SCIENCE CHINA
Information Sciences

• Supplementary File •

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

Zhong JI[1,2], Ci LIU[1], Jingren LIU[1*], Chen TANG[1*], Yanwei PANG[1,2] & Xuelong LI[3]

[1]*School of Electrical and Information Engineering, Tianjin Key Laboratory of Brain-inspired Intelligence Technology, Tianjin University, Tianjin* 300072*, China*
[2]*Shanghai Artificial Intelligence Laboratory, Shanghai* 200232*, China*
[3]*Institute of Artificial Intelligence (TeleAI), China Telecom Corp Ltd, Beijing* 100033*, China*

Appendix A Optimal transport theory

VLMs face a fundamental challenge: the asymmetry in information density between visual and textual modalities. This disparity leads to performance degradation in multi-modal classification, particularly in fields such as remote sensing. Simple textual information often fails to capture the complex visual semantics. To address this, we propose a PEFT method for the CLIP model, leveraging OT theory to compensate for the deficiencies in textual data by utilizing the rich visual modality and approximate theoretically ideal Platonic representations for remote sensing tasks.

Optimal Transport theory seeks to efficiently redistribute a commodity from m supply points $\{x_1, x_2, \ldots, x_m\}$, with quantities $\{a_i\}$, to n demand points $\{y_1, y_2, \ldots, y_n\}$, with demands $\{b_j\}$, while minimizing the total transportation cost. Given the cost c_{ij} of transporting one unit from x_i to y_j, the objective is to determine the optimal transport plan f_{ij}:

$$\min_{f_{ij}} \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} f_{ij}, \quad f_{ij} \geqslant 0, \tag{A1}$$

subject to the constraints:

$$\sum_{j=1}^{n} f_{ij} = a_i, \quad \forall i \in [m], \quad \sum_{i=1}^{m} f_{ij} = b_j, \quad \forall j \in [n]. \tag{A2}$$

The classical OT solution has a computational complexity of $O(\max\{m^3, n^3\})$. Entropy-regularized algorithms, such as the Sinkhorn-Knopp algorithm [1], reduce this complexity to $O(\max\{m^2, n^2\}/\epsilon^2)$, where ϵ controls the precision.

In real-world scenarios, when extended to probability measures P_x and P_y, the Monge formulation [2] can seek a transport map $T : X \to Y$ that minimizes the expected cost:

$$\inf_{T} \mathbb{E}_{x \sim P_x}[c(x, T(x))]. \tag{A3}$$

However, the Monge formulation assumes absolute continuity, which may not always hold. The Kantorovich relaxation [3] allows for a probabilistic mapping by minimizing the transportation cost over couplings $\Gamma \in \Pi(P_x, P_y)$:

$$W_c(P_x, P_y) = \inf_{\Gamma \in \Pi(P_x, P_y)} \mathbb{E}_{(x,y) \sim \Gamma}[c(x, y)], \tag{A4}$$

where $\Pi(P_x, P_y)$ is the set of joint distributions with marginals P_x and P_y. When $c(x, y) = \|x - y\|^p$, the resulting $W_c(P_x, P_y)$ defines the p-th Wasserstein distance $W_p(P_x, P_y)$ [4].

In this work, to address the modality asymmetry in the original CLIP model, we introduce an optimization method inspired by the Sinkhorn-Knopp algorithm [1] to achieve the complementarity and alignment of the visual distribution P_x and the textual distribution P_y as described in Eq. A4. By integrating the OT-based adapters into the CLIP model, our method bridges the information gap and approximates the ideal Platonic representations [5] needed for effective adaptation.

Appendix B Method

Building upon the foundational tenets of the Platonic Representation Hypothesis [5], which emphasizes the criticality of unified representations in multimodal encoding, we propose the OTAT framework for FS-RSSC. As illustrated in Figure B1, OTAT employs OT theory to extract intricate cross-modal features, thereby minimizing the discrepancies between modalities and achieving quintessential Platonic representations. To better optimize our OTAT framework on remote sensing datasets, we utilize the Sinkhorn-Knopp algorithm [6] to perform optimization. In addition, we introduce the EAW loss that dynamically adjusts the alignment process by weighting image-text

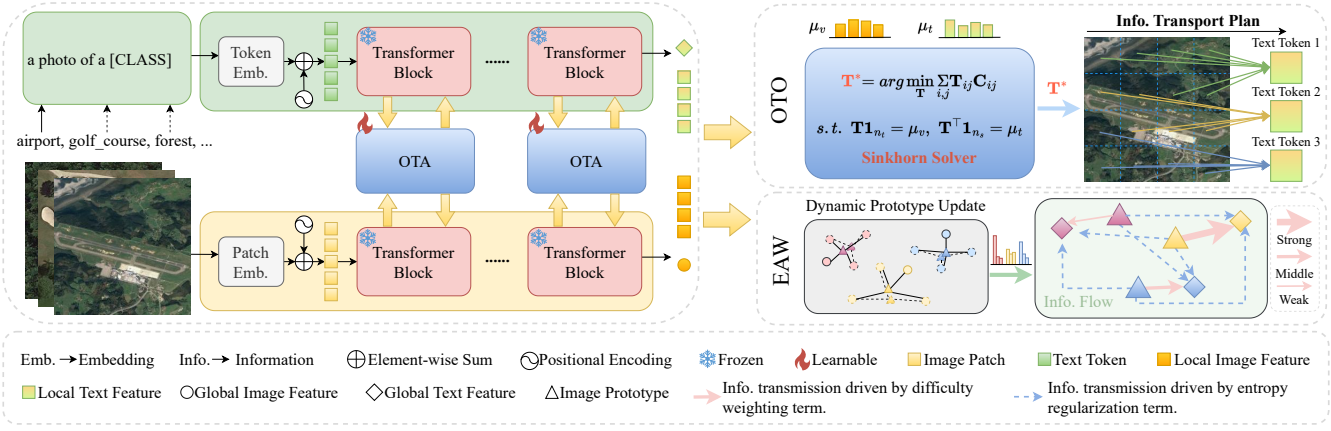


Figure B1 Illustration of the proposed **O**ptimal **T**ransport **A**dapter **T**uning (OTAT) framework, comprising: (1) A frozen CLIP model and a trainable OTA structure for multimodal feature extraction; (2) OTO, which leverages visual knowledge to augment textual information and optimizes information transfer between the two modalities; and (3) EAW, integrating adaptive weight adjustment and entropy regularization to derive the optimal OT solver.

similarity scores according to their complexity and incorporating entropy regularization. This integrated approach transforms the initial CLIP-derived representations into their ideal Platonic counterparts, significantly enhancing performance in remote sensing applications.

Formally, let us consider a dataset $\mathcal{D} = \{(v_i, t_c)\}_{i=1}^{N}$, where $N = C \times K$ denotes the total number of samples, encompassing C distinct classes, each with K exemplars. Here, v_i represents the i-th image datum, and t_c is the textual description corresponding to the c-th class, shared by all image samples within that class. For notational clarity, we denote a single remote sensing (RS) image as $\mathbf{V} \in \mathbb{R}^{H\times W\times 3}$, and its associated text as $T = \{w_j\}_{j=1}^{M}$, where $H \times W \times 3$ specifies the spatial and channel dimensions of the image, and w_j signifies the j-th lexical unit in the text sequence.

Appendix B.1 CLIP representations: initial embeddings

Given the remarkable generalization of CLIP model [7], we select it as our multimodal backbone. The CLIP model integrates a visual and textual encoder, achieving robust cross-modal alignment through contrastive pre-training. However, in specialized domains such as remote sensing [8], medical imaging [9], and embodied intelligence [10], the significant differences in information density between modalities present substantial challenges to optimal performance [5].

Appendix B.1.1 *CLIP encoders*

The visual encoder in the CLIP model adopts a Vision Transformer (ViT) architecture [11]. The image $\mathbf{V}$ is divided into patches and a learnable [CLS] token is prepended, generating patch embeddings. After passing through multiple transformer blocks, the [CLS] token of the final output, $\mathbf{V}_{L_V}^{\text{cls}}$, is projected via a linear transformation $\mathbf{W}_V \in \mathbb{R}^{D_1\times D}$ into the shared embedding space:

$$\mathbf{v} = \frac{\mathbf{V}_{L_V}^{\text{cls}}\mathbf{W}_V}{\left\|\mathbf{V}_{L_V}^{\text{cls}}\mathbf{W}_V\right\|_2}. \tag{B1}$$

This ensures that the global image representation, encapsulated in the [CLS] token, is aligned within the shared embedding space, while other tokens remain unaffected.

