

TopicGeo: An Efficient Unified Framework for Geolocation

Xin Wang Xinlin Wang* Shuiping Gou*
Xidian University, Xi'an, China

{xinwangai@stu., wangxinlin@, shpgou@mail.}@xidian.edu.cn

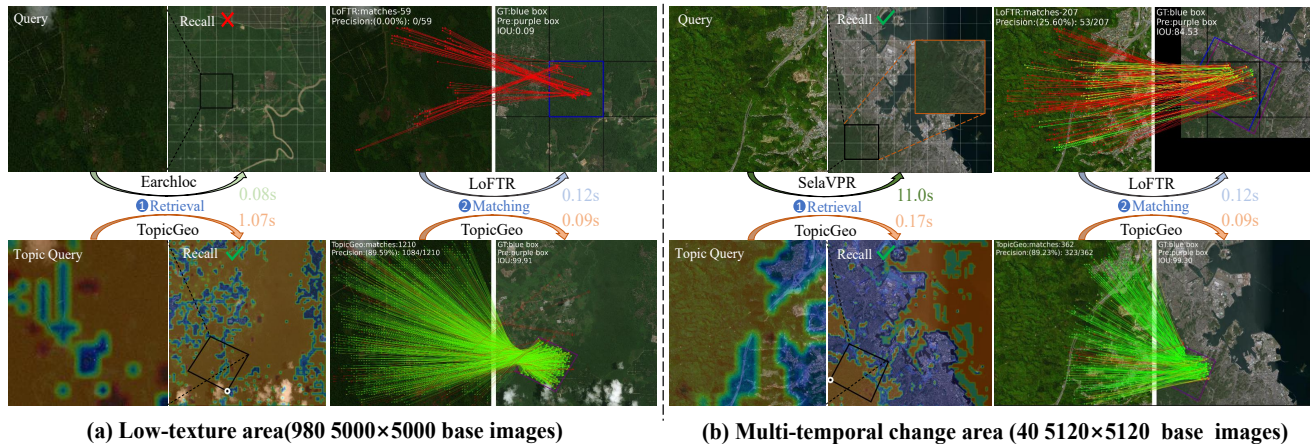


Figure 1. The visual geolocation results of our **TopicGeo**, and typical retrieval and matching approaches: Earthloc [4], SelaVPR [29], and LoFTR [47]. "Recall" represents retrieval results. The color-coded lines show matching correctness: red for higher-error matches, green for higher-precision ones. In various scenes, our method accurately retrieves the target base image containing the query image from the database, while establishing pixel-level correspondence between the smaller query and large-scale retrieved image.

Abstract

Vision-based geolocation techniques that establish spatial correspondences between smaller query images and larger georeferenced images have gained significant attention. Existing approaches typically employ a separate "retrieve-then-match" paradigm, whereas such paradigms suffer from computational inefficiency or precision limitations. To this end, we propose *TopicGeo*, a unified framework for direct and precise query-to-reference image matching via three key innovations. The textual object semantics, called topics, distilled from CLIP prompt learning are embedded into the geolocation framework to eliminate intra-class and inter-class distribution discrepancies while also enhancing processing efficiency. Center-based adaptive label assignment and outlier rejection mechanisms as a joint retrieval-matching optimization strategy ensure task-coherent feature learning and precise spatial correspondences. A multi-level fine matching pipeline is introduced to refine matching from quality and quantity. Evaluations on large-scale synthetic

and real-world datasets illustrate that *TopicGeo* achieves state-of-the-art performance in retrieval recall and matching accuracy while maintaining a balance in computational efficiency.

1. Introduction

Image matching through establishing spatial correspondence between localized query images and extensive georeferenced databases has emerged as a critical technique for vision-based geolocation, playing a vital role in environmental monitoring [13, 53], human activities [26, 54]. However, this task faces inherent challenges. Sensor variations and temporal differences induce substantial discrepancies [31] and nonlinear geometric deformations, complicating cross-image accurate matching. Moreover, repetitive low-saliency regions, such as forests, deserts, and water bodies in georeferenced images, pose matching ambiguities.

Recent advances in satellite image retrieval have focused on mitigating these challenges through global retrieval approaches [4, 24, 48]. These methods encode entire images

* Corresponding author

as high-dimensional vectors for rapid retrieval of similar images from a database. However, such methods exhibit certain critical limitations. Global feature aggregation dilutes spatially fine-grained discriminative patterns, causing failures in low-saliency regions, such as uniform forests or barren land. In addition, when pairing large-size reference images, images are often cropped into smaller and partially overlapped pieces [4], which causes redundant computation. Local re-ranking methods [18, 29, 43] address these challenges by re-ranking candidates based on local matches. However, false positive candidates that are merely visually similar do not contribute to performance improvements. Retrieval-based methods exhibit limited geolocation accuracy, posing significant challenges for researchers in performing time-consuming and laborious manual corrections on multi-source imagery. A natural approach would involve first extracting candidate regions via a retrieval method, then refining for pixel-level matching with a specialized model [3, 31, 45]. However, a critical limitation of this approach lies in the decoupled retrieval and matching pipelines, which inherently prioritize different features in different tasks. In addition, the matching remains heavily dependent on the retrieval model’s effectiveness.

Recent studies [45, 49] apply a single end-to-end matching model for joint retrieval and matching. While enhanced by fine-grained features to boost accuracy, these methods face a huge time overhead [45], preventing real-time geolocation. The above geolocation methods rely on visual features for location estimation. However, remote sensing images exhibit a pronounced imbalance of areal distribution of terrestrial features, inducing inherent bias in visual information-dependent models toward overrepresented categories. Moreover, geometric distortions between multi-source images and cross-temporal visual discrepancies lead to variations in the distribution of geographic features at the same location. Some methods [43, 55] learn more robust representations using natural images by incorporating semantic segmentation. In remote sensing, datasets [8, 39] and segmentation models [32, 33, 50] offer semantic information but face generalization limitations. Recent studies [15, 16] leverage self-supervised learning of image topics to enhance visual features. However, the resulting topic distributions often lack explicit semantic information.

