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

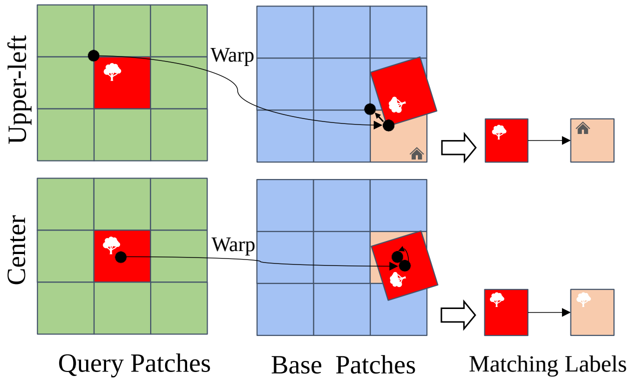

Figure 1. Visualized comparison between AdaMatcher's top-left corner-based label adaptive assignment and our proposed center-based label adaptive assignment (ACA) strategy.

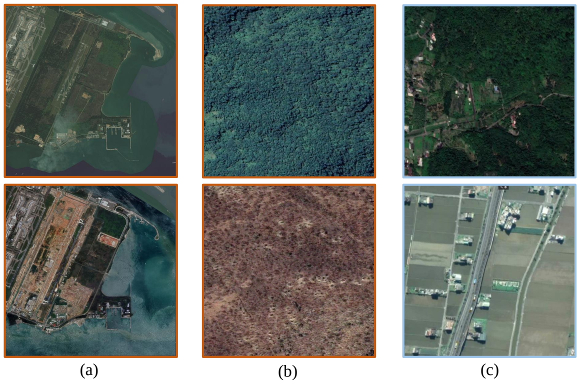

Figure 2. Challenges in remote sensing images. (a) Multi-source heterogeneity. (b) Indistinctive textures. (c) Spatial distribution imbalance of land cover types.

1. Network Details

We employ a lightly modified ResNet50 as our backbone network. Note that CLIP is solely used as an external tool for offline embedding extraction and is excluded during inference. As LoFTR [5] shares structural similarities with our approach, we provide a detailed comparison between the two methods in Table 1.

Input: Unlike LoFTR, we employ stacked 4×4 patches for parameter-free 4× downsampling. The deeper downsampling strategy of LoFTR requires an additional feature extraction stage, yet our progressive channel increase (starting with fewer channels) ensures computational efficiency at the current stage.

Multi-source feature interaction: Our topic-based retrieval and coarse matching module replaces LoFTR's interleaved self-cross attention with a more efficient manner. For multi-level fine matching, we adopt a 256-channel architecture fused with features from preceding stages, mitigating challenges posed by heterogeneous multi-source image variations.

Parameters: Despite having 9.03M more parameters than LoFTR (11.6M vs. 20.63M), 10.5M parameters originate from the topic-based cross-attention mechanism. This parameter increase is primarily attributed to the 512-dimensional input encoding, as CLIP[3] typically utilizes higher-dimensional encodings to preserve richer semantic information. As shown in Table 3 of ablation studies in the manuscript, these topic-related parameters are allocated to highly efficient and high-reward processes, while the overall architecture remains optimized for retrieval tasks.

Matching label assignment: Building upon AdaMatcher's adaptive label assignment [2], we improve the traditional top-left corner-based label assignment by adopting a center-based label adaptive assignment (ACA) approach, shown in Figure 1. This modification enhances robustness against large rotational misalignments inherent in remote sensing imagery.

2. Data Description

Challenges in Data Accurately localizing objects across multi-source remote sensing images pose significant technical challenges. As illustrated in Figure 2, we present three representative cases spanning two datasets: the TZC

[*] Corresponding author

Table 1. The detailed architectures of the proposed method (TopicGeo) and the most related baseline (LoFTR). Our design improves the computational efficiency for dual-task: i) feature extraction, ii) coarse matching, and iii)multi-level fine matching. Note that a ResBlock includes two convolutional blocks along with a residual connection. K, S, C, SA, and CA are denoted for the kernel size, stride, size of output channel, self-attention, and cross-attention, respectively. Note that fine encoder 2 utilizes the upsampled features of ${F_1}^{out}$.