Similarly, the textual encoder processes the input text T using a Transformer [12], embedding each word w_j into $\mathbf{e}_j \in \mathbb{R}^{D_2}$ with positional embeddings. After processing through multiple transformer layers, only the [END] token is mapped into the shared space:

$$\mathbf{t} = \frac{\mathbf{T}_{L_T}^{\text{end}}\mathbf{W}_T}{\left\|\mathbf{T}_{L_T}^{\text{end}}\mathbf{W}_T\right\|_2}. \tag{B2}$$

Appendix B.1.2 *Contrastive pre-training*

Finally, the CLIP model is trained using a symmetric cross-entropy loss derived from the InfoNCE [13], which promotes higher similarity for matching image-text pairs. The loss for a batch of N pairs is:

$$\mathcal{L} = \frac{1}{2N}\left(\sum_{i=1}^{N}\mathcal{L}_{\text{image}}^{i} + \sum_{i=1}^{N}\mathcal{L}_{\text{text}}^{i}\right), \tag{B3}$$

where $\mathcal{L}_{\text{image}}^{i}$ and $\mathcal{L}_{\text{text}}^{i}$ are based on cosine similarity between image and text embeddings.

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

Appendix B.1.3 *Adapter-based fine-tuning*

To adapt the CLIP model for the remote sensing domain, we integrate adapters into both the visual and textual encoders. Remote sensing imagery, with its high spatial resolution and specific land-use categories, differs significantly from the natural images used in CLIP's pre-training. Thus, employing PEFT techniques is essential for effective domain adaptation without full model retraining. Among PEFT approaches, adapters are notably effective [14–16], leading us to select them for this purpose.

For the visual encoder, the adapters are inserted in parallel with the MHSA and FFN modules of each transformer block. They are implemented with down-projection $\mathbf{W}_{\text{down}}$, non-linear activation σ, and up-projection $\mathbf{W}_{\text{up}}$, while the adapted representation of the l-th block is:

$$\mathbf{V}_l^{\text{adapted}} = \mathbf{V}_l + \alpha \mathbf{W}_{\text{up}} \sigma \left(\mathbf{W}_{\text{down}} \mathbf{V}_{l-1} \right), \tag{B4}$$

where α is a weight coefficient and $\mathbf{V}_{l-1}$ is the $(l-1)$-th representation.

Similarly, the textual encoder also integrates the adapters in parallel with the MHSA and FFN modules:

$$\mathbf{T}_l^{\text{adapted}} = \mathbf{T}_l + \alpha \mathbf{W}_{\text{up}} \sigma \left(\mathbf{W}_{\text{down}} \mathbf{T}_{l-1} \right). \tag{B5}$$

Despite the parameter efficiency of this construction, it does not fully address the challenges posed by the imbalance in semantic density between remote sensing images and text. These characteristics constrain CLIP's cross-modal alignment capabilities, even with fine-tuned adapters. To surmount these limitations, we employ OT theory to reengineer the structure and optimization of the adapters, generating embeddings that more closely approximate the ideal Platonic representations.

Appendix B.2 Platonic representations: optimal transport adapters

To approximate the ideal Platonic representations—abstract, modality-agnostic embeddings that capture the essence of concepts across different modalities, we introduce an innovative architecture grounded in OT theory. Our objective is to minimize discrepancies between visual and textual feature distributions by efficiently transporting information across modalities, thereby creating a unified embedding space. At the core of our framework is the Optimal Transport Adapters (OTA) structure, which leverages the cross-modal attention mechanism to enrich textual features with visual information. By formulating the alignment between modalities as an OT problem, we aim to find the most efficient mapping of visual features onto textual ones, minimizing the "cost" of this transfer and ensuring harmonious integration.

Specifically, given an image $\mathbf{V}$, we extract its local patch features $\left\{ \mathbf{V}_{L_V}^i \right\}_{i=1}^{L_1}$ from the final transformer block of the CLIP visual encoder. These features are then mapped into a shared embedding space using a linear transformation $\mathbf{W}_V$, resulting in the visual feature map $\mathbf{V}_{\text{local}} = \left[\mathbf{v}^1, \mathbf{v}^2, \ldots, \mathbf{v}^{L_1} \right]^\top \in \mathbb{R}^{L_1 \times D}$. Similarly, we process the textual data by mapping the outputs of the text encoder into the same embedding space, yielding the textual feature map $\mathbf{T}_{\text{local}} = \left[\mathbf{t}^{\text{start}}, \mathbf{t}^1, \mathbf{t}^2, \ldots, \mathbf{t}^{L_2}, \mathbf{t}^{\text{end}} \right]^\top \in \mathbb{R}^{(L_2+2) \times D}$.

Appendix B.2.1 *OTA structure*

Due to the significant information density differences between image and text modalities in FS-RSSC, direct information transfer fails to perfectly achieve the desired alignment. To overcome this, we propose the OTA structure, which leverages cross-modal attention mechanism to enrich the textual information with image features. As shown in Figure B2, we enhance the basic adapter structure by placing the adapters parallel to the MHSA module, and integrating a cross-modal attention mechanism alongside the FFN module. This design not only mitigates the textual information deficiency but also serves as a medium for aligning cross-modal information, preparing the model for further optimization through dynamic, adaptive information transfer driven by the OT plan.

First, in each transformer block, the input features undergo the following transformation:

$$\mathbf{T}_l^{\text{res}} = \mathbf{T}_l^{\text{att}} + \alpha \mathbf{W}_{\text{up}} \sigma \left(\mathbf{W}_{\text{down}} \mathbf{T}_{l-1} \right), \tag{B6}$$

where α is a scaling factor, $\mathbf{W}_{\text{up}}$ and $\mathbf{W}_{\text{down}}$ are the adapter parameters, and σ is the activation function.

To accomplish modal information transfer, we utilize cross-modal attention operations between the textual $\mathbf{T}_l^{\text{res}}$ and visual $\mathbf{v}_i$ from multiple images within the same class. These features are projected into a common query-key-value space:

$$\mathbf{Q}_l^{\text{text}} = \mathbf{W}_Q^{\text{text}} \mathbf{T}_l^{\text{res}}, \ \mathbf{K}_l^{\text{image}_i} = \mathbf{W}_K^{\text{image}} \mathbf{v}_i, \ \mathbf{V}_l^{\text{image}_i} = \mathbf{W}_V^{\text{image}} \mathbf{v}_i, \tag{B7}$$

where $\mathbf{W}_Q^{\text{text}}, \mathbf{W}_K^{\text{image}}, \mathbf{W}_V^{\text{image}} \in \mathbb{R}^{D \times D}$ represent trainable projection matrices that facilitate mapping the input features into a unified latent space for effective cross-modal interaction.

The attention weights, which are conducive to subsequent transmission, are then computed as:

$$\mathbf{W}_l^{\text{cmam}_i} = \text{softmax} \left(\frac{\mathbf{Q}_l^{\text{text}} (\mathbf{K}_l^{\text{image}_i})^\top}{\sqrt{D}} \right). \tag{B8}$$

As a medium for information transfer, this structure is not only optimized for the subsequent OT plan optimization, it also determine how textual features attend to visual features, effectively transferring information from the visual modality. The enriched textual features are then aggregated:

$$\mathbf{T}_l^{\text{cmam}} = \mathbf{T}_l^{\text{res}} + \gamma \cdot \frac{1}{K} \sum_{i=1}^{K} \mathbf{W}_l^{\text{cmam}_i} \mathbf{V}_l^{\text{image}_i}, \tag{B9}$$

where γ is a scaling factor, and K is the number of visual samples from the same category.

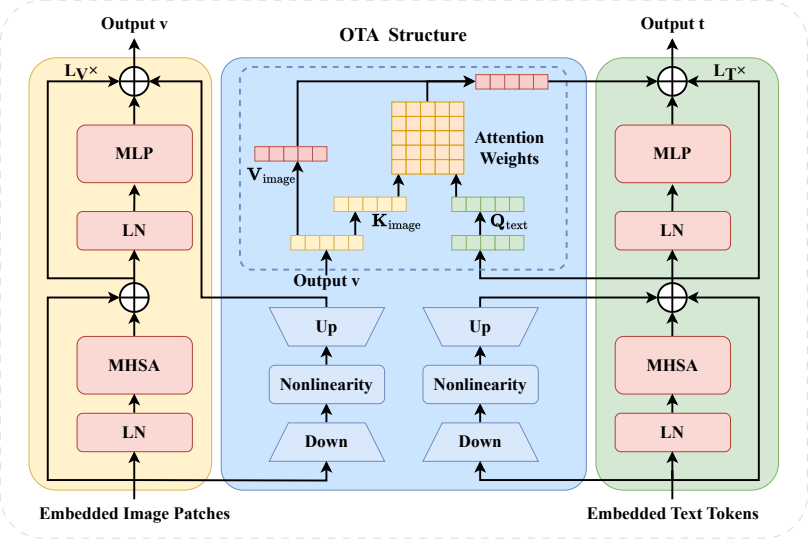


Figure B2 Implementation details of the OTA structure. In the image encoder, adapter layers are placed parallel to both MHSA and FFN blocks. In the text encoder, adapter layers are positioned parallel to MHSA blocks, with a cross-modal attention mechanism added parallel to the FFN block.