In this study, we propose a model that jointly optimizes retrieval and matching tasks for geographic positioning. Specifically, we employ asymmetric-input architecture to minimize redundant computation in large-displacement and large-scale scenarios. The core innovation lies in hierarchical textual embeddings for prompt learning and a topic-aware encoder trained via knowledge distillation and spatial alignment, yielding efficient topic features that are beneficial for both image retrieval and coarse matching. Building on this foundation, a multi-level refinement pipeline it-

eratively resolves positional cues across scales, improving quality and quantity of matching points. Our unified framework is jointly optimized through effective label supervision and quality-aware rejection mechanisms. The contributions of the study are as follows:

- We propose a unified retrieval and matching framework with asymmetric large-scale input, boosted by embedded topics for efficiency and robustness to intra-class discrepancies.
- The center-based label assignment is designed to optimize the contrastive loss, which is further weighted through topic-based category perception to balance inter-class discrepancies.
- The local spatial consensus filter and the global geometric verification module are co-designed to eliminate mismatched correspondences, thereby enhancing retrieval recall and matching precision in geolocation tasks.

2. Related Work

Visual Language Model. The pioneering visual-language model (VLM) CLIP [38] employs vision and text encoders to project images and texts into a shared semantic space, which achieves joint embedding of images and texts, making cross-modal retrieval and zero-shot learning possible. Since then, VLMs have rapidly advanced and been applied across fields like open-vocabulary detection [6], image-text retrieval [30], and action recognition [52]. Menon et al. [34] apply large language models (LLMs) to represent object classes as semantic attribute sets for enhanced classification. Recently, Liu et al. present the first remote sensing visual-language foundation model, RemoteCLIP[25], learning semantically rich visual features aligned with text embeddings for seamless downstream applications. Our method leverages the textual semantics information of objects to capture invariant matching features across multi-source and multi-temporal satellite images.

Image Retrieval. The image retrieval task focuses on identifying images similar to a query image within a database, encompassing two primary approaches: global and local feature-based retrieval. In global retrieval, the fundamental challenge lies in effectively aggregating image features from diverse regions into a cohesive high-dimensional vector. Traditional methods accomplish this by aggregating handcrafted local features [2, 5, 28] into global representations through methods such as Fisher Vectors [35], and then performing efficient retrieval through nearest neighbor search algorithms. Recent advancements leverage deep learning for feature aggregation [1, 37], employing strategies such as margin-based classification loss functions [9, 27] and clustering training samples [4]. However, global retrieval often results in a low recall rate during the initial retrieval phase. To address the issue, local matching is often integrated to re-rank retrieval results [18, 29, 43]. Geomet-

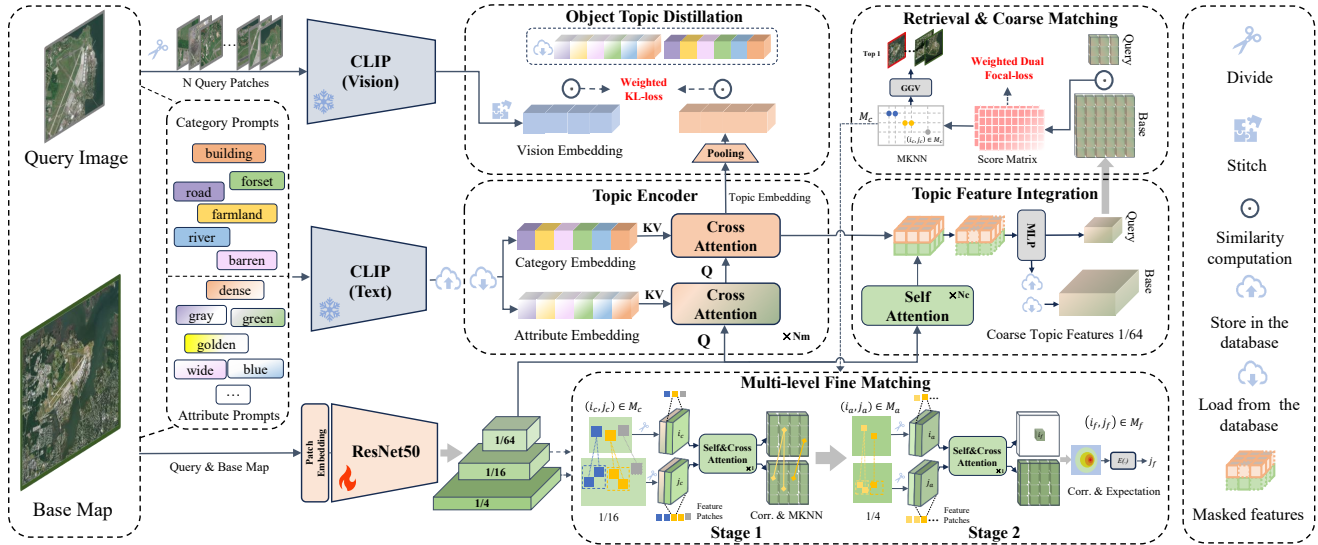


Figure 2. Overview of TopicGeo. (a) Multi-level features are extracted from input images. (b) CLIP-generated prompt embeddings and visual features are injected into cross-attention layers and then distilled to infer the topic distribution. The extracted topic embeddings are integrated with global context features from self-attention layers. Preliminary matches are obtained using dual-softmax and local mutual k-nearest neighbor (MKNN) filter, followed by global geometric verification (GGV) for retrieval. (c) Adaptive matching is performed on cropped patches at a secondary resolution, with further coordinate refinement through fine-grained feature patch cropping.

ric verification methods like RANSAC [14] are employed to identify reliable candidates, though they frequently suffer from computational inefficiency. Fast verification approaches based on matching pair counts [29] face additional challenges in remote sensing applications, where weakly textured regions often lead to significant false positives. Unlike these retrieval-specific pipelines, this paper establishes a unified framework that seamlessly integrates retrieval and matching processes and introduces an efficient geometric verification mechanism.