Stage	**TopicGeo**	**Output size**	**Stage**	**LoFTR**	**Output size**
F_1	Reshape($\frac{H}{4}$, $\frac{W}{4}$)+Conv[K : 7, S : 2, C : 32]+GN+ReLU [ResBlock[K:3,S:1,C:32]+GN+RLU]$_{\times 2}$	$\frac{H}{8} \times \frac{W}{8} \times 32$	F_1	Conv[K : 7, S : 2, C : 128]+BN+ReLU [ResBlock[K:3,S:1,C:128]+BN+RLU]$_{\times 2}$	$\frac{H}{2} \times \frac{W}{2} \times 128$
F_2	ResBlock[K:3,S:2,C:64]+GN+RLU ResBlock[K:3,S:1,C:64]+GN+RLU	$\frac{H}{16} \times \frac{W}{16} \times 64$	F_2	ResBlock[K:3,S:2,C:196]+BN+RLU ResBlock[K:3,S:1,C:196]+BN+RLU	$\frac{H}{4} \times \frac{W}{4} \times 196$
F_3	ResBlock[K:3,S:2,C:128]+GN+RLU ResBlock[K:3,S:1,C:128]+GN+RLU	$\frac{H}{32} \times \frac{W}{32} \times 128$	F_3	ResBlock[K:3,S:2,C:256]+BN+RLU ResBlock[K:3,S:1,C:256]+BN+RLU	$\frac{H}{8} \times \frac{W}{8} \times 256$
${F_4}^{out} = F_c$	ResBlock[K:3,S:2,C:256]+GN+RLU ResBlock[K:3,S:1,C:256]+GN+RLU	$\frac{H}{64} \times \frac{W}{64} \times 256$	${F_3}^{out} = F_c$	Conv[K : 3, S : 1, C : 256]	$\frac{H}{8} \times \frac{W}{8} \times 256$
${F_3}^{\text{out}}$	$\text{Conv}[K:1, S:1, C:256](F_3) \oplus \text{Up}({F_4}^{\text{out}})$ $\text{Conv}[K:3, S:1, C:256] + \text{GN} + \text{LReLU}$	$\frac{H}{32} \times \frac{W}{32} \times 256$	${F_2}^{\text{out}}$	$\text{Conv}[K:3, S:1, C:256](F_2) \oplus \text{Up}({F_3}^{\text{out}})$ $\text{Conv}[K:3, S:1, C:256] + \text{BN} + \text{LReLU}$ $\text{Conv}[K:3, S:1, C:196]$	$\frac{H}{4} \times \frac{W}{4} \times 196$
${F_2}^{out} = F_a$	$\text{Conv}[K:1, S:1, C:256](F_2) \oplus \text{Up}({F_3}^{out})$ $\text{Conv}[K:3, S:1, C:256] + \text{GN} + \text{LReLU}$	$\frac{H}{16} \times \frac{W}{16} \times 256$	${F_1}^{out} = F_f$	$\text{Conv}[K:3, S:1, C:196](F_1) \oplus \text{Up}({F_2}^{out})$ $\text{Conv}[K:3, S:1, C:196] + \text{BN} + \text{LReLU}$ $\text{Conv}[K:3, S:1, C:128]$	$\frac{H}{2} \times \frac{W}{2} \times 128$
${F_1}^{out}$	$\text{Conv}[K:1, S:1, C:256](F_1) \oplus \text{Up}({F_2}^{out})$ $\text{Conv}[K:3, S:1, C:256] + \text{GN} + \text{LReLU}$	$\frac{H}{8} \times \frac{W}{8} \times 256$	-	-	-
Topic module	$Linear[C:512](F_c)$ $[SA[C:256, \text{head}:8](F_c, F_c)]_{\times 4}$ $[CA[C:512, \text{head}:8](F_c, F_{prompts})]_{\times 4}$ $Linear[C:256](concat(F_{SA}, F_{CA}))$	$\frac{H}{64} \times \frac{W}{64} \times 256$	Coarse encoder	$\begin{bmatrix} SA[C:256, \text{head}:8](F_c, F_c) \\ CA[C:256, \text{head}:8](F_c^A, F_c^B) \end{bmatrix}_{\times 4}$	$\frac{H}{8} \times \frac{W}{8} \times 256$
Fine encoder 1	$SA[C:256, \text{head}:8]$ $CA[C:256, \text{head}:8]$	$\frac{H}{16} \times \frac{W}{16} \times 256$	Fine encoder	$SA[C:128, \text{head}:8]$ $CA[C:128, \text{head}:8]$	$\frac{H}{2} \times \frac{W}{2} \times 128$
Fine encoder 2	$SA[C:256, \text{head}:8]$ $CA[C:256, \text{head}:8]$	$\frac{H}{4} \times \frac{W}{4} \times 256$	-	-	-

dataset and the MTGL40-5 dataset. These challenges manifest as follows: (1) **Multi-source heterogeneity**: Substantial appearance discrepancies caused by varying imaging conditions, acquisition times, and sensor modalities. (2) **Indistinctive textures**: Large homogeneous regions with insufficient distinctive features for reliable feature extraction and matching. (3) **Spatial distribution imbalance of land cover types**: Highly skewed land-cover type distributions that hinder balanced feature learning and robust matching.