The enriched features are processed via the FFN module:

$$\mathbf{T}_l^{\text{enhanced}} = \text{FFN}\left(\text{LN}\left(\mathbf{T}_l^{\text{res}}\right)\right) + \beta \mathbf{T}_l^{\text{cmam}}, \tag{B10}$$

where β balances the contribution of the enriched features, and LN denotes the layer normalization.

By integrating the OT structure within adapters, we ensure efficient and theoretically grounded feature transport between modalities. This structure creates a unified, modality-agnostic space, enriching textual information with relevant visual information, leading to features that closely approximate ideal Platonic forms and enhancing the cross-modal interactions.

Appendix B.2.2 *OTA optimization*

To facilitate smoother interactions between modalities, we propose an optimization approach based on OT solver. We consider the visual and textual features as discrete probability distributions:

$$P_v = \sum_{i=1}^{L_1} a_i \delta_{\mathbf{v}^i}, \quad P_t = \sum_{j=1}^{L_2+2} b_j \delta_{\mathbf{t}^j}, \tag{B11}$$

where $\delta_{\mathbf{x}}$ is the Dirac delta function at $\mathbf{x}$, and a_i and b_j are uniform weights, i.e., $a_i = \frac{1}{L_1}$ and $b_j = \frac{1}{L_2+2}$.

The cost function $c(\mathbf{v}^i, \mathbf{t}^j)$ measures the dissimilarity between $\mathbf{v}^i$ and $\mathbf{t}^j$ via cosine distance:

$$c(\mathbf{v}^i, \mathbf{t}^j) = 1 - \frac{\langle \mathbf{v}^i, \mathbf{t}^j \rangle}{\|\mathbf{v}^i\| \|\mathbf{t}^j\|}. \tag{B12}$$

With the cost fixed, the optimal transport problem seeks a transport plan $\mathbf{T} = [T_{ij}] \in \mathbb{R}_+^{L_1 \times (L_2+2)}$ that minimizes the total transportation cost from P_v to P_t:

$$W_c(P_v, P_t) = \min_{\mathbf{T} \in \Pi(P_v, P_t)} \sum_{i=1}^{L_1} \sum_{j=1}^{L_2+2} T_{ij}\, c(\mathbf{v}^i, \mathbf{t}^j), \tag{B13}$$

subject to the marginal constraints:

$$\Pi(P_v, P_t) = \left\{ \mathbf{T} \in \mathbb{R}_+^{L_1 \times (L_2+2)} \;\middle|\; \sum_{j=1}^{L_2+2} T_{ij} = a_i,\ \sum_{i=1}^{L_1} T_{ij} = b_j \right\}. \tag{B14}$$

To ensure stability and efficient computation, we introduce the Kantorovich relaxation [3]:

$$W_c^{\lambda}(P_v, P_t) = \min_{\mathbf{T} \in \Pi(P_v, P_t)} \sum_{i,j} T_{ij}\, c(\mathbf{v}^i, \mathbf{t}^j) - \lambda H(\mathbf{T}), \tag{B15}$$

where $H(\mathbf{T}) = -\sum_{i,j} T_{ij} \log T_{ij}$ is the entropy of the transport plan, and $\lambda > 0$ is the regularization parameter.

We solve this problem using the Sinkhorn algorithm [6], which iteratively approximates the entropy-regularized OT plan while satisfying the marginal constraints.

First, we define the cost matrix $\mathbf{C} \in \mathbb{R}^{L_1 \times (L_2+2)}$:

$$\mathbf{C}_{ij} = c(\mathbf{v}^i, \mathbf{t}^j) = 1 - \frac{\langle \mathbf{v}^i, \mathbf{t}^j \rangle}{\|\mathbf{v}^i\| \|\mathbf{t}^j\|}. \tag{B16}$$

Next, we define the kernel matrix $\mathbf{K}$:

$$\mathbf{K}_{ij} = \exp(-\lambda \mathbf{C}_{ij}). \tag{B17}$$

We then initialize vectors $\mathbf{p} = 1_{L_1}$ and $\mathbf{q} = 1_{L_2+2}$, and iteratively update them until convergence:

$$\mathbf{p} \leftarrow \frac{a}{\mathbf{K}\mathbf{q}}, \quad \mathbf{q} \leftarrow \frac{b}{\mathbf{K}^\top \mathbf{p}}, \tag{B18}$$

where $a \in \mathbb{R}^{L_1}$ and $b \in \mathbb{R}^{L_2+2}$ denote the marginal distributions vectors corresponding to the visual and textual feature distributions, respectively. Consistent with Eq. B11, their elements are set as uniform weights ($a_i = 1/L_1$ and $b_j = 1/(L_2+2)$).

The convergence criterion is based on the Frobenius norm, where we stop when the difference between successive transport plans is below a threshold ε_1, i.e., $\|\mathbf{T}^{(k+1)} - \mathbf{T}^{(k)}\|_F < \varepsilon_1$.

Upon convergence, the optimal transport plan $\mathbf{T}^*$ is given by:

$$\mathbf{T}^* = \mathrm{diag}(\mathbf{p}) \mathbf{K} \mathrm{diag}(\mathbf{q}). \tag{B19}$$

Finally, the entropy-regularized optimal transport distance is computed as:

$$W_c^\lambda(P_v, P_t) = \langle \mathbf{T}^*, \mathbf{C} \rangle = \sum_{i=1}^{L_1} \sum_{j=1}^{L_2+2} \mathbf{T}_{ij}^* \mathbf{C}_{ij}. \tag{B20}$$

Using the Sinkhorn algorithm, we efficiently and stably solve the entropy-regularized OT problem, facilitating robust information exchange between visual and textual features.

After computing the entropy-regularized OT distance $W_c^\lambda(P_v, P_t)$, we define the probability of a match between an image $\mathbf{v}_i$ and text $\mathbf{t}_j$ as:

$$p_{\mathrm{OT}}(y = j|\mathbf{v}_i) = \frac{\exp\left((1 - W_c^\lambda(P_i, P_j))/\tau\right)}{\sum_{c=1}^{C} \exp\left((1 - W_c^\lambda(P_i, P_c))/\tau\right)}, \tag{B21}$$

where $W_c^\lambda(P_i, P_j)$ is the OT distance between the i-th image and the text of the j-th category.

To enhance cross-modal information transport, with the fixed optimal transport plan $\mathbf{T}^*$, we apply the loss function:

$$\mathcal{L}_{\mathrm{OTA}} = -\frac{1}{N} \sum_{i \in N, j \in C} y_{i,j} \log p_{\mathrm{OT}}(y = j|\mathbf{v}_i). \tag{B22}$$

This approach ensures precise, high-quality cross-modal alignment, guiding the model towards ideal, modality-independent Platonic representations.

Appendix B.3 Regularizing pathway: entropy-aware weighted loss

With the aid of (B22), the CLIP model is initially optimized to approximate Platonic representations. However, this optimization assumes uniform treatment of all samples within a batch, which may be suboptimal in FS-RSSC due to significant sample variability. To alleviate this, we introduce the Entropy-Aware Weighted (EAW) loss, which dynamically adjusts the optimization process based on sample complexity. Unlike the batch-level entropy regularization in (B15), EAW loss incorporates difficulty-weighted similarity and entropy regularization at the sample level, offering more granular control over information flow.

Appendix B.3.1 *Dynamic prototype update*

We compute the class prototypes using a dynamic update strategy. The class prototypes $\mathbf{p}_c$ are initialized to zero and updated using the average of the normalized visual embeddings for each class c as:

$$\mathbf{p}_c = \frac{1}{n_c} \sum_{j=1}^{n_c} \frac{\mathbf{v}_{(j)}}{\|\mathbf{v}_{(j)}\|}, \tag{B23}$$

where n_c is the number of samples in class c.