Image Matching Image matching establishes accurate spatial correspondences between two images, essential for tasks like visual localization [21, 41] and 3D reconstruction [23]. The core of image matching is keypoint detection and description. Classical methods [2, 5, 28] employ hand-crafted features to detect and describe keypoints. Recent advances leverage deep learning for improved sparse keypoint detection [10, 40, 44], and self-supervised descriptor learning [17]. SuperGlue [42] harnesses the graph neural network with attentional mechanism to address significant occlusion or viewpoint changes, but relying on the quality of keypoints. RoMa [12] significantly enhances accuracy via Markov chain-based dense matching but incurs high computational cost. Detector-free approaches like LoFTR [47] use position embedding and cross-attention for semi-dense matching, which has gained traction in remote sensing [3, 45, 49], yet struggle with large-scale or point-of-view transformation. AdaMatcher [20] proposes an adap-

tive label assignment strategy to solve this problem, but introduces inconsistent learning-driven global mismatch filtering. TopicFM [15] sets up learnable topic embeddings for topic perception, while the unknown topics diverge from human perception and fail in complex scenes. Most existing methods rely on cross-attention for feature interaction, which limits computational efficiency. In contrast, our approach leverages distilled topics for feature enhancement and a learning-free local filter for efficient correspondence refinement.

3. Proposed Approach

Our goal is to retrieve a reliable base image from a non-overlapping image database using a small query image and then match the query image to the retrieved larger-scale base image. Let $R = \{B_i | i \in [1, m]\}$ be a retrieval database containing m high resolution base images, and A be a query image. We establish a unified retrieval and matching framework, which encodes R into features $F_R = \{F_{B_i} | i \in [1, m]\}$ storing as a feature database, and A into F_A . Then A is retrieved, and the corresponding target base image B is obtained. Finally, the fine-grained features of both A and B are extracted to refine the matching and compute the homography H . Figure 2 gives an overview of our pipeline.

3.1. Topic-aware Feature Extraction

This paper tackles distribution discrepancies of land objects by establishing an effective topic perception process that integrates primitive visual features with high-level text representations independent of vision. Firstly, we extract visual features of the query image A and the base image B . To be efficient and preserve sufficient information, both of them undergo processing through patch embedding [11]. Then the downsampled images are fed into a shared ResNet [19] to extract coarse-grained visual features, $(F_A^{1/64}, F_B^{1/64})$, and fine-grained visual features $(F_A^{1/16}, F_B^{1/16})$ and $(F_A^{1/4}, F_B^{1/4})$.

Hierarchical topic prompts. To accelerate convergence and integrate richer information, we design dual-level prompts for land objects. Specifically, generic and stable categories are extracted from established land cover datasets [50]. These categories are enriched through object attribute prompts generated by the GPT-4 language model. Subsequently, the embeddings of the augmented object prompts are encoded using the RemoteCLIP[25] text encoder. This process generates category embeddings $\mathbf{F}_{\text{category}} \in \mathbb{R}^{C \times d}$ and attribute embeddings $\mathbf{F}_{\text{attribute}} \in \mathbb{R}^{C' \times d}$.

Object topic distillation. The obtained attribute and category prompt embeddings are embedded into visual features in sequence via cross-attention layers, where visual features serve as queries while these prompt embeddings act as keys and values. This process synthesizes the fused textual and visual representation to generate topic embeddings $\mathbf{F}_{\text{topic}}^{1/64}$. Notably, the sparsity of textual semantic tokens drastically lowers the computational cost of cross-attention.

However, such embeddings fails to align category-specific and attribute-related information with their corresponding visual semantics. Leveraging the inherent alignment between visual and textual features in RemoteCLIP, we employ its frozen vision encoder to further extract the visual embeddings $\mathbf{F}_{\text{vision}} \in \mathbb{R}^{N \times d}$ of N patches from A . Specifically, the patch-level category perception is established by computing the category probability distribution $p \in \mathbb{R}^{N \times C}$:

$$p = \text{softmax} \left(\frac{\mathbf{F}_{\text{vision}} \mathbf{F}_{\text{category}}^T}{t} \right), \quad (1)$$

where t is the temperature parameter.

To transfer the category-specific information from $\mathbf{F}_{\text{vision}}$ to topic embedding $\mathbf{F}_{\text{topic}}^{1/64}$, the latter are spatially aligned with $\mathbf{F}_{\text{vision}}$ via 2×2 pooling, resulting in resized topic embedding $\mathbf{F}_{\text{topic}}^p \in \mathbb{R}^{N \times d}$. The corresponding category probability distribution for the pooled topic embedding is computed analogously to p , denoted as $q \in \mathbb{R}^{N \times C}$. Thus, the weighted Kullback-Leibler (KL) divergence loss $Loss_{\text{category}}$ is employed to transfer the category-specific in-

formation in $\mathbf{F}_{\text{vision}}$ to topic embeddings, expressed as:

$$Loss_{\text{category}} = \sum_{i=0}^N \left(\sum_{j=0}^C \left(p_{ij} \cdot \log \left(\frac{p_{ij}}{q_{ij}} \right) \right) \cdot w_i \right) \cdot t^2, \quad (2)$$

where w_i , the normalized class frequency weight of the i -th image patch, is derived from p . Analogously, attribute-related information is transferred using $Loss_{\text{attribute}}$. Then the topic embeddings are enriched by minimizing the combined loss:

$$L_d = Loss_{\text{category}} + Loss_{\text{attribute}}. \quad (3)$$

Note that distillation operates exclusively on query images due to the fully shared parameters and training efficiency.

Topic feature integration. To capture global context absent from $\mathbf{F}_{\text{topic}}$, the linear self-attention[51] with rotary position embedding[46] is further employed. Thus, the outputs from parallel encoding branches are integrated via a multilayer perceptron (MLP), with features from one branch stochastically masked during optimization to mitigate MLP-induced training bias. The final topic features, denoted as $\hat{F}_A^{1/64}$ and $\hat{F}_B^{1/64}$, are utilized for downstream retrieval and matching.

