LEARNING TO LOCALIZE BY RETRIEVAL: SPECTRAL-SPATIAL CONTRASTIVE REPRESENTATION LEARNING FOR MULTISPECTRAL SATELLITE IMAGERY

Anonymous authors
Paper under double-blind review

ABSTRACT

In most cases, supervised object localization in multispectral images is based on dense bounding-box annotations which is not always possible when these annotations are not available or are costly. We formulate localization as similarity search in a learned embedding space and avoid learning to predict a bounding box, thereby eliminating the need for bounding-box supervision for detector-free localization. We train a four channel multispectral encoder with self-supervised momentum contrastive learning to learn the discriminative patch representations. A partition of the target imagery is constructed into overlapping patches, which are incorporated into a normalized representation space, and then indexed so that it is easy to retrieve the patch corresponding to the nearest imagery. Similar patches are retrieved and translated into localized regions by similarity filtering and spatial consolidation, given a query region. We test the learned representation from four complementary viewpoints: embedding separability, retrieval without labels, localization consistency despite query perturbations and retrieval scalability with the growth of the indexed corpus. Based on our results, self-supervised contrastive learning yields structured multispectral representations, which enables repeatable query retrieval based localization at larger corpses and exhibit a low query latency. The results obtained in this study show the viability of an alternative approach to localization in data-constrained multispectral remote-sensing environments based on detector-free retrieval, instead of using annotation intensive detection.

1 INTRODUCTION

Object-level annotations, with the corresponding high cost, are the typical means by which supervised object localization is performed in remote-sensing (RS) multispectral image data, but they are difficult to acquire in regions, conditions and classes Fasana et al. (2022). This challenge becomes more complex in multi-spectral imagery as useful information can be present in multiple spectral bands and not just RGB Mañas et al. (2021); Cong et al. (2022). This provides motivation for representation learning from unlabeled imagery for spatial retrieval. Self-supervised contrastive learning offers an alternative approach to generating visual representations without object-level ground truth, by promoting consistency across different views of the same image, while pushing apart different images. In remote sensing, these representations are used to represent the spatial, spectral and temporal structure, but may also start a detector, classifier or segmentation model via supervised training. We instead examine the possibility of direct linkage between the visual similarity and spatial localization using a learned multi-spectral representation.

We convert a query exemplar into an embedding and look for similar patches in a target gallery. The identifier and spatial coordinates are preserved in each gallery, which can be used to transform patches into candidate regions using class-specific filtering and spatial suppression. The formulation is learned in four channels B, G, R, and NIR with a modified ResNet-50 model He et al. (2016). This results in 128-dimensional embeddings can be used for similarity search using FAISS Douze et al. (2024). Evaluation is conducted on 221 query chips belonging to eight classes and 32,542 patch embeddings in a target gallery. With no target ground truth annotations provided, we investigate the retrieval process representation similarity, retrieval concentration, and query stability in addition to multi-query consistency, inference sensitivity, and retrieval computation and ask if self-supervised

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

multispectral representations are a useful basis for retrieval-driven candidate localization under these constraints.

2 Related Work

2.1 Self-supervised visual representation learning

Self-supervised visual representation learning: Contrastive learning allows visual learning to be achieved without using explicit labels Chen et al. (2020a); He et al. (2020). SimCLR trains representations by comparing the representation of an image with enhanced views of the image with other images Chen et al. (2020a); on the other hand, MoCo trains representations by comparing the representation of an image with enhanced views of the image with a few negative representations in a momentum-updated encoder He et al. (2020). To further enhance this framework, MoCo-v2 introduces new augmentations and training enhancements Chen et al. (2020a). We do not use it as pretraining for a supervised detector, but rather it represents the retrieval-based localization pipeline as the representation layer.

2.2 Remote-sensing representation learning

Another source of variability is the earth observation imagery, which varies in spectral content, acquisition conditions, spatial scale, and geographic distribution. It has been shown in previous work how to leverage these features in the context of self-supervised learning for unlabeled remote-sensing observations (Seasonal Contrast) Mañas et al. (2021) and temporal and multispectral representation learning (SatMAE). While our study is based on a 4-channel ResNet-50 representation, we aim to directly leverage this representation in exemplar-based spatial retrieval Cong et al. (2022).