The update rule incorporates a weighting factor μ, which increases over time, stabilizing near 1:

$$\mathbf{p}_c^{\mathrm{updated}} = \mu \mathbf{p}_c^{\mathrm{old}} + (1 - \mu)\, \mathbf{p}_c^{\mathrm{new}}. \tag{B24}$$

Appendix B.3.2 *EAW loss*

The key contribution of EAW loss is its sample-level entropy regularization, which promotes a balanced similarity distribution across all pairs, ensuring smoother information flow between image and text modalities. The entropy regularization term is defined as:

$$r = -\frac{1}{C} \sum_{c,j \in C} p_{c,j} \log \left(p_{c,j} + \varepsilon_2 \right), \tag{B25}$$

where

$$p_{c,j} = \frac{\exp\left(\text{sim}\left(\mathbf{p}_c^{\text{updated}}, \mathbf{t}_j\right)/\tau\right)}{\sum_{j=1}^{C} \exp\left(\text{sim}\left(\mathbf{p}_c^{\text{updated}}, \mathbf{t}_j\right)/\tau\right)}, \tag{B26}$$

and ε_2 ensures numerical stability.

In addition to entropy regularization, EAW loss incorporates a difficulty-weighting mechanism to further refine the alignment of challenging samples. The difficulty-weighted similarity loss adjusts the focus based on the hardest examples:

$$\mathcal{L}_{\text{difficulty}} = -\frac{1}{C} \sum_{c,j \in C} k_c p_{c,j}, \tag{B27}$$

where $k_c = 1 - \max_j p_{c,j}$. This ensures that more challenging image-text pairs receive greater attention.

The total EAW loss is formulated as:

$$\mathcal{L}_{\text{EAW}} = \mathcal{L}_{\text{difficulty}} - \zeta r, \tag{B28}$$

where ζ balances the similarity distribution uniformity and the focus on challenging samples.

By prioritizing sample-level entropy regularization, EAW loss maintains a balanced similarity distribution while adapting to the complexity of individual samples. The difficulty-weighting mechanism complements this by emphasizing challenging examples. Together, these components enhance the effectiveness of the OTA structure, refining CLIP-derived representations to better approximate ideal Platonic forms and improving the model's performance in remote sensing tasks.

Appendix B.4 Total loss

In addition to the OTA loss and EAW loss, we incorporate a cosine similarity loss, $\mathcal{L}_{\cos}$, to maintain the base classification performance in FS-RSSC, which is defined as:

$$\mathcal{L}_{\cos} = -\frac{1}{N} \sum_{i \in N, j \in C} y_{i,j} \log \frac{\exp\left(\text{sim}(\mathbf{v}_i, \mathbf{t}_i)/\tau\right)}{\sum_{j=1}^{C} \exp\left(\text{sim}(\mathbf{v}_i, \mathbf{t}_j)/\tau\right)}. \tag{B29}$$

These three losses collectively refine the CLIP-derived features into an ideal Platonic features. Specifically, the OTA loss enhances textual information by leveraging visual knowledge and optimizes the information transfer between modalities. The EAW loss introduces entropy regularization and adaptive weight adjustment to address sample importance imbalances, ensuring more effective and balanced feature alignment. The cosine similarity loss $\mathcal{L}_{\cos}$ ensures the maintenance of base performance. Formally, the total loss is expressed as:

$$\mathcal{L}_{train} = \mathcal{L}_{\cos} + \xi \mathcal{L}_{\text{OTA}} + \nu \mathcal{L}_{\text{EAW}}, \tag{B30}$$

where ξ, ν control the contributions of each component.

Appendix C Experiments results and analysis

In this section, we conduct a detailed analysis and experiments to evaluate the performance of our OTAT framework in the FS-RSSC scenarios. We also perform thorough ablation studies, comparative experiments, and visualizations to validate the effectiveness of each component and our motivations.

Appendix C.1 Experimental setup

Appendix C.1.1 *Datasets*

In the experimental section, we evaluate our proposed OTAT framework on four prominent remote sensing datasets: UC Merced [17], NWPU-RESISC45 [18], AID [19], and WHU-RS19 [20]. Inspired by CoOp [21], we divide each dataset into training, validation, and testing sets, comprising 30%, 20%, and 50% of the data per class.

• **UC Merced [17]**: Released by University of California, Merced. It includes 2100 images in 21 classes, and the pixel size per image is 256×256.

• **NWPU-RESISC45 [18]**: Originating from Northwestern Polytechnical University, it consists of 45 scene categories, with 700 images per category, each of size 256×256 pixels.

• **AID [19]**: Co-released by Wuhan University and Huazhong University of Science and Technology, this dataset features 10,000 images across 30 categories, with resolutions of 600×600 pixels.

• **WHU-RS19 [20]**: Compiled by Wuhan University utilizing Google Earth, this dataset comprises 1,005 images, each 600×600 pixels, covering 19 categories with no fewer than 50 images per category.

Table C1 Few-shot classification accuracy (%) comparisons on the UC MERCED and WHU-RS19 datasets.

Approach	Backbone	Setting	UC MERCED							WHU-RS19						
			1-shot	2-shot	4-shot	5-shot	8-shot	16-shot	32-shot	1-shot	2-shot	4-shot	5-shot	8-shot	16-shot	32-shot
SCL-MLNet$_{\text{TGRS}'22}$ [22]			51.37	-	-	68.09	-	-	-	63.36	-	-	77.62	-	-	-
TSC$_{\text{ISPRS}'22}$ [23]			55.11	-	-	69.20	-	-	-	70.99	-	-	82.18	-	-	-
MPCL-Net$_{\text{TGRS}'23}$ [24]			56.46	-	-	76.57	-	-	-	61.84	-	-	80.34	-	-	-
TDNet$_{\text{TGRS}'23}$ [25]	ResNet12	5-way	-	-	-	-	-	-	-	64.24	-	-	84.15	-	-	-
TeAW$_{\text{ICASSP}'23}$ [26]			56.94	-	-	77.50	-	-	-	-	-	-	-	-	-	-
ACL-Net$_{\text{TGRS}'24}$ [27]			59.74	-	-	74.89	-	-	-	78.30	-	-	90.43	-	-	-
AGPLS$_{\text{SSRN}'26}$ [33]	CLIP	5-way	77.19	-	-	92.99	-	-	-	94.34	-	-	97.61	-	-	-
OTAT (Ours)	CLIP	5-way	**96.90**	-	-	**99.11**	-	-	-	**95.79**	-	-	**98.53**	-	-	-
Zero-shot*$_{\text{ICML}'21}$ [7]						65.81							71.40			
Full fine-tuning*$_{\text{ICML}'21}$ [7]			82.86	83.43	92.29	92.38	92.95	96.38	96.67	93.19	94.94	96.50	98.64	98.64	99.22	99.42
Linear Probe*$_{\text{ICML}'21}$ [7]			63.32	76.29	85.53	87.52	90.32	93.02	93.51	79.98	90.19	94.30	96.07	97.30	97.65	97.67
CoOp*$_{\text{IJCV}'22}$ [21]			76.10	83.52	86.38	88.76	91.24	95.14	96.57	87.94	91.44	95.14	96.30	96.11	97.86	97.28
CoCoOp*$_{\text{CVPR}'22}$ [31]			74.10	78.19	82.10	85.14	87.52	89.05	91.81	87.94	90.08	94.16	93.39	95.72	96.89	95.91
MaPLe*$_{\text{CVPR}'23}$ [32]	CLIP	full-way	64.83	74.51	74.70	75.11	79.05	83.87	85.46	80.67	82.36	86.45	87.03	91.96	93.32	94.94
Tip-Adapter-F*$_{\text{ECCV}'22}$ [28]			78.67	82.10	85.90	86.19	88.67	92.95	95.05	92.02	92.83	96.50	96.69	96.50	97.47	98.64
CLIP-Adapter*$_{\text{IJCV}'24}$ [29]			69.71	71.05	75.43	77.52	81.24	83.33	90.19	76.46	78.21	85.02	83.46	91.44	94.75	94.36
RemoteCLIP$_{\text{TGRS}'24}$ [8]			-	-	-	-	-	-	-	40.88	-	60.88	-	80.75	89.45	93.38
LP++*$_{\text{CVPR}'24}$ [30]			78.76	84.53	88.32	88.58	91.21	93.36	93.63	90.58	94.38	96.58	96.83	97.70	97.80	98.13
DA-CLIP$_{\text{J-STARS}'25}$ [34]			-	-	-	-	-	-	-	86.03	90.57	90.60	-	94.50	95.00	–
OTAT (Ours)			**84.67**	**88.10**	**92.67**	**95.62**	**95.90**	**97.14**	**97.81**	**94.55**	**95.13**	**97.28**	**98.83**	**99.03**	**99.42**	**99.61**

* denotes our reproduced results. Unless otherwise specified, all methods adopt ViT-B/16. Notably, RemoteCLIP utilizes ViT-B/32 and DA-CLIP uses ResNet-50, as reported in their original publications.