3.2. Retrieval & Coarse Matching

The multimodal topic features of query and base images, $\hat{F}_A^{1/64}$ and $\hat{F}_B^{1/64}$, enable efficient retrieval and coarse matching. Inspired by [20], the coarse-level matching probability matrix \mathcal{S} and the bidirectional matching probability matrixes \mathcal{P} are calculated as :

$$\mathcal{S}(i, j) = \left\langle \hat{F}_{1/64}^A(i), \hat{F}_{1/64}^B(j) \right\rangle, \quad (4)$$

$$\mathcal{P}^{AB}(i, j) = \text{Softmax}(\mathcal{S}(i, \cdot))_j, \quad (5)$$

$$\mathcal{P}^{BA}(i, j) = \text{Softmax}(\mathcal{S}(\cdot, j))_i, \quad (6)$$

where $\langle \cdot, \cdot \rangle$ is the inner product. Based on \mathcal{P}^{AB} , an index pair (\tilde{i}, \tilde{j}) is selected as the match if its corresponding confidence value exceeds the threshold θ and is the maximum in its row. Thus, matches \mathcal{M}_{AB} is obtained as:

$$\mathcal{M}_{AB} = \{(\tilde{i}, \tilde{j}) \mid \mathcal{P}^{AB}(\tilde{i}, \tilde{j}) = \max_k \mathcal{P}^{AB}(\tilde{i}, k), \mathcal{P}^{AB} \geq \theta\}. \quad (7)$$

Analogously, \mathcal{M}_{BA} is derived by applying identical thresholding and column-wise maximization operations to \mathcal{P}^{BA} , completing the bidirectional coarse matching process.

Local MKNN filter. While LoFTR [47] employs mutual nearest neighbor (MNN) filtering to suppress local mismatches, this strategy exhibits inherent limitations in one-to-many mismatches prevalent in resolution-scaled discrepancies. Instead, we propose mutual k -nearest neighbor (MKNN) filtering, which relaxes mismatches constraints by retaining matches that satisfy k -neighbor spatial consistency across bidirectional correspondence sets \mathcal{M}_{AB} and \mathcal{M}_{BA} . Specifically, a candidate match pair $(i, j) \in \mathcal{M}_{AB}$

is preserved if there exists a reverse-matching pair $(i', j) \in \mathcal{M}_{BA}$, and the 2D coordinate pair (x, y) and (x', y') corresponding to points i and i' are k -nearest neighbors. Analogously, \mathcal{M}_{BA} is handled in the same manner. Finally, the refined coarse matches are combined using the OR operation to obtain \mathcal{M}_c , and the overall process is formulated as follows:

$$\mathcal{M}_c = \text{MKNN}(\mathcal{M}_{AB}, \mathcal{M}_{BA}). \quad (8)$$

Global geometric verification. It is known that correct correspondences have geometric spatial consistency. Therefore, we utilize the rule to filter false positive samples in the database efficiently. Inspired by [36], we independently analyze distance and angle compatibility for each correspondence pair $\mathcal{M}_c(n)$ among the N_c total pairs. For distance compatibility, we calculate pairwise Euclidean distances between each point and all other $N_c - 1$ points in both the query and base image coordinate systems, calculating cross-coordinate distance ratios and normalizing it. For angle compatibility, we evaluate the local geometric consistency by computing relative angles among each point's k -nearest neighbors in both coordinate systems, and measuring their cross-coordinate absolute angular difference c_n . Finally, the distance compatibility α_n and angle compatibility β_n of each matching pair $\mathcal{M}_c(n)$ are yielded to jointly quantify spatial coherence, defined as:

$$\alpha_n = \max\left(0, \min\left(Z - \frac{\text{Var}(d_n)}{\delta_d^2}, Z - 1\right)\right), \quad (9)$$

$$\beta_n = \max\left(0, \min\left(Z - \frac{c_n^2}{\delta_c^2}, Z - 1\right)\right), \quad (10)$$

where δ_d and δ_c are the maximum acceptable values, and Z is a scalar, used to truncate outliers. The spatial coherence confidences weighted by λ is utilized to re-weight the per-pair bidirectional matching probabilities E_n , expressed as:

$$E_n = (\lambda\beta_n + (1 - \lambda)\alpha_n) \times (P^{AB}(\tilde{i}_n, \tilde{j}_n) + P^{BA}(\tilde{i}_n, \tilde{j}_n)). \quad (11)$$

The global confidence E for the base image is then obtained by aggregating all individual matching probabilities. Finally, the base image only with the highest overall confidence score is selected as the target retrieval image.

The joint retrieval-matching loss L_c is defined as:

$$L_c = \left(\text{FL}(\mathcal{P}^{AB}, \hat{\mathcal{P}}^{AB}) + \text{FL}(\mathcal{P}^{BA}, \hat{\mathcal{P}}^{BA})\right) \times w_{token}, \quad (12)$$

where $\hat{\mathcal{P}}$ is the ground truth matching matrix. We allow one-to-many and many-to-one label assignments [20] for \mathcal{P}^{AB} and \mathcal{P}^{BA} . FL refers to the Focal Loss [22]. w_{token} is derived from the copy of w in the early training period, and the weight calculated by the category perception at the $\mathbf{F}_{\text{topic}}^{1/64}$ is used in the subsequent training.

3.3. Adaptive Center Assignment

Downsampling operations inherently yield sparse matching pairs, exacerbated by resolution-scaled discrepancies and asymmetric input. The adaptive label assignment [20] and MKNN filter alleviate the issue, while its reliance on the upper-left coordinates of the warped grid [42, 47] for label assignment often induces suboptimal matching patch overlaps, especially under large rotational misalignments in remote sensing images. This paper proposes a center-based label adaptive assignment strategy. Labels are assigned by distorting grid center coordinates instead of upper-left points, which are later reconverted to the upper-left coordinates for feature patch indexing, thereby enhancing robustness in challenging conditions.

3.4. Multi-level Fine Matching

The retrieval-oriented low-resolution encoding fundamentally limits matching precision. To achieve enhanced correspondence density through operationally consistent processing with the coarse matching stage, we employ the same adaptive local window matching on the finer-grained feature patches $(\hat{F}_A^{1/16}, \hat{F}_B^{1/16})$ corresponding to \mathcal{M}_c , thereby obtaining \mathcal{M}_a . Subsequently, following LoFTR[47], higher-resolution feature patches $(\hat{F}_A^{1/4}, \hat{F}_B^{1/4})$ extracted from $(F_A^{1/4}, F_B^{1/4})$ guided by \mathcal{M}_a are utilized for final expectation-based fine-grained matching, obtaining \mathcal{M}_f , achieving sub-pixel matching accuracy. For adaptive local window matching, the loss is formulated as:

$$\text{Loss}_a = \frac{1}{|\mathcal{M}_a|} \left(\text{FL}(\mathcal{P}_a^{AB}, \hat{\mathcal{P}}_a^{AB}) + \text{FL}(\mathcal{P}_a^{BA}, \hat{\mathcal{P}}_a^{BA})\right), \quad (13)$$

where $\hat{\mathcal{P}}_a$ represents the ground truth matching matrix. The expectation-based loss Loss_f refers to [47]. Thus, the final fine-grained loss is the sum of both:

$$L_f = \text{Loss}_a + \text{Loss}_f. \quad (14)$$

3.5. Training

The model is trained by optimizing the topic distillation loss L_d , the joint retrieval-coarse matching loss L_c , and fine-grained matching loss L_f , defined as:

$$L = L_d + L_c + L_f. \quad (15)$$

4. Experiments

4.1. Datasets and Implementation Details

TZC dataset [7] consists of 980 base maps with a spatial resolution of 1m, each with a high resolution of 5000×5000 pixels. This dataset captures various land cover types, including dense forests, barren land, water bodies, agricultural fields, and human settlements.

Table 1. Quantitative comparison of retrieval approaches. Top 1, 5, 10 recall are shown on the TZC dataset, and top 1, 2, 3 recall are shown on the MTGL40-5 dataset. Re-ranking thresholds are set at 100 for TZC and 50 for MTGL40-5.

Retrieval mode	Method	TZC				MTGL40-5			
		R@1	R@5	R@10	Time(s)	R@1	R@2	R@3	Time(s)
Global retrieval	Patch-NetVLAD	63.1	63.1	63.4	0.391	58.2	60.7	62.3	0.352
	SelaVPR	68.9	69.2	69.5	0.046	66.7	66.9	67.4	0.029
	EarthLoc	70.4	71.6	72.4	0.082	67.2	67.5	68.2	0.064
Local Re-ranking	Patch-NetVLAD	65.6	66.2	66.4	21.5	62.3	64.2	64.8	10.7
	SelaVPR	72.6	73.1	73.4	0.274	70.1	71.5	72.6	0.133
	SP-SuperGlue	71.4	72.8	73.5	12.1	69.3	70.7	73.6	5.87
Local retrieval	SelaVPR	74.4	75.6	75.9	3.2×10^2	73.3	73.6	75.5	11.0
	LoFTR_Coarse	81.6	83.2	84.4	8.7×10^3	76.4	78.4	80.5	3.9×10^2
	TopicFM_Coarse	83.7	83.8	84.5	9.5×10^3	81.4	83.1	84.5	4.3×10^2
	AdaMatcher_Coarse	87.2	89.1	91.5	9.7×10^3	84.6	86.2	87.8	4.5×10^2
Asymmetric local retrieval	Our(LoFTR)	86.1	88.4	90.1	34.2	82.3	83.1	83.7	1.81
	Our(TopicFM)	88.6	90.2	92.1	42.4	86.4	88.3	90.0	2.17
	Our	90.6	93.2	93.7	1.071	91.6	93.1	93.9	0.165

MTGL40-5 dataset [31] focuses on 20 carefully selected scenes of ports and airports. Each scene has been standardized to 5120×5120 pixels, with a spatial resolution of 0.5m, and comprises five temporally distinct images. The appendix comprehensively overviews both datasets, including the generated prompt texts.

Parameter setting. We employ the Adam optimizer for model training, set the learning rate as 5.0×10^{-4} and the batch size as 2. The threshold θ is fixed at 0.2. The window size for Multi-level Fine Matching is 4. The spatial compatibility parameter Z is 10. The focal loss parameters α and γ are configured as 0.25 and 2, respectively. For the MTGL40-5 dataset, γ is reduced to 1 to account for temporal variations. Additionally, during the category and attribute distillation, temperature parameters are respectively set as $1/5$ and 1 to facilitate effective knowledge transfer.

4.2. Retrieval Performance

Quantitative comparison. We evaluate retrieval performance of various state-of-the-art methods on TZC and MTGL40-5 datasets, as shown in Table 1. The global retrieval approaches, such as Patch-NetVLAD[18], SelaVPR[29], and EarthLoc[4], exhibit minimal latency but suffer from significantly low recall rates on both datasets.

Local re-ranking strategies built upon global retrieval improve performance, where SP-SuperGlue re-ranks EarthLoc candidates by matching SuperPoint [10] features through SuperGlue [42] matcher. In the case, SelaVPR has significant gains, achieving 3.7% and 3.4% R@1 gains on TZC and MTGL40-5 datasets, respectively. Compared to similar methods, SelaVPR reduces computational overhead associated with geometric verification by leveraging matching-pair counts as similarity scores. However, these

enhancements remain constrained by the limited positive sample coverage in global retrieval candidate sets.

Moreover, we compare methods that rely solely on local features for retrieval. It is obvious that the local feature retrieval-based SelaVPR incurs a two-order-of-magnitude computational cost increase, primarily due to redundant symmetric processing, which expands 980 base images into approximately 10^5 database images. In contrast, our asymmetric-input architecture directly processes the 980 original images, significantly improving efficiency. Coarse matcher-based local retrieval approaches like LoFTR_Coarse [47], TopicFM_Coarse [15], and AdaMatcher_Coarse [20], outperform SelaVPR through fine-grained supervision but incur three-order-of-magnitude higher online retrieval latency. Furthermore, the coarse matching encoders of LoFTR and TopicFM are integrated into our framework to illustrate the effectiveness of the asymmetric-input architecture on enhancing performance and efficiency.

In addition, compared to the performance on both datasets, most methods underperform on the MTGL40-5 dataset. Since multi-source modality gaps in the MTGL40-5 dataset challenge robust features extraction, even though it contains significantly fewer base images than the TZC dataset. Conversely, our paradigm outperforms other methods on retrieval performance and efficiency on both datasets, which illustrates the effect of topic features on learning consistent and invariant representations.