2.3 Retrieval and localization

Vector similarity search is a natural way to compare learned image representations, and FAISS offers efficient retrieval over the indexed embeddings Douze et al. (2024). Another area of dense contrastive learning has been representations that are able to capture local correspondences for dense prediction Wang et al. (2021). In our formulation, both retrieval is coupled with explicit spatial metadata: each of the patches in the galleries carries its image identity and its coordinates, which enable converting the similarity-ranked patches into spatial candidate regions. The resulting task is a mix of visual retrieval and object localization, while the similarity of the representations is used to rank the candidates and subsequent filtering and spatial aggregation gave rise to object localization hypotheses.

3 Problem Formulation

Let a four-channel multispectral image be $I \in \mathbb{R}^{4\times H\times W}$, with channels corresponding to B, G, R, and NIR. The image is partitioned into patches $p_i = T(I; x_i, y_i, P, S)$, where (x_i, y_i) denotes the upper-left coordinate, P is the patch size, and S is the extraction stride. For the target gallery, $P = 128$ and $S = 64$, corresponding to 50% spatial overlap.

Let E_θ denote the full self-supervised embedding encoder. Each patch is mapped to an L2-normalized embedding, and a query exemplar q is encoded into the corresponding query embedding:

$$\begin{aligned} z_i &= E_\theta(p_i), \qquad \|z_i\|_2 = 1, \\ z_q &= E_\theta(q), \\ \mathcal{G} &= \{(z_i, m_i)\}_{i=1}^{N}, \end{aligned} \tag{1}$$

where m_i stores the source-image identifier and spatial metadata associated with patch i.

For a query-gallery pair, similarity is computed as $s_i = z_q^\top z_i$, which is equivalent to cosine similarity because the embeddings are L2-normalized. For retrieval depth K_r, the candidate retrieval set is

$$\begin{aligned} \mathcal{R}_{K_r}(q) &= \mathrm{TopK}_{i\in\{1,\dots,N\}} \left(z_q^\top z_i\right), \\ b_i &= [x_i, y_i, x_i + P, y_i + P]. \end{aligned} \tag{2}$$

The retrieved candidates are then filtered using a class-specific similarity threshold τ_c and consolidated using class-wise non-maximum suppression. The final output is therefore a set of candidate spatial regions rather than a directly predicted bounding box.

The overall formulation is

Multispectral Patch → Representation → Similarity Retrieval → Spatial Aggregation → Candidate Regions.

4 METHOD

4.1 OVERALL FRAMEWORK

We propose an offline representation-learning and gallery-construction stage and an online query-retrieval and spatial-localization stage in our framework. In the representation-learning stage, four-channel multispectral imagery is cropped into 128×128 patches and doubled as two augmented views. A modified ResNet-50 encoder projects the patches via momentum-contrastive learning to a normalized 128-dimensional representation space. The encoder is frozen after training and then used for inference.