Appendix C.1.2 *Implementation details*

In this study, all experimental procedures are executed on a single NVIDIA RTX 4090 GPU, employing the pretrained CLIP ViT-B/16 as the foundational model. To meet the input size requirements, remote sensing images are resized to 224 × 224 pixels. Accordingly, the textual prompt template adopts the form "a photo of a [CLS]". For the hyper-parameters in our OTAT framework, the down-projection dimension of the adapter is set to 8, the weight coefficient α is designated as 1.0, and the regularization strength λ is configured to 10. The iterative process of (B18) for the OTA optimization is limited to a maximum of 100 cycles. For dynamic prototype updating, the initial value of μ is 0.5, which is incremented by 0.02 at each step. Training configurations include an initial learning rate of 0.001, leveraging the AdamW optimizer, and the network is trained with a batch size of 64 across 50 epochs. To facilitate adaptive learning rate adjustments, a cosine annealing schedule is implemented throughout the training period.

Appendix C.2 Comparisons with state-of-the-art approaches

To elucidate the advantages of our proposed framework, FS-RSSC experiments are conducted across four distinct datasets, facilitating a direct comparison with state-of-the-art methodologies. These benchmark evaluations encompass both CNN-based approaches, such as SCL-MLNet [22], TSC [23], MPCL-Net [24], TDNet [25], TeAW [26], and ACL-Net [27], as well as ViT-based approaches, including zero-shot CLIP [7], full fine-tuning CLIP [7], Linear Probe [7], Tip-Adapter-F [28], CLIP-Adapter [29], LP++ [30], CoOp [21], CoCoOp [31], MaPLe [32], RemoteCLIP [8], AGPLS [33], and DA-CLIP [34]. This comprehensive assessment ensures a rigorous validation of our OTAT framework's efficacy and robustness in diverse few-shot scenarios.

To ensure a fair comparison with CNN-based approaches, we adopt the meta-learning paradigm for 5-way 1-shot and 5-way 5-shot experiments, following the data partitioning from [35] and [22]. For the ViT-based approaches, we perform full-way classification across 1- to 32-shot scenarios. Additionally, we categorize the ViT-based approaches into zero-shot, full fine-tuning, adapter-based, and prompt-based approaches to underscore the distinct advantages of our proposed framework. The comparative results in Table C1 and Table C2 demonstrate the superior performance of ours approach across four datasets.

As shown in Tables C1 and C2, OTAT demonstrates superior performance across all scenarios. Compared to CNN-based approaches, OTAT yields substantial gains, highlighting the advantage of large-scale pre-trained backbones. More critically, against recent remote sensing-specific methods like AGPLS and DA-CLIP which utilize similar or heavier backbones, OTAT consistently achieves higher accuracy (e.g., surpassing AGPLS by +19.71% on UC Merced 1-shot). This confirms that our OT-based optimization provides significant gains beyond standard task-specific adaptations. Furthermore, OTAT outperforms existing adapter and prompt-based PEFT methods, validating the efficacy of optimal transport in bridging the modality gap.

Notably, OTAT frequently surpasses full fine-tuning, challenging the notion that full fine-tuning is the performance upper bound. This suggests that approximating theoretically optimal Platonic representations is more effective than empirical parameter updates, especially in low-resource regimes. While performance margins narrow as shot numbers increase, OTAT maintains a distinct advantage in 1-shot to 8-shot settings, offering a robust and efficient alternative for few-shot remote sensing classification. Collectively, these results verify that constructing ideal Platonic representations via OT theory significantly enhances cross-modal alignment and generalization.

Appendix C.3 Ablation study

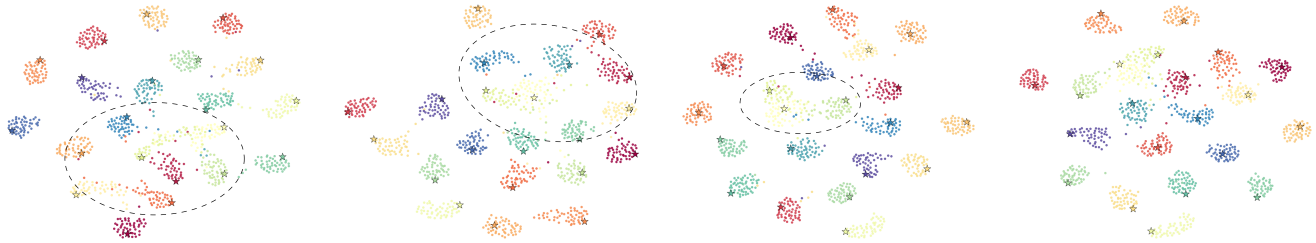
Appendix C.3.1 *Representation learning across various module combinations*

To evaluate the contribution of each component in guiding the CLIP model toward an approximate ideal Platonic representation, we visualize the learned visual and textual features and employ Mutual Nearest Neighbors (MNN) as a metric to assess cross-modal

Table C2 Few-shot classification accuracy (%) comparisons on the NWPU-RESISC45 and AID datasets.

Approach	Backbone	Setting	NWPU-RESISC45							AID						
			1-shot	2-shot	4-shot	5-shot	8-shot	16-shot	32-shot	1-shot	2-shot	4-shot	5-shot	8-shot	16-shot	32-shot
SCL-MLNet$_{\text{TGRS'22}}$ [22]			62.21	-	-	80.86	-	-	-	59.46	-	-	76.31	-	-	-
TSC$_{\text{ISPRS'22}}$ [23]			73.26	-	-	84.62	-	-	-	68.21	-	-	81.72	-	-	-
MPCL-Net$_{\text{TGRS'23}}$ [24]			55.94	-	-	76.24	-	-	-	60.61	-	-	76.78	-	-	-
TDNet$_{\text{TGRS'23}}$ [25]	ResNet12	5-way	65.85	-	-	82.16	-	-	-	67.48	-	-	80.56	-	-	-
TeAW$_{\text{ICASSP'23}}$ [26]			70.23	-	-	85.57	-	-	-	70.35	-	-	84.62	-	-	-
ACL-Net$_{\text{TGRS'24}}$ [27]			76.13	-	-	86.54	-	-	-	70.86	-	-	85.37	-	-	-
AGPLS$_{\text{SSRN'26}}$ [33]	CLIP	5-way	88.21	-	-	91.26	-	-	-	79.71	-	-	83.56	-	-	-
OTAT (Ours)	CLIP	5-way	**93.28**	-	-	**95.96**	-	-	-	**96.42**	-	-	**98.56**	-	-	-
Zero-shot*$_{\text{ICML'21}}$ [7]						57.00							57.20			
Full fine-tuning*$_{\text{ICML'21}}$ [7]			70.36	78.25	**82.36**	83.84	**85.86**	**87.42**	89.14	77.22	82.46	87.68	**90.94**	91.64	**94.44**	95.36
Linear Probe*$_{\text{ICML'21}}$ [7]			52.45	64.72	75.31	77.35	81.79	85.46	88.22	64.66	78.38	85.96	87.25	91.05	93.78	94.84
CoOp*$_{\text{IJCV'22}}$ [21]			63.31	70.11	76.30	77.68	80.29	84.62	87.63	72.12	82.14	82.90	86.66	87.34	90.72	92.70
CoCoOp*$_{\text{CVPR'22}}$ [31]			69.88	71.53	74.36	75.78	79.15	80.72	82.10	71.54	73.50	79.84	84.06	84.58	87.68	89.68
MaPLe*$_{\text{CVPR'23}}$ [32]	CLIP	full-way	57.10	59.43	63.93	63.48	68.51	72.58	76.16	60.17	65.65	64.86	69.91	74.90	80.83	84.59
Tip-Adapter-F*$_{\text{ECCV'22}}$ [28]			67.33	71.30	76.13	78.24	82.98	86.48	88.43	78.86	82.78	87.22	86.76	90.30	93.32	95.08
CLIP-Adapter*$_{\text{IJCV'24}}$ [29]			63.37	66.74	68.69	70.99	74.65	79.89	84.24	64.02	62.66	69.56	70.42	74.30	87.60	92.30
RemoteCLIP†$_{\text{TGRS'24}}$ [8]			42.22	-	60.86	-	70.97	75.88	81.83	36.98	-	65.62	-	75.71	81.07	86.72
LP++*$_{\text{CVPR'24}}$ [30]			71.59	76.07	80.37	81.80	84.00	86.19	88.17	77.68	84.76	88.33	89.31	91.57	93.44	94.32
DA-CLIP$_{\text{J-STARS'25}}$ [34]			58.04	66.51	69.44	-	75.00	79.42	-	70.00	76.54	80.00	-	84.55	88.36	–
OTAT (Ours)			**72.02**	**78.81**	81.70	**84.03**	84.76	87.37	**89.61**	**79.12**	**85.50**	**88.62**	90.12	**92.50**	94.22	**95.70**

* denotes our reproduced results. Unless otherwise specified, all methods adopt ViT-B/16. Notably, RemoteCLIP utilizes ViT-B/32 and DA-CLIP uses ResNet-50, as reported in their original publications.