4.3. Matching Performance

Quantitative comparison. Table 2 shows the matching performance and time of state-of-the-art methods. To avoid retrieval preference, positive samples from SelaVPR’s lo-

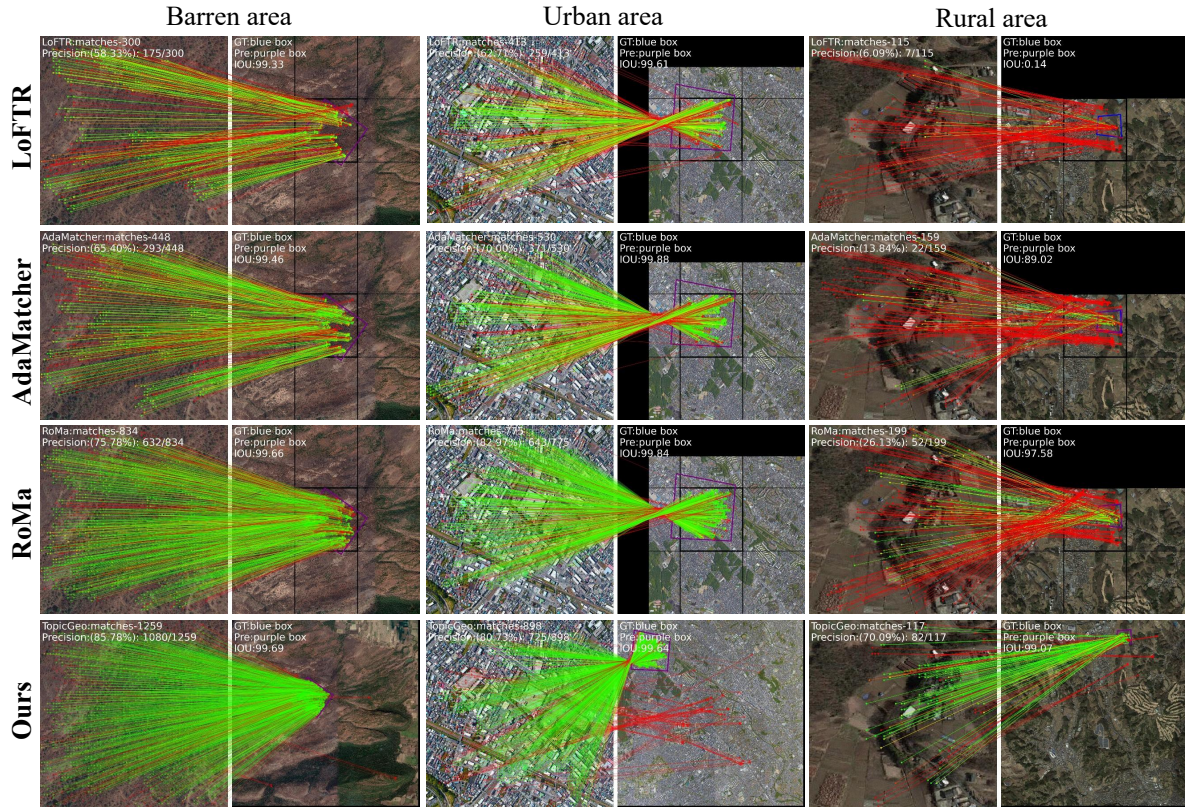


Figure 3. A visualized comparison of the state-of-the-art approaches on TzC dataset. Our unified framework generates numerous accurate correspondences in both texture-scarce barren areas and texture-dense urban areas. For images with large resolution-scaled discrepancies, our approach matches regions that are easier to match although they are sparse.

Table 2. Comparison of matching approaches. $n = 1$ represents the original matching, while $n = 4$ indicates four iterations of the EarthMatch protocol. **IoU** is used as an effective evaluation metric.

Category	Method	TzC		MTGL		Time(s)	
		$n=1$	$n=4$	$n=1$	$n=4$	$n=1$	$n=4$
Detector-based	SIFT-SuperGlue	82.4	89.5	70.6	78.6	0.24	2.81
	SP-SuperGlue	84.4	91.2	77.4	82.4	0.14	2.37
Dense matcher	RoMa	92.2	96.1	86.7	90.2	0.79	5.17
Detector free	LoFTR	85.7	90.3	77.3	84.1	0.12	2.18
	AdaMatcher	91.4	93.7	83.3	88.4	0.15	2.51
	TopicFM	89.2	92.4	78.5	84.5	0.13	2.32
	Ours	98.9	99.1	91.1	92.7	0.09	2.02

cal re-ranking are used for matching, while the EarthMatch [3] protocol is applied to iteratively optimize matching and compensate for the limited receptive field of other methods. It is obvious that our method achieves state-of-the-art performance on both protocols while significantly improving

efficiency. The proposed asymmetric-input direct matching framework reduces computational time by 1/20 of the EarthMatch protocol with merely 1.6% accuracy loss on TzC dataset. Notably, while recent dense matcher RoMa [12] delivers competitive accuracy (90.2% on the MTGL dataset), it requires $2.5\times$ longer for per match than our solution. The multi-temporal discrepancies in MTGL dataset pose greater challenges than TzC’s synthetic data, where our matching method effectively handles heterogeneous differences through consistent semantics and adaptive matching mechanisms.

Visual comparison. Figure 3 displays the matching results of the crucial first iteration under the EarthMatch protocol. Notably, our method searches over a broader range compared to others. It is seen that our method excels in texture-scarce (such as barren areas) and texture-dense (like urban regions) scenarios through adaptive refinement and relaxed matching constraints in MKNN. Under extreme scale variations, existing methods like LoFTR distribute numerous matching pairs on non-discriminative natural terrains, which leads to low matching rates and disruption of iter-

Table 3. Ablation of topic-related cross-attention layers and weighted loss on TZC dataset. The baseline, C, and T are the self-attention encoder, the cross-attention layers in LoFTR, and the topic-related layers of TopicFM, respectively.

Base	C/T	Category	Attribute	WLoss	R@1	Time(s)
✓					82.4	0.87
	✓				86.1/88.6	34/42
		✓			85.8	0.94
		✓	✓		88.1	1.07
		✓	✓	✓	90.6	1.07

Table 4. Ablation of topic-independent components on TZC dataset. A baseline is used that incorporates the topic encoder, original adaptive label assignment [20], and expectation-based fine-grained matcher [47].