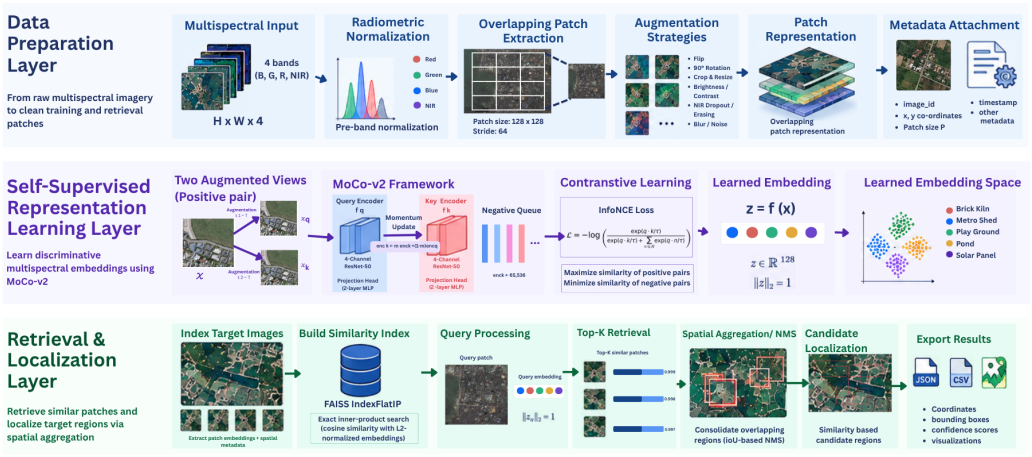

Figure 1: **Retrieval-based multispectral localization framework.**

The frozen encoder is then applied to the target imagery, which is tiled systematically. The patches are extracted at 128×128 resolution with a stride of 64 pixels, and the resulting embeddings are saved alongside their image source ID and spatial information. The embeddings form the target gallery and are indexed with FAISS IndexFlatIP Douze et al. (2024).

When making an inference, the query exemplar goes through the same frozen encoder. It is compared with the gallery and $top - K$ results are retrieved. The retrieved embeddings are associated with their saved metadata, which is used to map them back to 128×128 spatial boxes. Class-wise similarity thresholds eliminate lower-similarity candidates, while class-wise spatial NMS groups together overlapping candidates Neubeck & Van Gool (2006). The offline path is the computationally expensive gallery encoding and indexing stage, while the online path consists of query encoding, retrieval, thresholding, and spatial suppression.

4.2 MULTISPECTRAL REPRESENTATION LEARNING

There are four channels per input patch: B, G, R, and NIR. The first convolution of ResNet-50 is changed from 3 to 4 input channels He et al. (2016; 2020); Chen et al. (2020a). The new input channel is initialized with the average of the original convolutional weights for the RGB channels. The classification layer is discarded, and a 2048-dimensional backbone representation is obtained.

A two-layer projection head maps this representation according to $2048 \to 2048 \to 128$, with Batch Normalization and ReLU between the projection layers. The full embedding encoder E_θ consists of the modified ResNet-50 backbone, projection head, and L2 normalization:

$$E_\theta(p) = \frac{g(f_\theta(p))}{\|g(f_\theta(p))\|_2}. \tag{3}$$

Raw pixel values are scaled by 10,000 and subsequently normalized using channel-wise statistics estimated from the development imagery. The final inference configuration uses $\mu = (0.10424, 0.09753, 0.06988, 0.24941)$ and $\sigma = (0.06071, 0.04603, 0.03621, 0.09987)$ for B, G, R, and NIR, respectively.

In the self-supervised training, two views are constructed by applying a variety of geometric and radiometric transformations to each patch sample, such as random crop-and-resize, horizontal/vertical flip, rotation by multiples of 90, varying brightness and contrast, blurring, additive noise, NIR-channel dropout, and random erasing. These augmentations are meant to induce the representation to maintain patch identity when it is subjected to more carefully controlled changes of appearance.

4.3 Momentum-Contrastive Training

We use a momentum-contrastive training framework following MoCo-v2 He et al. (2020); Chen et al. (2020b). For a training patch, two augmented views are encoded by a query encoder and a momentum-updated key encoder. Let zq denote the query embedding, zk+ the corresponding positive key, and

$$\{z_k^j\}_{j=1}^{K_q} \tag{4}$$

the negative keys stored in the contrastive queue. We use a queue of $K_q = 65{,}536$ negative entries and a temperature of $\tau = 0.2$. The contrastive objective is

$$\mathcal{L} = -\log \frac{\exp\left(z_q^\top z_k^+/\tau\right)}{\exp\left(z_q^\top z_k^+/\tau\right) + \sum_{j=1}^{K_q} \exp\left(z_q^\top z_k^j/\tau\right)}. \tag{5}$$

The query encoder is optimized using AdamW [15] with weight decay 10-4. Mixed-precision training and cosine learning-rate scheduling are used in the training implementation Loshchilov & Hutter (2019). The model used for all experiments is the checkpoint saved as `encoder_moco_resnet50_4ch_ep360.pt`.

We distinguish the contrastive queue size Kq from the retrieval depth Kr used during inference.