Image dataset setup For the TZC dataset[1], we divide all base maps into training, validation, and testing sets. During testing, all base maps in TZC dataset serve as the search space. For the MTGL-40-5 dataset[4], we follow the protocol of method [4]: base maps with ID 0, 1, and 2 are used for training, while ID 3 and 4 are used for testing, with ID 3 specifically serving as the search space.

To effectively simulate the geometric transformation between images in real-world scenarios, for both datasets, query images are randomly cropped from the base maps in both the training and testing sets. Specifically, the cropping scale is randomly selected between 341 and 3072 pixels, followed by a perspective transformation and a random rotation between 0° and 360°. The final cropped regions are resized to a fixed resolution of 1024×1024, and the corresponding homography matrices are saved. For the TZB dataset, we further apply color augmentation and introduce Gaussian noise with an intensity randomly sampled between 0 and 0.1.

Finally, we generate 20,580 training query images and 588 testing query images from the respective base maps in TZC dataset, and 12,000 training query images and 522 test query images in MTGL-40-5 dataset. Notably, we filter out textureless query images, which is particularly common in the MTGL-40-5 dataset due to the presence of large water bodies with no discernible texture.

Data setup for other methods To ensure a fair comparison and acknowledge the limitations of other methods in aggregating information from the entire basemap at once, we adopted standard preprocessing strategies for basemap data. Specifically, the base maps are divided into image

Table 2. Topic Prompts

Category Prompts	Attribute Prompts
Building Road River Barren Forest Farmland	Grey, White, Green, Golden, Yellow, Blue, Red, Black, Tall, Long, Wide, Vast Thick, Narrow, Dense, Clear, Dusty, Large, Small, Agricultural, Barren, Shimmering, A, Few, Some

patches with a 25% area overlap, which are then used for retrieval and matching. Furthermore, one-to-one positive image patches are designated for training both the retriever and the matcher, while evaluating retrieval efficacy via map-wide recall metrics. During the retrieval, methods without geometric validation use the classic RANSAC approach.

Topic prompts As introduced in our methodology, we establish two distinct types of topic prompts: *Category Prompts* and *Attribute Prompts*. Including categories of small-scale and variational objects, like cars and people, may introduce noise. Instead, our method extracts stable and common topics statistically derived from a large-scale remote sensing dataset[7]. The attribute generation process employs a data-driven strategy, avoiding manual bias. The GPT-4 is first applied to generate comprehensive visual descriptions of these land covers. Then, descriptions are linguistically refined into descriptive keywords (as attribute prompts) of typical geographical features. The generated prompts are shown in Table 2. Our taxonomy design illustrates three key characteristics: 1) *Semantic granularity adaptation* through separate categorical and descriptive prompts, 2) *Data-driven description generation* via large language model analysis, and 3) *Domain-specific adaptation* through customizable topic sets. This structured approach enables effective knowledge representation while maintaining flexibility for various remote sensing applications.

3. Parameters Details

Geometric consistency provides a reliable confidence measure for retrieval, which involves the truncation parameter Z. It also depends on the distance compatibility parameter δ_d, the angular compatibility parameter δ_c, and the sparsity weight λ. Due to the large number of parameters, finding the optimal values is challenging. Therefore, we adopt empirical values based on the data distribution. The parameter δ_d is set to 0.2, meaning that a length ratio variation within ± 0.2 is considered normal. Similarly, δ_c is set to 0.175, corresponding to $\pi/18$. The truncation parameter Z is set to 10 to prevent excessive confidence decay. Finally, λ is set to 0.2.

Table 3. Effect of different topics on retrieval and matching performance on TZC dataset.

Category	R@1	IoU	Attribute	R@1	IoU
All	**90.6**	**90.5**	100%	**90.6**	**90.5**
Building	88.7	83.3	80%	90.1	88,9
Road	87.9	82.1	60%	89.2	86.6
River	87.2	81.0	40%	88.5	85.5
Barren	85.1	77.4	20%	88.1	85.2
Farmland	86.7	79.8	12%	88.0	85.2
Forest	85.7	77.9	8%	88.0	84.9

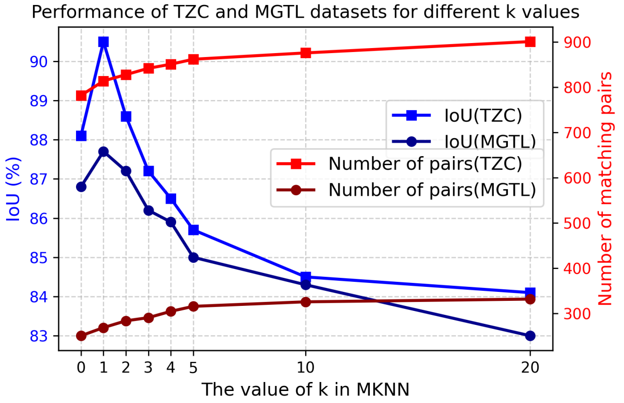

Figure 3. The relationship between the matching index (IoU), the matches number, and the parameter k of MKNN. When k=1, the matching index IoU achieves its optimal value. As k increases, the number of matching pairs grows. However, it also introduces more mismatches, leading to a decline in overall matching performance.