Base	ACA	MKNN	GGV	MFM	R@1	IoU	Pairs	Time(s)
✓					84.9	72.3	155	0.94/0.07
	✓				86.5	75.4	171	0.94/0.07
	✓				86.5	80.7	171/192	0.94/0.08
	✓			✓	90.6	85.1	171/192	1.07/0.08
	✓			✓	87.1	81.2	192	1.07/0.08
	✓			✓	90.6	90.5	171/814	1.07/0.09

ative process. Whereas our topic-aware weighted loss encourages the model to match regions that are easier to match although they are sparse, thereby enhancing reliability.

4.4. Ablation Study

The effect of topics. Our method uses distilled topics to improve efficiency and performance. Therefore, we separately evaluate the effect of the category and attribute, and the derived category weighted loss on the TZC dataset, as illustrated in Table 3. The cross-image perception of LoFTR improves $R@1$ by 3.7%, while TopicFM’s self-supervised topic achieves 6.2% gains. Both encoders that require cross-attention for information interaction cause $40\times$ slower online retrieval than the baseline (0.87s). Our category embeddings brings 3.4% improvement, attribute embeddings adds 2.7%, and weight loss contributes 2.5%. Moreover, the lightweight parallel architecture avoids significant time consumption. Figure 4 further visually compares topic semantics of remote sensing land cover generated by our method and the self-supervised approach TopicFM. It is seen that our method’s deterministic land-cover topic modeling better distinguishes terrain features than self-supervised approach, enhancing geolocation precision.

The effect of other modules. Our evaluation first examines adaptive center assignment (ACA) and MKNN modules, followed by global geometric verification (GGV) and multi-level fine matching (MFM). During testing, the retrieval and matching is in sequence: failed retrieval yields zero matching IoU. When GGV is disabled, match count

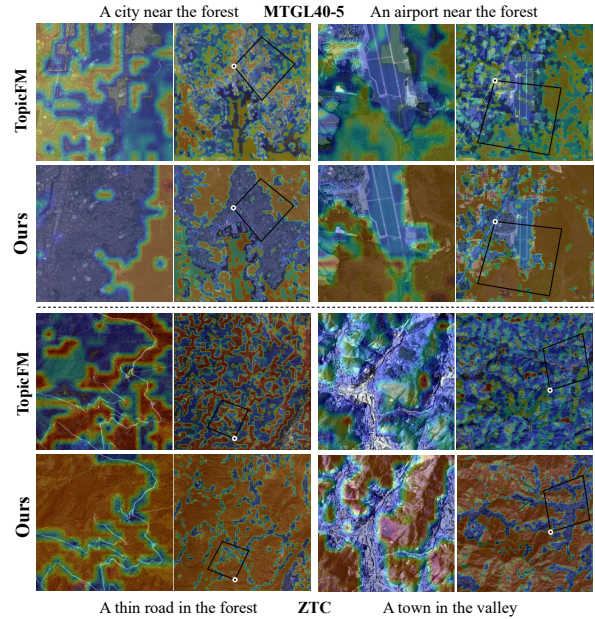


Figure 4. Visualization of category topics in remote images. Our human-perception-aligned topic distributions encourage multi-source feature alignment for geospatial retrieval and matching, whereas self-supervised structural modeling struggles to adapt to complex remote sensing scenarios.

directly determines retrieval scores. The two values of MKNN correspond to the k value in retrieval and coarse matching. As Table 4 shows, ACA increases initial matching pairs by 10.3% over the original adaptive top-left assignment. MKNN ($k=1$) further elevates this by 12.3%. The GGV enhances retrieval accuracy by 4.1% with only an additional time consumption (0.13s). One-to-many with $k=1$ during retrieval makes the GGV score unstable and the recall rate decreases. MFM amplifies matching pairs by $5\times$ with 5.4% IoU improvement, while maintaining low time overhead. Crucially, correctly retrieved samples achieve 99% (90.5(matching)/90.6(retrieval)) average matching IoU, indicating the task-coherent feature learning of retrieval and matching tasks.

5. Conclusion

We introduced a unified asymmetric-input retrieval-matching framework for geolocation. By combining object semantic modeling with image retrieval and matching, our method learns a robust set of primary features that can be stored offline. Moreover, we design a center-based label adaptive assignment and a multi-level fine-grained local matching pipeline, optimizing the matching precision. Compared to state-of-the-art methods, the proposed approach is both efficient and interpretable.

6. Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No.62372358 and 62302355, the Fundamental Research Funds for the Central Universities under Grant No.XJSJ24071 and No.XJSJ24072, and the Xidian University Specially Funded Project for Interdisciplinary Exploration under Grant No. TZJH2024026.

References

- [1] Artem Babenko and Victor Lempitsky. Aggregating local deep features for image retrieval. In *Proceedings of the IEEE international conference on computer vision*, pages 1269–1277, 2015. [2](#)
- [2] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008. [2](#), [3](#)
- [3] Gabriele Berton, Gabriele Goletto, Gabriele Trivigno, Alex Stoken, Barbara Caputo, and Carlo Masone. Earthmatch: Iterative coregistration for fine-grained localization of astronaut photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4264–4274, 2024. [2](#), [3](#), [7](#)
- [4] Gabriele Berton, Alex Stoken, Barbara Caputo, and Carlo Masone. Earthloc: Astronaut photography localization by indexing earth from space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12754–12764, 2024. [1](#), [2](#), [6](#)
- [5] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. Brief: Binary robust independent elementary features. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5–11, 2010, Proceedings, Part IV 11*, pages 778–792. Springer, 2010. [2](#), [3](#)
- [6] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16911, 2024. [2](#)
- [7] "Tianzhi Cup" Artificial Intelligence Challenge Organizing Committee. zhihuidiqiu2024, 2024. [5](#)
- [8] Rodrigo Caye Daudt, Bertrand Le Saux, Alexandre Boulch, and Yann Gousseau. Multitask learning for large-scale semantic change detection. *Computer Vision and Image Understanding*, 187:102783, 2019. [2](#)
- [9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. [2](#)
- [10] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. [3](#), [6](#)
- [11] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. [4](#)
- [12] Johan Edstedt, Qiyu Sun, Georg Bökman, Márten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19790–19800, 2024. [3](#), [7](#)
- [13] JR Elliott. Earth observation for the assessment of earthquake hazard, risk and disaster management. *Surveys in geophysics*, 41(6):1323–1354, 2020. [1](#)
- [14] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. [3](#)
- [15] Khang Truong Giang, Soohwan Song, and Sungho Jo. Topicfm: Robust and interpretable topic-assisted feature matching. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2447–2455, 2023. [2](#), [3](#), [6](#)
- [16] Khang Truong Giang, Soohwan Song, and Sungho Jo. Topicfm+: Boosting accuracy and efficiency of topic-assisted feature matching. *IEEE Transactions on Image Processing*, 2024. [2](#)
- [17] Pierre Gleize, Weiyao Wang, and Matt Feiszli. Silk: Simple learned keypoints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22499–22508, 2023. [3](#)
- [18] Stephen Hausler, Sourav Garg, Ming Xu, Michael Milford, and Tobias Fischer. Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14141–14152, 2021. [2](#), [6](#)
- [19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. [4](#)
- [20] Dihe Huang, Ying Chen, Yong Liu, Jianlin Liu, Shang Xu, Wenlong Wu, Yikang Ding, Fan Tang, and Chengjie Wang. Adaptive assignment for geometry aware local feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5425–5434, 2023. [3](#), [4](#), [5](#), [6](#), [8](#)
- [21] Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. Cotr: Correspondence transformer for matching across images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6207–6217, 2021. [3](#)
- [22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. [5](#)
- [23] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-perfect structure-from-motion with featuremetric refinement. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5987–5997, 2021. [3](#)
- [24] Chao Liu, Jingjing Ma, Xu Tang, Fang Liu, Xiangrong Zhang, and Licheng Jiao. Deep hash learning for remote