4.4 Gallery Construction

The inference encoder is frozen and then applied to the target imagery after training. The target images are divided into 128×128 patches with a stride of 64 pixels, and each patch is fed to the model. The gallery thus contains overlapping representations rather than a partition of the imagery.

For every patch, the system records $m_i = (\text{image_id}, x_i, y_i, P)$. The final gallery used for the main evaluation contains **32,542 indexed patch embeddings** extracted from **40 target images**.

The resulting gallery is $\mathcal{G} = \{(z_i, m_i)\}_{i=1}^N$.

An exact FAISS IndexFlatIP index is constructed over the L2-normalized embeddings Douze et al. (2024). Because the vectors are normalized, inner-product ranking is equivalent to cosine-similarity ranking.

4.5 Query-Based Localization

A query exemplar is transformed using the same preprocessing as the gallery patches and encoded using the frozen model, $z_q = f_\theta(q)$. The gallery is ranked using $s_i = z_q^\top z_i$. The baseline retrieval configuration uses $K_r = 500$. The retrieved patches are converted to candidate boxes through their stored metadata, $b_i = [x_i, y_i, x_i + 128, y_i + 128]$.

Class-specific similarity thresholds are then applied. For each class and target image, the remaining candidate boxes are sorted by similarity and spatially consolidated through non-maximum suppression Neubeck & Van Gool (2006). The baseline NMS IoU threshold is $\theta_{\mathrm{NMS}} = 0.40$. The implementation retains at most 50 detections per class per target image.

This method does not result in detector scores calibrated to target ground truth. Rather, each final bounding box is an interpretation of one or more high-similarity gallery matches.

5 Experimental Setup

5.1 Dataset and Evaluation Protocol

The framework is evaluated in a four-channel multispectral satellite-imagery challenge setting. The development set contains 150 unlabeled four-band TIFF images for self-supervised representation learning, while 40 target images are tiled into 128×128 patches with a stride of 64 pixels, producing 32,542 indexed embeddings. Query exemplars are constructed from the available class-wise annotations, yielding 221 query chips across eight classes: Brick Kiln, Metro Shed, Play Ground, Pond-1, Pond-2, STP, Sheds, and Solar Panel.

Independent ground-truth annotations for the target imagery are unavailable. Therefore, target-box IoU, AP, mAP, precision, and recall are not reported; the query annotations are used only to construct exemplars. The evaluation covers representation similarity, retrieval concentration, query stability, multi-query consistency, inference sensitivity, and retrieval computation. Inference sensitivity is examined through retrieval depth K_r, similarity-threshold offsets, NMS IoU, and indexed gallery size. No matched RGB-only retraining, alternative self-supervised objective, alternative architecture, or public benchmark is evaluated.

5.2 Implementation Configuration

The main representation-learning and inference configuration used throughout the experiments is summarized in Table 1.

Table 1: Main implementation and inference configuration.

Component	**Configuration**
Input	B, G, R, NIR
Patch size	128×128
Target stride	64 px
Target overlap	50%
Backbone	ResNet-50, 4-channel
Backbone feature	2048-D
Projection head	2048-128-128
Embedding	128-D, L2 normalized
Training objective	Momentum contrastive / MoCo-v2
Queue size K_q	65,536
Temperature	0.2
Optimizer	AdamW
Weight decay	10^{-4}
Retrieval index	FAISS IndexFlatIP
Baseline retrieval depth K_r	500
NMS IoU	0.40
Detection cap	50 per class per target image

6 Results

6.1 Representation Similarity

First, we discuss the similarity structure of the learned embedding space. The top-10 retrieved pairs are compared to pairs of embeddings drawn at random in Figure 2, showing the cosine similarities

between the two. The retrieved pairs are clustered at the high end of the cosine-similarity range, while random pairs are distributed throughout the cosine-similarity range.

This behavior can be exploited when using nearest-neighbor retrieval to find highly similar regions in this representation, which is produced by the encoder. Meanwhile, the clustering of random similarities at the high end shows how important a basic problem of overhead imagery is: many different areas of the country may be alike in both spectral and spatial characteristics.