Since our method employs center-based label assignment, the minimum window size is inherently set to an even number (4) during the multi-level fine matching stage, rather than the odd number (5) used in top-left label assignment. This further reduces computational overhead.

4. Ablation Study

CLIP encoders We compare the RemoteCLIP [3] with RS vision language models, SkyScript [8], where base and larger pre-trained models (CLIP ViT-B32 and ViT-L14) are used. Table 4 illustrates that for both CLIP models, the larger version provides more sufficient semantics. However, most metrics achieved by the base and larger versions of RemoteCLIP are relatively higher. Moreover, the performance gap between base and larger versions of RemoteCLIP is more smaller, which shows its robustness.

Topic selection TopicGeo achieves higher performance by estimating the topics through the interaction of category embeddings and attribute embeddings. Consequently, the selection of these embedding schemes directly governs the overall efficacy. To investigate category and attribute embedding influences, we evaluate TopicGeo under different category embeddings and varying amounts of attribute embeddings. Table 3 presents the retrieval and matching performance of a specific category embedding on TZC testing dataset, as well as the effect of selecting different proportions of detail embeddings at random. The results show that ”building” has the most significant impact on performance, followed by ”road”, while other categories exhibit comparatively marginal effects. Furthermore, the larger the number of attribute topics, the better the performance. The overall performance monotonically improves with increasing embeddings granularity. Note that our model dynamically prioritizes important and robust matching semantics through topic distillation and semantic matching during training. The model gradually becomes resilient to certain unavoidable noise.

The parameter k of MKNN In our approach, adaptive center assignment of matching labels optimizes the pipeline. Intuitively, this assignment enables one-to-many matching, which mitigates resolution-scale discrepancies' detrimental effects. However, it is important to note that not all matching candidates are equally valuable, as their corresponding patches differ in overlap area, center distance, and multi-source feature characteristics. Fortunately, this parameter is independent of training, allowing for adjustments during testing. In the adaptive local window matching and retrieval stages, we set $k = 0$ since, at a fine-grained level, there are already sufficient matching candidates, and both stages benefit from higher precision. In the coarse matching stage, the choice of k is more flexible. As shown in Figure 3, our experiments on two datasets indicate that setting $k = 1$ yields the highest IoU, leading to better homography estimation. However, as k increases, the filtering effect weakens. While a larger k introduces more match pairs, it also introduces additional noise, ultimately degrading the quality of the homography estimation. Essentially, $k = 1$ permits unrestricted correspondence between a query patch and patches within a 3×3 window on the base map, signifying effective match filtering capability under $3\times$ resolution scale discrepancies between query and base imagery.

Image size We evaluate the impact of image size on the performance of various retrieval and matching methods using the MTGL40-5 dataset. For the retrieval, we employ center-cropped resizing to standardize image dimensions, while for the matching, we directly scale the images and compare the matching performance. As shown in Figure 4 (left), as the image size decreases, the performance of local feature-based retrieval methods gradually declines due to the reduced number of matching pairs. In contrast, global feature retrieval initially benefits from a more comprehensive representation, but ultimately deteriorates due to insufficient information. As shown in Figure 4 (right), the matching accuracy of most models decreases as the image size reduces. While our method demonstrates inferior performance to RoMa-like models at the standard 512×512 resolution, it achieves notably superior accuracy on high-resolution imagery, aligning with the resolution demands predominant in real-world applications.

Table 4. Performance comparison of different CLIP encoders.