(a) (b) (c) (d)

Figure C1 The UMAP visualization of generated features under different configurations: (a) Baseline, where the original adapters are optimized with standard cross-entropy loss, (b) OT Optimization (OTO), applying our OT optimization to the original adapters, (c) OTA + OTO, integrating OT optimization into our OTA structure, and (d) OTA + OTO + EAW, further incorporating EAW loss. Dots denote visual features, stars denote textual features, and dashed lines indicate regions of severe class mixing.

alignment performance. Experiments are conducted under four configurations: (1) a baseline configuration where the original adapters are inserted into each transformer block of the frozen CLIP model and optimized with standard cross-entropy loss, (2) applying our OT optimization (OTO) to optimize the original adapters, (3) applying OTO to our proposed OTA structure, and (4) incorporating the EAW loss into the optimization process.

As shown in Figure C1, the baseline demonstrates significant inter-class overlap, with multiple feature points clustered into other classes. Notably, the text features are mostly distributed along the boundaries of clusters, indicating insufficient alignment with visual features and limited semantic representativeness. Introducing OT theory effectively mitigates these issues by redistributing the weights between image and text features through information transfer, enhancing intra-class consistency and reducing inter-class confusion. Building on this foundation, adding the OTA structure significantly enhances the uniformity of feature distributions. More importantly, the text features are progressively drawn closer to the centers of their corresponding visual clusters, demonstrating that they are effectively supplemented with discriminative visual information rather than remaining marginal representations. This improvement arises from the cooperative optimization of both modalities, where text features are enriched with complementary image information. Finally, integrating EAW loss refines the feature learning process even further, nearly eliminating misclassifications and improving separability between challenging categories. In this stage, the star-shaped text features exhibit strong overlap with cluster centers and align tightly with image features, providing clear evidence that our framework successfully enhances text representations and achieves better cross-modal consistency.

Additionally, to demonstrate that we have achieved the construction of theoretically optimal Platonic representations, we provide the change curve of modal alignment performance in Figure C2. With the introduction of OTO, modal alignment is significantly enhanced, indicating that text modality information is effectively enriched and consistent features are learned through image-text information transmission. Building on this, the inclusion of the OTA structure facilitates smoother alignment, with the alignment performance stabilizing at a higher level, highlighting the importance of modality information interaction in optimal transport methods. Finally, incorporating EAW loss as a sample-level regularization term ensures the necessary conditions for solving the entropy-regularized OT problem. Our framework achieves high-quality modal alignment while creating the most ideal Platonic space.

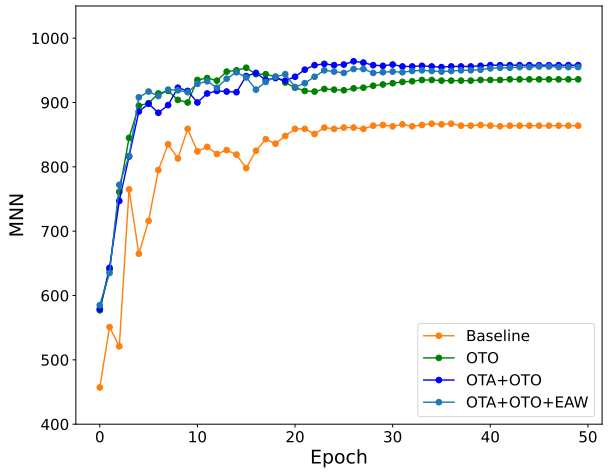


Figure C2 Impact of module combinations on cross-modal alignment. The MNN metric is used to measure alignment over training.

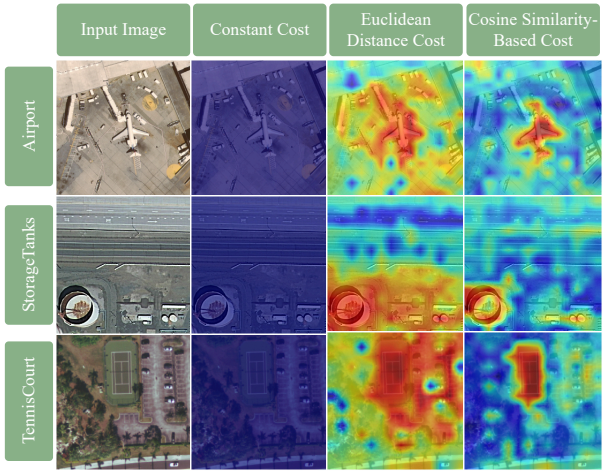


Figure C3 Ablation study on the optimal transport cost function. Warm colors indicate regions with smaller cumulative transport distances to text tokens, highlighting their greater contribution to textual information.

Table C3 Ablation study (%) on component effectiveness under the full-way 5-shot setting.

OTO	OTA structure	EAW	UC MERCED	WHU-RS19	NWPU-RESISC45	AID
			91.52	97.67	80.59	85.84
✓			93.90	97.86	83.10	87.54
✓	✓		94.38	98.64	83.43	89.78
✓		✓	92.00	98.64	83.70	86.72
✓	✓	✓	95.62	98.83	84.03	90.12

These results validate the effectiveness of each component in advancing cross-modal representation learning, culminating in a unified and aligned Platonic representation space with superior performance.

Appendix C.3.2 *Classification performance with different module combinations*

To further evaluate the effectiveness of each component, we conduct an ablation study under the full-way 5-shot setting. The results, shown in Table C3, highlight the significant contributions of each component. Introducing OTO leads to notable performance improvements of 0.19%-2.51% across four datasets, demonstrating its crucial role in enhancing information transfer and compensating for weaker modality. Compared to using OTO alone, adding the OTA structure further improves performance by 0.33% - 2.24%, as the OTA structure minimizes transmission losses and facilitates more robust multi-modal interactions.

To isolate the contribution of the OTA structure, we additionally evaluate the combination of OTO and EAW on a standard adapter architecture. As shown in Table C3, this variant exhibits mixed performance, while it improves over OTO in some cases, it slightly degrades in others. These observations suggest that EAW mainly acts as an auxiliary optimization regularizer rather than an independent alignment module. Its effectiveness largely depends on the OT-guided information transport enabled by the OTA structure.

In contrast, when the OTA structure is introduced, combining all components yields the best and most consistent performance across all datasets. This demonstrates that the proposed OTA structure is essential for stabilizing the optimization process and fully unleashing the benefits of the EAW loss. Overall, OTO, OTA, and EAW play complementary roles, jointly leading to more robust and effective feature representations.

Appendix C.3.3 *Optimal transport cost function*

To investigate the impact of the optimal transport cost function $c(\mathbf{v}^i, \mathbf{t}^j)$ on information transmission, we design three types of cost functions: constant cost (serving as a baseline control), Euclidean distance cost, and cosine similarity-based cost (used in our framework). As shown in Table C4, the Euclidean distance cost results in a noticeable drop in accuracy, highlighting its limitations in capturing nuanced relationships between features. In contrast, the cosine similarity-based cost effectively models the angular relationships between feature vectors, which is crucial for aligning multimodal representations. This superiority can be attributed to the ability of cosine similarity to emphasize relative directions over magnitudes, enabling a more precise transportation plan that enhances model performance. Furthermore, the marginal increase in training time for the cosine similarity-based approach is well-justified by its substantial performance gains, making it the most robust choice.

To investigate how OT theory directs information transfer between image patches and textual tokens, we visualize the cumulative OT distance on the original images as heatmaps, with the value for the i-th patch computed as $1 - \sum_{j=1}^{L_2+2} \mathbf{T}^*_{ij}\mathbf{C}_{ij}$. As shown in Figure C3, the results highlight the effectiveness of the cosine similarity-based cost function, which more precisely focused on class-specific key

Table C4 Ablation study on optimal transport cost function in the full-way 5-shot setting of the UC Merced dataset.