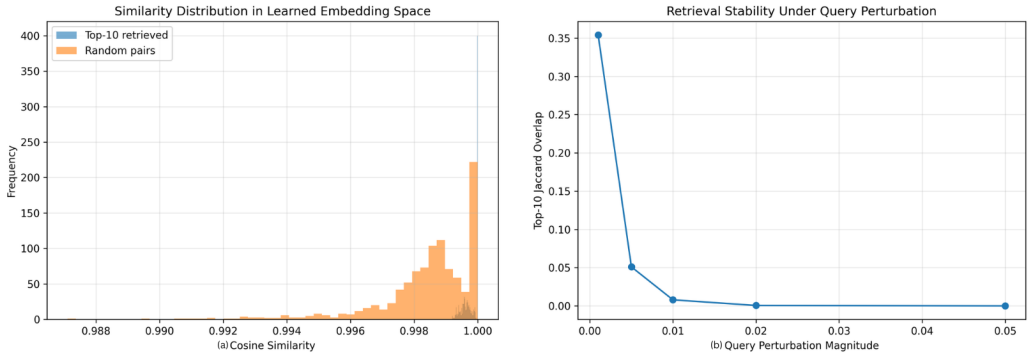

Figure 2: **Similarity distribution and retrieval stability.** (a) Cosine similarity of top-10 retrieved pairs compared with randomly paired embeddings; (b) mean Top-10 Jaccard overlap under increasing query-noise magnitude.

We extend the analysis of retrieval depth to study the spatial extent of retrieved candidates over $K \in \{1, 5, 10, 25, 50, 100, 250, 500\}$. The mean similarity is still high, ranging from 0.999453 at K=1 to 0.998901 at K=500, and the retrieved set is spread over more target images and larger spatial regions. The number of unique target images increases from 1.00 for K=1 to 22.20 for K=100 and 32.94 of 40 for K=500. The mean spatial range increases from 0 at $K = 1$ to about 4296 px in x and 3124 px in y at $K = 500$. These results indicate that high similarity does not imply spatial concentration. Highly similar regions continue to be retrieved with increasing retrieval depth, but across different target images and over wider spatial extents.

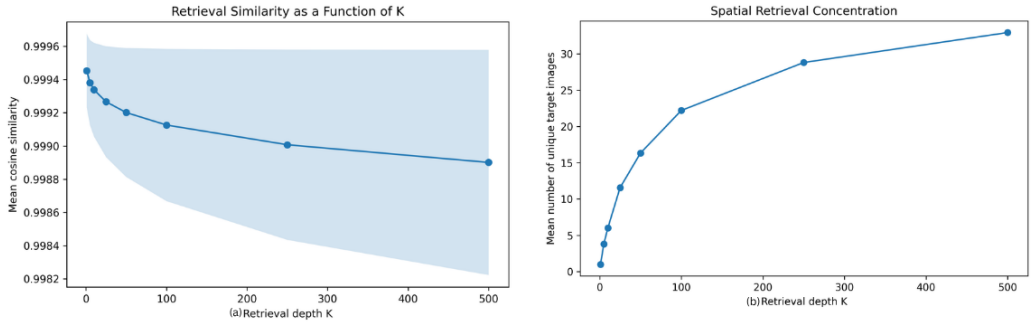

Figure 3: **Retrieval behavior as retrieval depth increases.** (a) Mean retrieval similarity; (b) mean number of unique target images represented in the retrieved set.

6.2 Retrieval-Stage Computational Behavior

We then discuss the computational complexity of the exact retrieval phase. Two gallery sizes are considered: 10K and 25K.

The index construction time for the measured index is 0.279 ms at 10K embeddings, the mean query latency is 0.014584 ms/query, and the embedding memory footprint is 4.883 MB. At 25,000 embeddings, the corresponding measurements are 0.619 ms, 0.015701 ms/query, and 12.207 MB.

As expected for an exact vector index, the query-time measurements are roughly constant over the range tested and memory grows with the number of stored embeddings.

These experiments should be understood as a retrieval-stage systems characterization at the sizes of the galleries tested. They are not intended to define performance for an arbitrary satellite archive, nor do they account for the offline cost to generate and store the full gallery.