CLIP Encoder	**ZTC**			**MTGL40-5**		
	R@1	R@5	R@10	R@1	R@2	R@3
RemoteClip (B)	90.6	93.2	93.7	91.6	93.1	93.9
RemoteClip (L)	**90.8**	**93.5**	**93.9**	91.7	93.3	**94.8**
SkyScript (B)	89.8	91.8	92.0	92.2	92.7	93.2
SkyScript (L)	90.5	92.3	93.4	**92.2**	**93.7**	94.4

Table 5. Resource consumption analysis (Training & Testing)

Model	**Training**	**Testing**	
	Mem(GB)	**FLOPs (G)**	**Mem (GB)**
LoFTR(1024)	23.2	192	10.2
Ours(1024)	16.8	123	7.2
LoFTR(512)	10.4	94	4.6
Ours(512)	8.6	71	3.4

5. Resource Consumption

As previously discussed, our method balances efficiency and recall during the retrieval stage. While high-resolution base maps may inherently incur elevated resource demands during the matching process, we explicitly address this potential concern through assessing resource consumption. The experiments are conducted on a workstation equipped with an Intel Core i7-12700KF CPU, 64GB RAM, and an Nvidia GeForce RTX 4090 GPU with 24GB of VRAM. We measure the computational load and memory consumption of our method and LoFTR when processing query images at resolutions of 1024 and 512 on a single workstation. As evidenced in Table 5, our architecture sustains minimized memory allocation and reduced computational overhead throughout both training and inference phases. This efficiency originates from the absence of cross-image dependency constraints and our multi-stage downsampling mechanism.

6. Method Extension

Generalization Our topic distillation introduces high-level semantics of objects, independent of visual information,

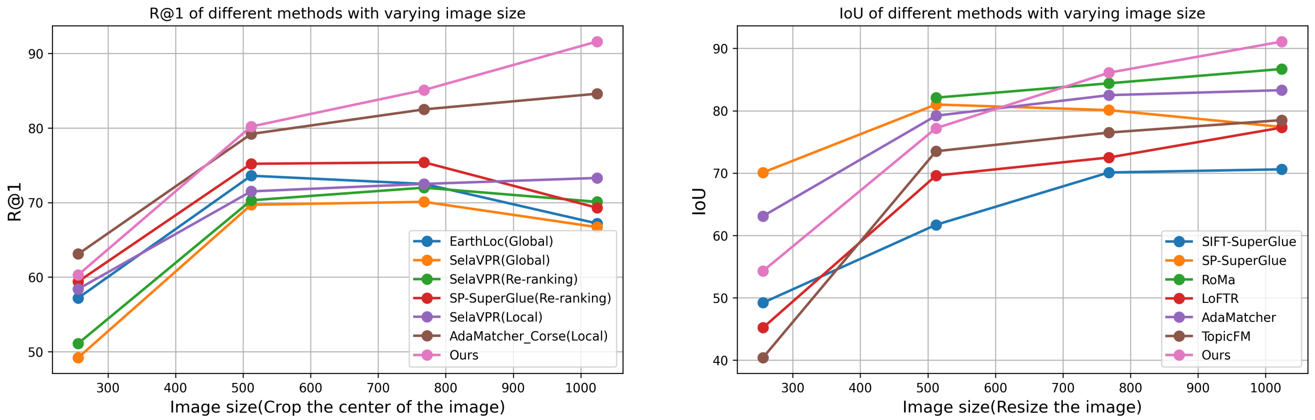

Figure 4. Changing image size on MTGL40-5. Image size variation has a significant impact on the retrieval performance $R@1$ and matching performance IoU of each method. Our method is more advantageous at higher resolutions.

into the retrieval-matching network, thereby enhancing generation performance. To evaluate the generalization, we conduct experiments on a more challenging multi-temporal aerial dataset, Hi-UCD [6], with a spatial resolution of 0.1m. The image is processed as a large size of 5120×5120 pixels. Firstly, we train SOTA models on the Hi-UCD dataset (w/Training) to evaluate the generation for different datasets (experimental settings are consistent with the satellite datasets). To further evaluate the generalization for cross-domain, we directly use models trained on the MTGL40-5 dataset to test the Hi-UCD testing dataset (w/o Training). Table 6 quantitatively verifies generation performance on cross-domain scenarios even without fine-tuning, comparing retrieval metrics and the average IoU of R@1 across all test samples.

Table 6. The comparison of SOTAs on the Hi-UCD dataset.

Method	**w/ Training**				**w/o Training**			
	R@1	R@2	R@3	IoU	R@1	R@2	R@3	IoU
AdaMatcher	80.1	81.5	82.4	65.7	68.4	69.8	71.3	51.1
Ours(TopicFM)	83.7	85.8	87.2	73.7	72.8	73.4	75.6	62.6
Ours	**86.2**	**87.4**	**88.1**	**79.6**	**80.2**	**82.6**	**83.5**	**71.4**

Larger Scenes The proposed scale-extended and performance-unbiased strategies enable to balance performance and hardware efficiency in larger scenes than training: 1) Our asymmetric processing allows large-patch and low-redundancy cropping (far lower than the typical 50% overlap) of the reference map. 2) The proposed retrieval-matching coupled structure allows coarse retrieval to first identify regions of interest (RoIs), followed by fine-grained matching focused solely on these RoI areas, thereby reducing fine-grained feature extraction operations over large irrelevant areas.

