and inserting trainable adapters into the Transformer blocks. Subsequently, we apply OT optimization coupled with the proposed EAW loss to facilitate cross-modal complementarity and synergy.

The OTA structure mitigates the modality gap by leveraging cross-modal attention mechanisms. In the visual encoder, we integrate a lightweight low-rank adaptation (LoRA)-based adapter in parallel with the multi-head self-attention (MHSA) and feed-forward network (FFN) modules to facilitate domain adaptation. Concurrently, within each Transformer block of the text encoder, a LoRA-based adapter operates alongside the MHSA, while a dedicated cross-modal attention module is established adjacent to the FFN. In this configuration, text features act as queries while image features serve as keys and values, enabling the model to enrich sparse textual representations with dense visual semantics from remote sensing imagery. We further validate the effectiveness of the OTA design through extensive ablation studies in Appendix C.3.4.

We formalize the alignment by treating features as discrete distributions P_v and P_t. The objective is to find an optimal transport plan $\mathbf{T} = [T_{ij}]$ that minimizes the total transportation cost

$$W_c(P_v, P_t) = \min_{\mathbf{T}\in\Pi(P_v,P_t)} \sum_{i=1}^{L_1}\sum_{j=1}^{L_2} T_{ij}\, c(\mathbf{v}^i, \mathbf{t}^j), \tag{1}$$

subject to the marginal constraints

$$\Pi(P_v, P_t) = \left\{ \mathbf{T} \in \mathbb{R}_+^{L_1\times L_2} \;\middle|\; \textstyle\sum_{j=1}^{L_2} T_{ij} = \frac{1}{L_1},\ \sum_{i=1}^{L_1} T_{ij} = \frac{1}{L_2} \right\}, \tag{2}$$

* Corresponding author (email: jrl0219@tju.edu.cn, tangchen@tju.edu.cn)

© Science China Press 2026

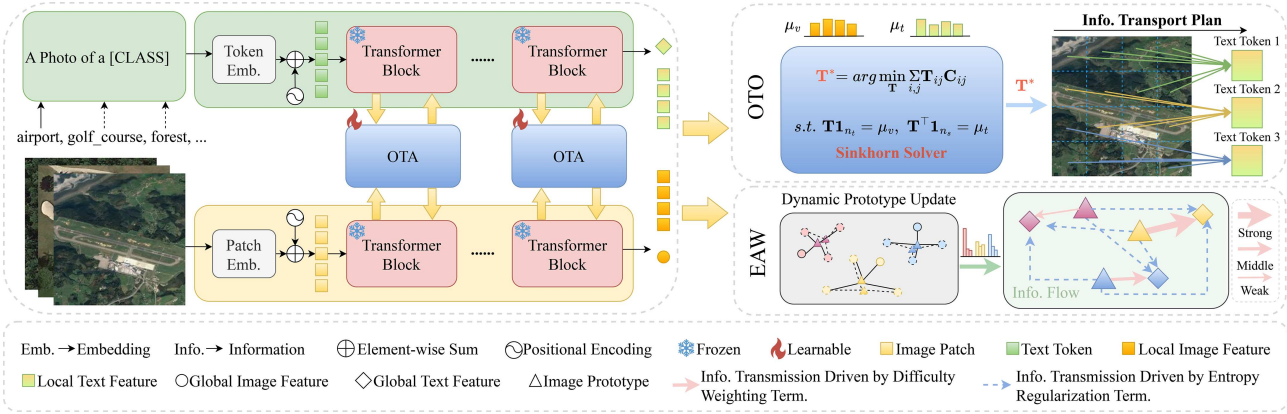


Figure 1 (Color online) Illustration of the proposed OTAT framework, comprising: (1) a frozen CLIP model and a trainable OTA structure for multimodal feature extraction; (2) OTO, which leverages visual knowledge to augment textual information and optimizes information transfer between the two modalities; (3) EAW, integrating adaptive weight adjustment and entropy regularization to derive the optimal OT solver.

where $W_c(P_v, P_t)$ is the OT distance between P_v and P_t, $c(\mathbf{v}^i, \mathbf{t}^j) = 1 - \frac{\langle \mathbf{v}^i, \mathbf{t}^j \rangle}{\|\mathbf{v}^i\| \|\mathbf{t}^j\|}$ is the cosine distance, L_1 is the number of image patches, and L_2 is the number of text tokens. To ensure stability and efficient computation, we introduce the Kantorovich relaxation [3] with entropy regularization and employ the Sinkhorn algorithm [4] to iteratively approximate the entropy-regularized transport plan.