Cost function	Accuracy(%)	Train time(s)
Constant cost	91.52	71.43
Euclidean distance cost	90.29	78.69
Cosine similarity-based cost	93.90	75.26

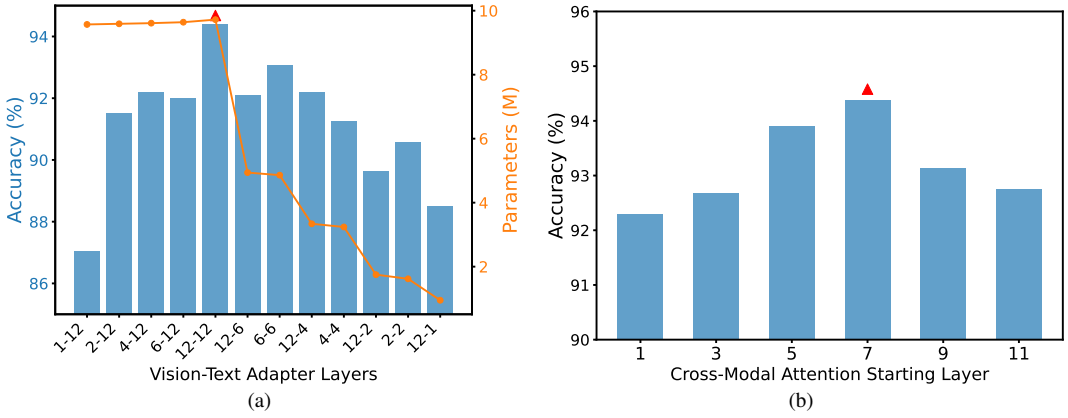


Figure C4 Ablation study on adapter insertion strategy. (a) The number of inserted adapters is varied, with "1V-12T" indicating 1 adapter in the visual encoder and 12 adapters in the text encoder. The curve represents the parameter count, while the histogram illustrates the classification accuracy. (b) The starting layer for the cross-modal attention mechanism is varied to assess its impact on classification performance.

features, such as the fuselage and wings of airplanes or the distinct layouts of tennis courts. This indicates a progressive enhancement in the alignment between visual and textual modalities. Compared to alternative cost functions, the cosine similarity-based approach dynamically refines feature mapping, enabling more precise and interpretable information transfer and complementarity.

Appendix C.3.4 *Adapter insertion strategy*

To further demonstrate the superiority of our OTA component, we conduct an analysis of adapter insertion strategies, focusing on two aspects: the number of inserted adapters and the starting layer of the cross-modal attention mechanism (CMAM). These strategies target different dimensions of adapter integration and collectively influence the model's ability to achieve optimal performance.

As shown in the left panel of Figure C4, the collaborative design of vis-adapters and text-adapters significantly impacts model performance and parameter efficiency. With limited textual information, increasing the number of text-adapters enhances text representation and overall performance. Conversely, adding vis-adapters enhances accuracy with abundant text-adapters, emphasizing the importance of visual information. A balanced adapter structure, such as 6 vis-adapters and 6 text-adapters, achieves 93.05% accuracy with 4.86M parameters, offering a strong trade-off. These findings demonstrate that flexible adapter configurations can optimize performance for specific scenarios while effectively managing resource consumption. As shown in the right panel of Figure C4, the starting position of the CMAM significantly influences performance. Starting at layer 7 achieves the best accuracy, balancing foundational feature extraction and higher-level semantic interaction. Starting too early disrupts low-level feature learning, while starting too late limits the refinement of cross-modal alignment. Intermediate layers, such as layer 5, also perform well, demonstrating their ability to balance both aspects effectively. These results underscore the importance of strategically selecting the CMAM starting layer to optimize performance.

Appendix C.4 Convergence analysis of loss components

To address the convergence behavior of the proposed loss functions and validate their effectiveness in bridging the modality gap, we visualize the training dynamics of each loss component on the UC Merced dataset (5-shot setting). As illustrated in Figure C5, we plot the Total Loss, Cosine Loss, OTA loss, and the components of the EAW Loss (Difficulty Weighted Term and Entropy Regularization).

From the figure, the total loss exhibits rapid decrease in early training followed by gradual stabilization, confirming that joint optimization remains stable. The cosine loss steadily converges to a low value, preserving baseline image-text matching, while the OTA loss drops sharply and approaches near-zero, demonstrating that OT theory effectively minimizes the visual-textual distribution discrepancy and rapidly reduces the modality gap. The EAW loss shows smooth convergence, maintaining a stable negative value as a structural regularizer. The entropy regularization term prevents over-confident matching and promotes uniform similarity distribution, while the difficulty-weighted term gradually stabilizes, indicating progressive improvement on hard samples.

The consistent convergence of all components demonstrates well-coordinated loss design: OTO accelerates cross-modal alignment, EAW stabilizes optimization, and cosine loss maintains baseline performance. This provides strong empirical evidence that the proposed losses effectively reduce the modality gap while ensuring training stability.

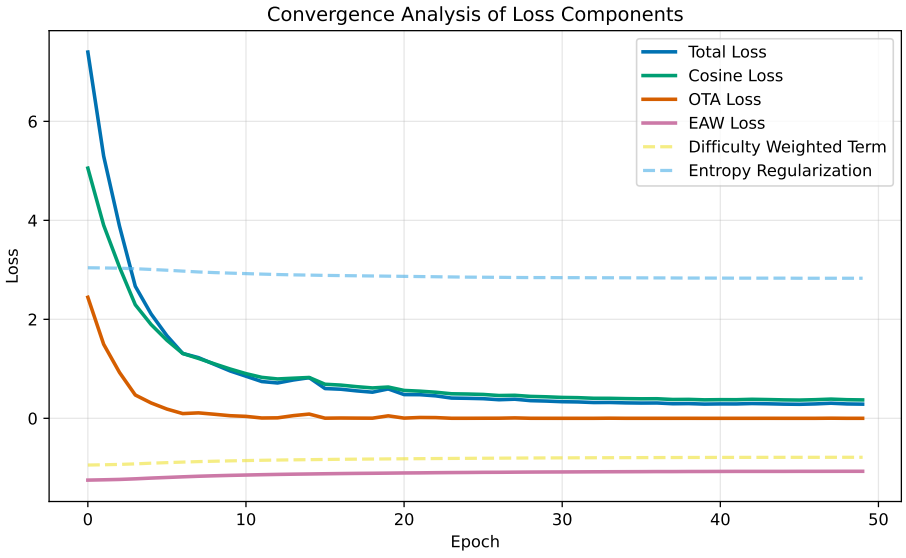


Figure C5 Convergence analysis of loss components on the UC Merced dataset (5-shot). The curves illustrate the training dynamics of Total Loss, Cosine Loss, OTA loss, and EAW Loss components. The rapid convergence of OTA loss validates the efficiency of the OT theory in bridging the modality gap, while the stable EAW Loss indicates effective regularization.

Appendix C.5 Weighting coefficients analysis

Appendix C.5.1 *Impact of OTA loss weight coefficient*

Table C5 highlight the critical impact of the $\mathcal{L}_{\text{OTA}}$ weight ξ on optimizing cross-modal classification accuracy. Accuracy increases from 90.29% to 92.29% in the low-weight region and peaks at 93.90% at $\xi = 1.0$, demonstrating the power of OT theory in guiding effective information complementarity between modalities. Beyond this optimal point, higher weights (e.g., 2.0–5.0) lead to moderate declines, and excessive emphasis (e.g., weight 10.0) disrupts this harmony, degrading accuracy to 90.00%. This may be due to excessive weights causing the network to focus solely on information complementarity and alignment between different modalities, thereby neglecting the classification performance.

Table C5 Impact of $\mathcal{L}_{\text{OTA}}$ weight coefficient ξ in the full-way 5-shot setting of the UC Merced dataset.

Metric	Value 1	Value 2	Value 3	Value 4	Value 5	Value 6
weight	0.01	0.1	1.0	2.0	5.0	10.0
Accuracy(%)	90.29	92.29	93.90	93.14	93.52	90.00

Appendix C.5.2 *Impact of EAW loss weight coefficient*

EAW ensures smooth cross-modal information flow by dynamically adjusting sample contributions. As shown in Table C6, the accuracy peaks at 95.62% when $\nu = 0.08$. Smaller coefficients fail to emphasize challenging samples adequately, while larger coefficients over-regularize the model, disrupting global feature learning. The optimal value strikes a balance between emphasizing challenging samples and maintaining robust global representations. This balance enables the model to refine cross-modal similarity distributions, achieving a more refined representation space and the highest accuracy.