7. Visualization Results

Matching visualization Figure 5 visually illustrates more matching results of TopicFM and our proposed TopicGeo on the MTGL40-5 dataset under various difficult matching scenarios. The EarthMatch protocol is employed, and results from the first iteration are shown. Red and green lines indicate incorrect and correct matches, respectively. It is seen that compared to TopicFM, our method achieves a significantly higher number of correct matches and superior accuracy.

Topic visualization Figure 6 and Figure 7 display topic visualizations across different scenarios on the MTGL40-5 and TZC datasets, respectively. Our prompt learning- and distillation-based approach effectively extracts topics that align well with visual features. Compared to existing self-supervised learning-based approach TopicFM, our method demonstrates higher semantic consistency between the query image and the reference area of the base image.

References

[1] ”Tianzhi Cup” Artificial Intelligence Challenge Organizing Committee. zhihuidiqiu2024, 2024. 2

[2] 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. 1

[3] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong 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. 1, 3

[4] Jingjing Ma, Shiji Pei, Yuqun Yang, Xu Tang, and Xiangrong Zhang. Mtgl40-5: A multi-temporal dataset for remote sens-

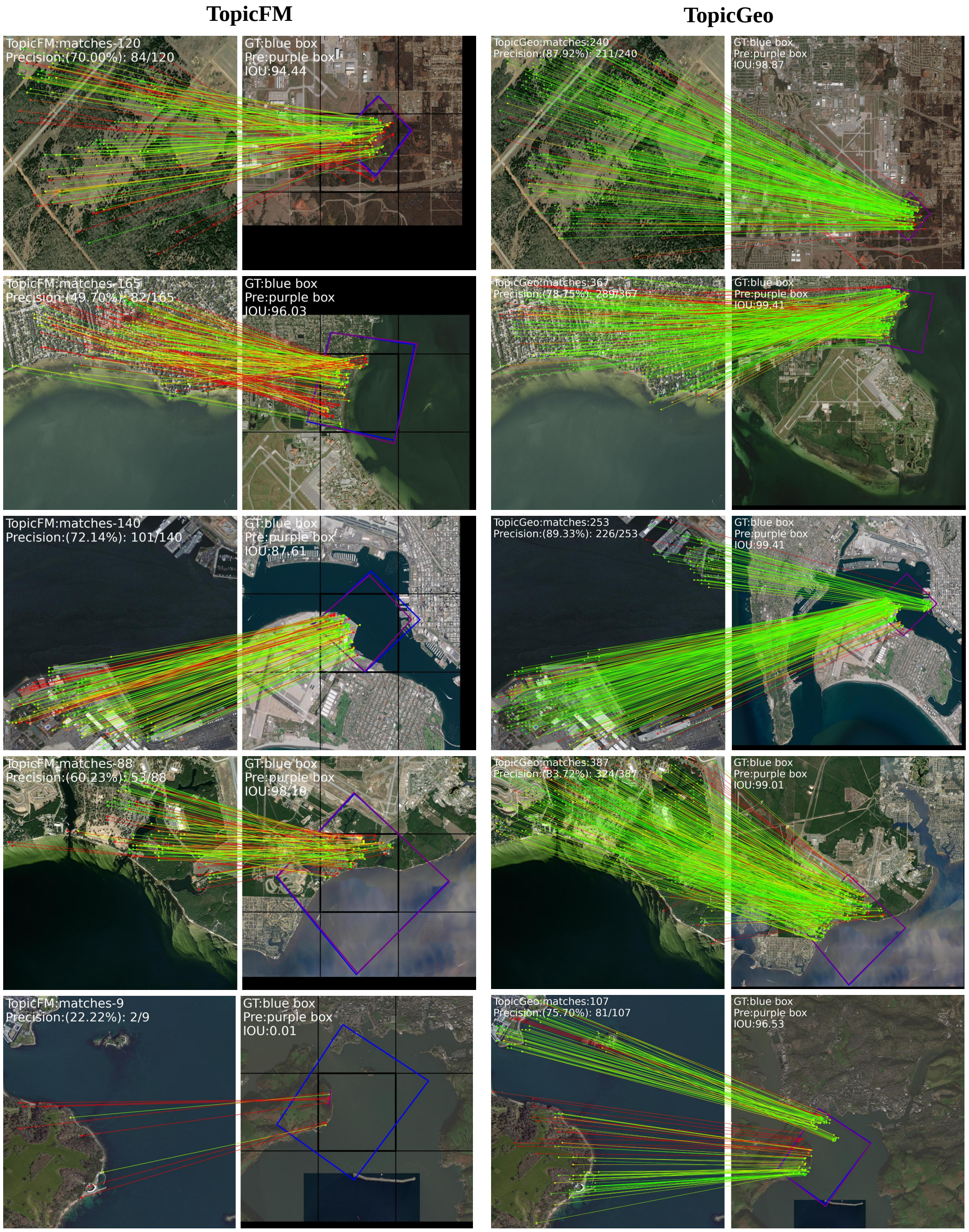

Figure 5. The visualized matching comparison between existing TopicFM and our TopicGeo on the MTGL40-5 dataset under challenging scenarios. The first iteration of the EarthMatch protocol is shown, where red and green lines denote incorrect and correct matches, respectively. Our approach outperforms TopicFM by simultaneously increasing the number of matches and boosting valid matches.