The resulting plan $\mathbf{T}$ defines the probability of a match between an image $\mathbf{v}^i$ and text $\mathbf{t}^j$ as

$$p_{\mathrm{OT}}(y = j|\mathbf{v}^i) = \frac{\exp\left((1 - W_c(P_i, P_j))/\tau\right)}{\sum_{c=1}^{C} \exp\left((1 - W_c(P_i, P_c))/\tau\right)}, \tag{3}$$

where $W_c(P_i, P_j)$ is the OT distance between the i-th image and the text of the j-th category, and τ is a temperature coefficient. This probability distribution guides our OT optimization $\mathcal{L}_{\mathrm{OTA}}$, defined via cross-entropy. This approach ensures precise, high-quality cross-modal alignment and guides the model toward ideal, modality-independent Platonic representations [5].

While $\mathcal{L}_{\mathrm{OTA}}$ establishes baseline cross-modal alignment, this optimization treats all samples uniformly within a batch, which may be suboptimal in FS-RSSC due to significant sample variability. To address this limitation, we introduce the EAW loss, which dynamically adjusts the optimization process based on sample complexity. Unlike the batch-level entropy regularization in Kantorovich relaxation, EAW incorporates both difficulty-weighted similarity and entropy regularization at the sample level. First, the difficulty-weighting term assigns higher weights to hard samples with lower classification confidence, calculated as $k_c = 1 - \max_j p_{c,j}$, where $p_{c,j}$ represents the predicted probability for class j under class prototype p_c. Second, the entropy regularization term $r = -\sum_{c,j} p_{c,j} \log p_{c,j}$ promotes uniformity across similarity distributions, preventing over-concentration on specific class pairs. These components jointly optimize the condition numbers for solving the underlying OT problem while refining cross-modal boundaries. By prioritizing difficult samples and balancing distribution entropy, $\mathcal{L}_{\mathrm{EAW}}$ enhances both local discriminability and global representation consistency. For a more detailed description of our method, please refer to Appendix B.

Experiments. To verify the superiority of OTAT, extensive experiments were conducted on four benchmark datasets against both CLIP-based PEFT methods and specialized remote sensing approaches. Our framework consistently achieves state-of-the-art performance, particularly in low-shot scenarios where it significantly outperforms full fine-tuning. Comprehensive ablation studies validate the contribution of each component, while cross-dataset experiments demonstrate robust generalization. Detailed quantitative results, visualizations, and sensitivity analyses are provided in Appendix C.

Conclusion and future work. This study proposed the OTAT framework for FS-RSSC. By reformulating adapter tuning as an optimal transport problem, we successfully bridged the modality gap caused by information density disparity. The proposed OTA structure, together with the auxiliary EAW regularization loss, facilitates effective cross-modal alignment within a balanced shared representation space. Experiments demonstrate that OTAT establishes a new state-of-the-art for FS-RSSC, offering a promising direction for multimodal remote sensing understanding.

Future research will extend the OTAT framework to more complex scenarios. First, we plan to explore more lightweight and computationally efficient cross-modal interaction mechanisms to further improve the scalability of OT-based optimization. Second, we will investigate its applicability to cross-modal retrieval and domain generalization tasks in remote sensing. Furthermore, we intend to extend the framework to multi-source remote sensing fusion, such as integrating synthetic aperture radar (SAR) and optical modalities to leverage complementary physical characteristics.

Acknowledgements This work was supported by National Financial Fund Project of China (Grant No. KG2026174KYRW14) and National Natural Science Foundation of China (Grant Nos. 62441235, 12272080).

Supporting information Appendixes A–C. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Qiu C, Zhang X, Tong X, et al. Few-shot remote sensing image scene classification: recent advances, new baselines, and future trends. ISPRS J Photogramm Remote Sens, 2024, 209: 368–382
2 Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning, 2021. 8748–8763
3 Kantorovich L V. On the translocation of masses. J Math Sci, 2006, 133: 1381–1382
4 Sinkhorn R. Diagonal equivalence to matrices with prescribed row and column sums. Am Math Mon, 1967, 74: 402–405
5 Huh M, Cheung B, Wang T, et al. Position: the platonic representation hypothesis. In: Proceedings of the 41st International Conference on Machine Learning, 2024. 20617–20642