- sensing image retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 59(4):3420–3443, 2020. 1
- [25] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiacong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 2, 4
- [26] Ganchao Liu, Chao Li, Sihang Zhang, and Yuan Yuan. Vlmfl: Uav visual localization based on multi-source image feature learning. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 1
- [27] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017. 2
- [28] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 2, 3
- [29] Feng Lu, Lijun Zhang, Xiangyuan Lan, Shuting Dong, Yaowei Wang, and Chun Yuan. Towards seamless adaptation of pre-trained models for visual place recognition. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2, 3, 6
- [30] Christian Lülfi, Denis Mayr Lima Martins, Marcos Antonio Vaz Salles, Yongluan Zhou, and Fabian Gieseke. Clip-branches: Interactive fine-tuning for text-image retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2719–2723, 2024. 2
- [31] Jingjing Ma, Shiji Pei, Yuqun Yang, Xu Tang, and Xiangrong Zhang. Mtgl40-5: A multi-temporal dataset for remote sensing image geo-localization. *Remote Sensing*, 15(17):4229, 2023. 1, 2, 6
- [32] Xianping Ma, Xiaokang Zhang, and Man-On Pun. Rs 3 mamba: Visual state space model for remote sensing image semantic segmentation. *IEEE Geoscience and Remote Sensing Letters*, 2024. 2
- [33] Xianping Ma, Xiaokang Zhang, Man-On Pun, and Ming Liu. A multilevel multimodal fusion transformer for remote sensing semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 2
- [34] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*, 2022. 2
- [35] Florent Perronnin, Yan Liu, Jorge Sánchez, and Hervé Poirier. Large-scale image retrieval with compressed fisher vectors. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3384–3391. IEEE, 2010. 2
- [36] Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Yuxing Peng, and Kai Xu. Geometric transformer for fast and robust point cloud registration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11143–11152, 2022. 5
- [37] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1655–1668, 2018. 2
- [38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2
- [39] Maryam Rahnemoonfar, Tashnim Chowdhury, Argho Sarkar, Debvrat Varshney, Masoud Yari, and Robin Robertson Murphy. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*, 9:89644–89654, 2021. 2
- [40] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2d2: Reliable and repeatable detector and descriptor. *Advances in neural information processing systems*, 32, 2019. 3
- [41] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12716–12725, 2019. 3
- [42] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 3, 5, 6
- [43] Yanqing Shen, Sanping Zhou, Jingwen Fu, Ruotong Wang, Shitao Chen, and Nanning Zheng. Structvpr: Distill structural knowledge with weighting samples for visual place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11217–11226, 2023. 2
- [44] Edgar Simo-Serra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, Pascal Fua, and Francesc Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *Proceedings of the IEEE international conference on computer vision*, pages 118–126, 2015. 3
- [45] Alex Stoken and Kenton Fisher. Find my astronaut photo: Automated localization and georectification of astronaut photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6196–6205, 2023. 2, 3
- [46] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 4
- [47] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8918–8927, 2021. 1, 3, 4, 5, 6, 8
- [48] Xu Tang, Qiushuo Ma, Xiangrong Zhang, Fang Liu, Jingjing Ma, and Licheng Jiao. Attention consistent network for remote sensing scene classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:2030–2045, 2021. 1

- [49] Haoyang Wang, Fuhui Zhou, and Qihui Wu. Accurate vision-enabled uav location using feature-enhanced transformer-driven image matching. *IEEE Transactions on Instrumentation and Measurement*, 2024. 2, 3
- [50] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021. 2, 4
- [51] Shuhang Wang, Jin Zheng, Hai-Miao Hu, and Bo Li. Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE transactions on image processing*, 22(9):3538–3548, 2013. 4
- [52] Xiang Wang, Shiwei Zhang, Jun Cen, Changxin Gao, Yingya Zhang, Deli Zhao, and Nong Sang. Clip-guided prototype modulating for few-shot action recognition. *International Journal of Computer Vision*, 132(6):1899–1912, 2024. 2
- [53] Qinghua Ye, Yuzhe Wang, Lin Liu, Linan Guo, Xueqin Zhang, Liyun Dai, Limin Zhai, Yafan Hu, Nauman Ali, Xinhui Ji, et al. Remote sensing and modeling of the cryosphere in high mountain asia: A multidisciplinary review. *Remote Sensing*, 16(10):1709, 2024. 1
- [54] Hongsheng Zhang and Ru Xu. Exploring the optimal integration levels between sar and optical data for better urban land cover mapping in the pearl river delta. *International journal of applied earth observation and geoinformation*, 64:87–95, 2018. 1
- [55] Yesheng Zhang and Xu Zhao. Mesa: Matching everything by segmenting anything. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20217–20226, 2024. 2