ing image geo-localization. *Remote Sensing*, 15(17):4229, 2023. 2

[5] 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,

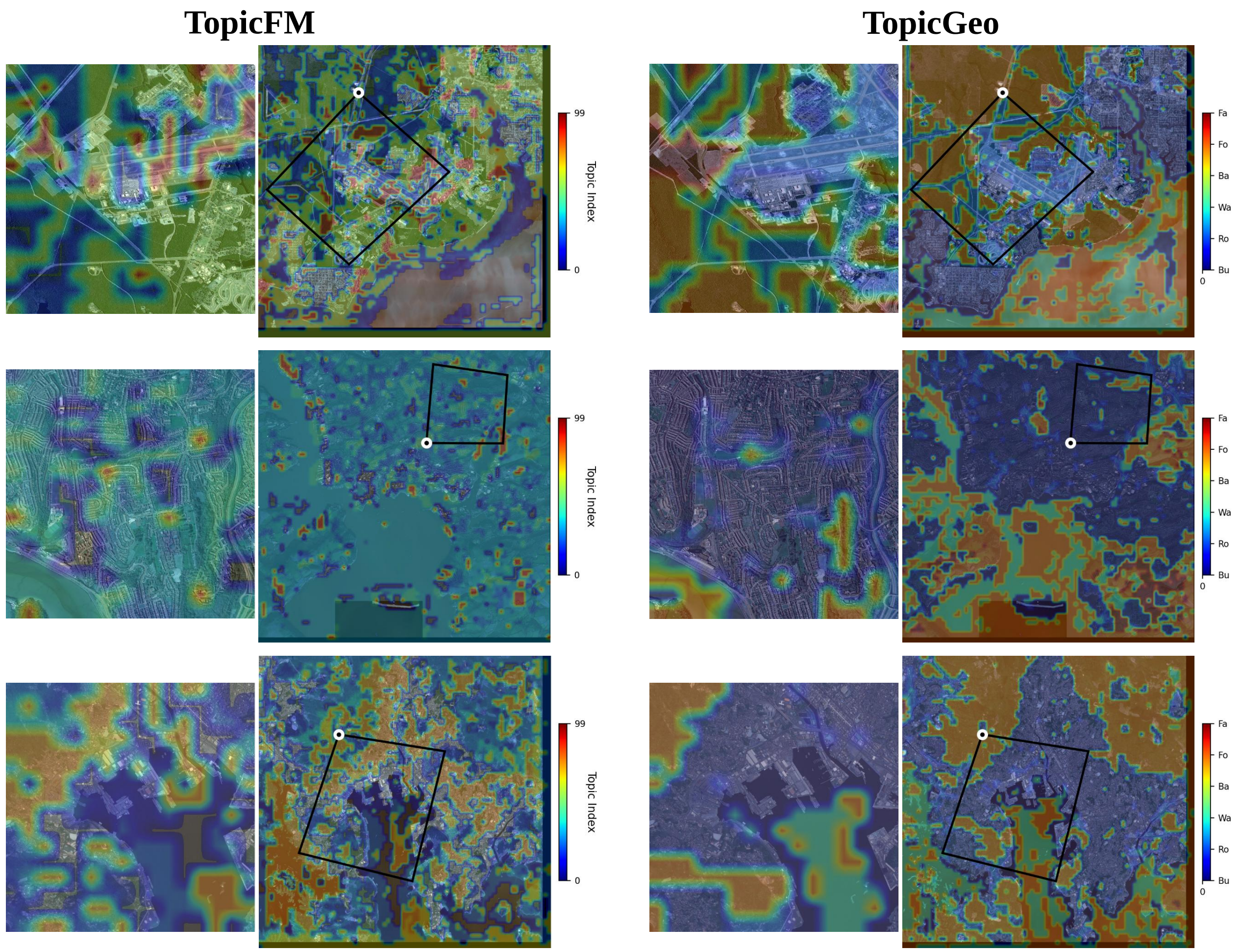

Figure 6. The cross-image topic visualization of existing TopicFM and our TopicGeo on the MTGL40-5 dataset. Our approach effectively extracts topics that align well with visual perception, demonstrating higher semantic consistency between the query image and the reference area of the base image.