Table C6 Impact of $\mathcal{L}_{\text{EAW}}$ weight coefficient ν in the full-way 5-shot setting of the UC Merced dataset.

Metric	Value 1	Value 2	Value 3	Value 4	Value 5	Value 6
weight	0.01	0.04	0.08	0.12	0.16	0.2
Accuracy(%)	93.33	93.52	95.62	94.29	92.38	93.14

Appendix C.5.3 *Impact of entropy weight coefficient*

To evaluate the effect of entropy weight ζ on model performance within EAW, we fix EAW weight to 0.08 and vary ζ. As shown in Table C7, accuracy improves to a peak of 95.62% at $\zeta = 0.1$. Further increases to 0.2 and 0.5 reduce accuracy to 93.52% and 92.19%, respectively, indicating the adverse effects of excessive weighting. These findings indicate that an entropy weight of 0.1 is optimal for the solvability of the Sinkhorn algorithm under Kantorovich relaxation conditions.

Table C7 Impact of entropy weight coefficient ζ in the full-way 5-shot setting of the UC Merced dataset.

Metric	Value 1	Value 2	Value 3	Value 4	Value 5	Value 6
weight	0.01	0.05	0.1	0.15	0.2	0.5
Accuracy(%)	93.71	93.05	95.62	94.10	93.52	92.19

Appendix C.6 Cross-dataset generalization

Additionally, to further demonstrate the strong generalization capabilities of the final generated Platonic representations, we present the results of cross-dataset generalization experiments in Table C8, comparing the performance of OTAT with several state-of-the-art methods. Using NWPU-RESISC45 as the source dataset, the models are evaluated on UC MERCED, WHU-RS19, and AID as target datasets. The results show that OTAT consistently achieves the best or second-best performance across all target datasets. Notably, OTAT achieves 87.35% on WHU-RS19 dataset and 84.48% on UC MERCED dataset, significantly outperforming other methods. While CoCoOp and APPLeNet exhibit competitive performance on specific datasets, their overall stability and generalization capabilities are inferior to our OTAT framework. Similarly, CoOp and CLIP-Adapter show limitations across different target datasets. The superior performance of our framework can be attributed to its innovative integration of OT-based information transfer and entropy-aware weighted loss design. These components effectively enhance cross-modal alignment and generate Platonic representations, resulting in improved generalization for remote sensing datasets.

Table C8 Comparison with state-of-the-art methods for cross-dataset generalization.

Approach	Source	Target		
	NWPU-RESISC45	UC MERCED	WHU-RS19	AID
CoOp* [21]	77.68	78.57	81.71	58.42
CoCoOp* [31]	75.78	80.57	<u>84.05</u>	**68.10**
CLIP-Adapter* [29]	70.99	76.57	80.35	65.54
APPLeNet* [36]	<u>76.29</u>	<u>80.86</u>	70.04	65.78
OTAT(Ours)	**84.03**	**84.48**	**87.35**	<u>68.02</u>

* indicates results reproduced by us.

References

1 Cuturi M. Sinkhorn distances: Lightspeed computation of optimal transport. Advances in Neural Information Processing Systems, 2013, 26
2 Brenier Y. Polar factorization and monotone rearrangement of vector-valued functions. Communications on Pure and Applied Mathematics, 1991, 44: 375–417
3 Kantorovich L V. On the translocation of masses. Journal of Mathematical Sciences, 2006, 133
4 Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of wasserstein gans. Advances in Neural Information Processing Systems, 2017, 30
5 Huh M, Cheung B, Wang T, et al. Position: The platonic representation hypothesis. In: Proceedings of Salakhutdinov R, Kolter Z, Heller K, et al., editors, Proceedings of the 41st International Conference on Machine Learning. PMLR, 2024. 20617–20642
6 Sinkhorn R. Diagonal equivalence to matrices with prescribed row and column sums. The American Mathematical Monthly, 1967, 74: 402–405
7 Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. In: Proceedings of Meila M, Zhang T, editors, Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021. 8748–8763
8 Liu F, Chen D, Guan Z, et al. Remoteclip: A vision language foundation model for remote sensing. IEEE Transactions on Geoscience and Remote Sensing, 2024
9 Liu J, Zhang Y, Chen J N, et al. Clip-driven universal model for organ segmentation and tumor detection. In: Proceedings of Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 21152–21164
10 Fan L, Wang G, Jiang Y, et al. Minedojo: Building open-ended embodied agents with internet-scale knowledge. Advances in Neural Information Processing Systems, 2022, 35: 18343–18362
11 Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of International Conference on Learning Representations, 2020
12 Vaswani A. Attention is all you need. Advances in Neural Information Processing Systems, 2017
13 He K, Fan H, Wu Y, et al. Momentum contrast for unsupervised visual representation learning. In: Proceedings of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 9729–9738
14 Cao B, Guo J, Zhu P, et al. Bi-directional adapter for multimodal tracking. In: Proceedings of Proceedings of the AAAI Conference on Artificial Intelligence, 2024. 927–935
15 Yuan Y, Zhan Y, Xiong Z. Parameter-efficient transfer learning for remote sensing image-text retrieval. IEEE Transactions on Geoscience and Remote Sensing, 2023
16 Wang Y, Ghamisi P. Rsadapter: Adapting multimodal models for remote sensing visual question answering. IEEE Transactions on Geoscience and Remote Sensing, 2024
17 Yang Y, Newsam S. Bag-of-visual-words and spatial extensions for land-use classification. In: Proceedings of Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Anformation Systems, 2010. 270–279
18 Cheng G, Han J, Lu X. Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE, 2017, 105: 1865–1883
19 Xia G S, Hu J, Hu F, et al. Aid: A benchmark data set for performance evaluation of aerial scene classification. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55: 3965–3981
20 Sheng G, Yang W, Xu T, et al. High-resolution satellite scene classification using a sparse coding based multiple feature combination. International Journal of Remote Sensing, 2012, 33: 2395–2412
21 Zhou K, Yang J, Loy C C, et al. Learning to prompt for vision-language models. International Journal of Computer Vision, 2022, 130: 2337–2348

22 Li X, Shi D, Diao X, et al. Scl-mlnet: Boosting few-shot remote sensing scene classification via self-supervised contrastive learning. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1–12
23 Zeng Q, Geng J. Task-specific contrastive learning for few-shot remote sensing image scene classification. ISPRS Journal of Photogrammetry and Remote Sensing, 2022, 191: 143–154
24 Ma J, Lin W, Tang X, et al. Multi-pretext-task prototypes guided dynamic contrastive learning network for few-shot remote sensing scene classification. IEEE Transactions on Geoscience and Remote Sensing, 2023
25 Wang B, Wang Z, Sun X, et al. Tdnet: A novel transductive learning framework with conditional metric embedding for few-shot remote sensing image scene classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2023, 16: 4591–4606
26 Cheng K, Yang C, Fan Z, et al. Teaw: Text-aware few-shot remote sensing image scene classification, 2023
27 Xu Y, Bi H, Yu H, et al. Attention-based contrastive learning for few-shot remote sensing image classification. IEEE Transactions on Geoscience and Remote Sensing, 2024
28 Zhang R, Zhang W, Fang R, et al. Tip-adapter: Training-free adaption of clip for few-shot classification. In: Proceedings of European conference on computer vision, 2022. 493–510
29 Gao P, Geng S, Zhang R, et al. Clip-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision, 2024, 132: 581–595
30 Huang Y, Shakeri F, Dolz J, et al. Lp++: A surprisingly strong linear probe for few-shot clip. In: Proceedings of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 23773–23782
31 Zhou K, Yang J, Loy C C, et al. Conditional prompt learning for vision-language models. In: Proceedings of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 16816–16825
32 Khattak M U, Rasheed H, Maaz M, et al. Maple: Multi-modal prompt learning. In: Proceedings of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 19113–19122
33 Dong Y, Fu Z, Wang S, et al. Agpls: Attribute-guided prompt learning with stability for few-shot remote sensing image scene classification. Available at SSRN 6381653
34 Deng D, Yao P. Dual-alignment clip: Task-specific alignment of text and visual features for few-shot remote sensing scene classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2025, 18: 19260–19272
35 Li L, Han J, Yao X, et al. Dla-matchnet for few-shot remote sensing image scene classification. IEEE Transactions on Geoscience and Remote Sensing, 2020, 59: 7844–7853
36 Singha M, Jha A, Solanki B, et al. Applenet: Visual attention parameterized prompt learning for few-shot remote sensing image generalization using clip. In: Proceedings of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023