2021. 1

[6] S. et al. Tian. Large-scale deep learning based binary and semantic change detection in ultra high resolution remote sensing imagery: From benchmark datasets to urban application. *ISPRS J. Photogramm. Remote Sens.*, 193:164–186, 2022. 5

[7] 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. 3

[8] Z. et al. Wang. Skyscript: A large and semantically diverse vision-language dataset for remote sensing. In *Proc. AAAI Conf. Artif. Intell.*, pages 5805–5813, 2024. 3

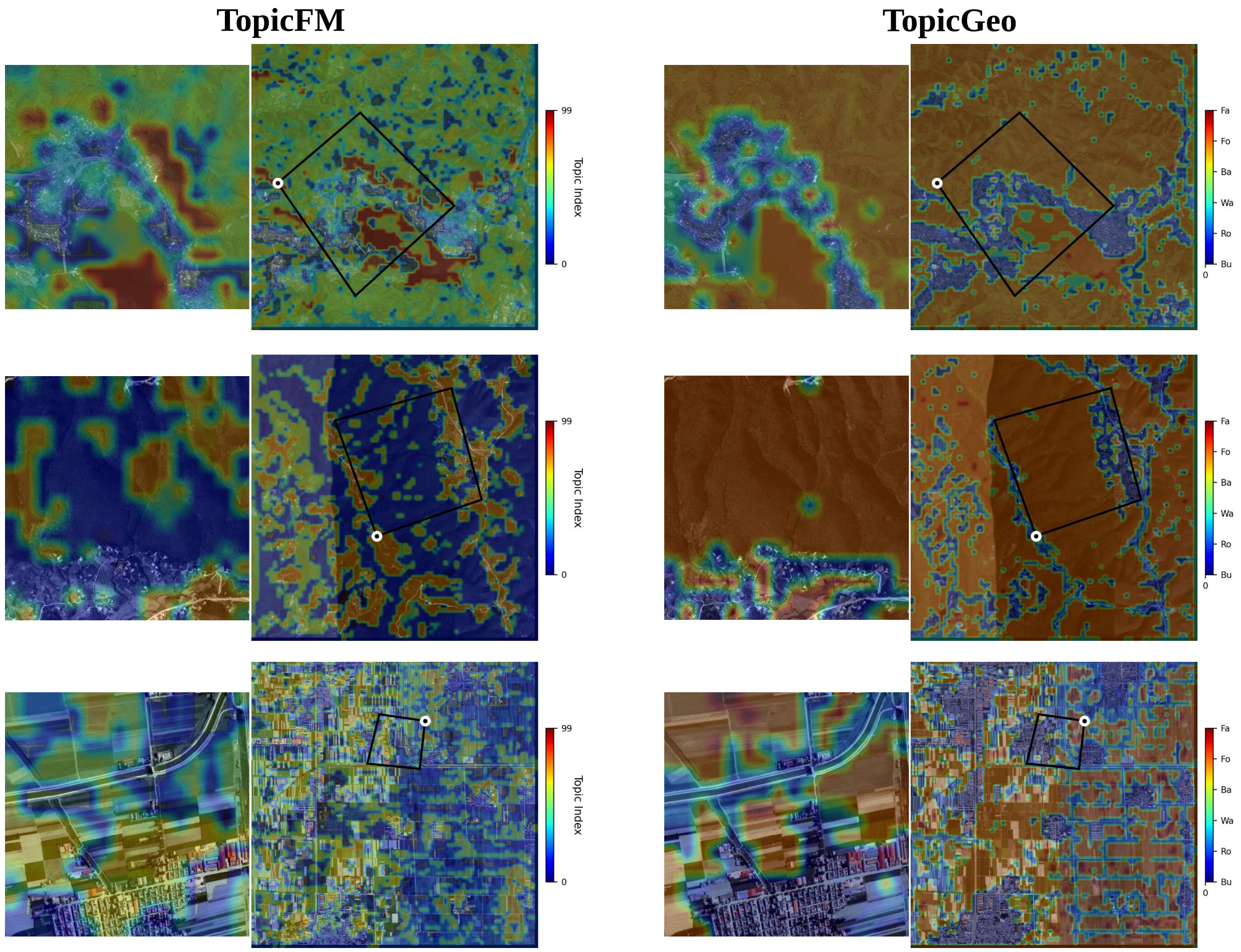

Figure 7. The cross-image topic visualization of existing TopicFM and our TopicGeo on the ZTC dataset. Our approach effectively extracts topics that align well with visual perception, demonstrating higher semantic consistency between the query image and the reference area of the base image.