7 Analysis and Sensitivity Studies

7.1 Query Perturbation Stability

We examine the consistency of the retrieval-to-localization response when the query exemplar is changed. Each of the 221 query chips undergoes four controlled perturbations: horizontal flipping, vertical flipping, brightness scaling, and contrast scaling. The transformations are given by the equations

$$x'_{c,y,w} = x_{c,y,W-1-w}, \tag{6}$$

for horizontal flipping,

$$x'_{c,y,w} = x_{c,H-1-y,w}, \tag{7}$$

for vertical flipping,

$$x' = \mathrm{clip}\left(1.10x,\, 0,\, \max(x)\right), \tag{8}$$

for brightness scaling, and

$$x' = 1.10(x - \mu_c) + \mu_c, \tag{9}$$

for contrast scaling, where c denotes the spatial mean of spectral channel c.

For each original/perturbed pair, both queries are processed through the complete Kr=500 retrieval, class-specific thresholding, and NMS pipeline. Detection sets are greedily matched at an IoU threshold of 0.5. Let M denote the number of matched detections. The forward agreement is

$$A_f = \frac{M}{|D_o|}, \tag{10}$$

where Do is the original detection set; reverse agreement is defined analogously using the perturbed detection set.

Table 2: Query perturbation stability.

Perturbation	Mean embedding cosine	Agreement	Reverse agreement	Mean matched IoU	Zero-match rate
Brightness	0.999728	0.5872	0.6199	0.9955	0.45%
Contrast	0.999989	0.9319	0.9243	1.0000	0.00%
Horizontal flip	0.996788	0.1297	0.1303	0.5656	43.44%
Vertical flip	0.998380	0.2803	0.2806	0.8054	19.46%

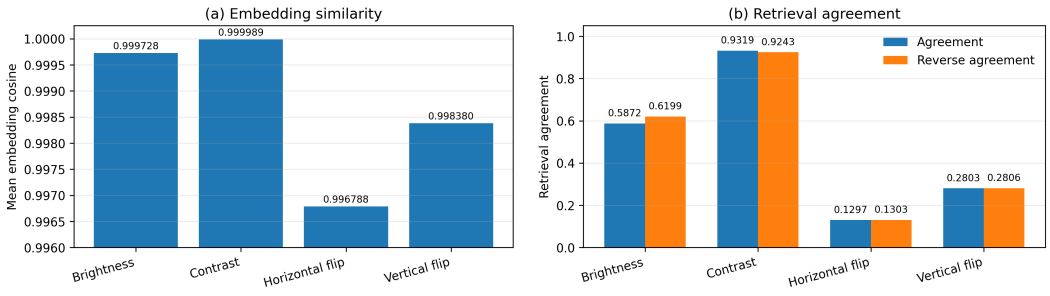

Figure 4: **Query perturbation stability.** (a) Mean embedding cosine; (b) retrieval agreement and reverse agreement under photometric and geometric query perturbations.

The results show a clear distinction between embedding-level stability and retrieval-level stability. The retrieval agreement is highest for contrast (0.9319), followed by brightness (0.5872), vertical flipping (0.2803), and horizontal flipping (0.1297). The agreement achieved by geometric flips is significantly lower, with horizontal flipping at 0.1297 and vertical flipping at 0.2803. Interestingly, embedding cosine is high for all four perturbations even though the agreement for horizontal flipping is only 0.1297. So, high embedding similarity is not necessarily correlated with stable retrieval neighborhoods. This difference is reflected in the zero-match rates as well, which are 43.44% for horizontal and 19.46% for vertical flips.

7.2 Noise Stability

We further assess retrieval repeatability by varying the magnitude of additive query noise. The mean Top-10 Jaccard overlap decreases rapidly as the noise magnitude increases, from 0.3541 at standard deviation 0.001 to 0.0509 at 0.005, 0.0079 at 0.010, 0.0005 at 0.020, and 0 at 0.050. The result matches the above perturbation analysis: retrieval neighborhoods drastically change when the query is perturbed, even given high embedding-level similarity. Jaccard overlap is thus a more stringent measure of retrieval repeatability than embedding cosine.

7.3 Inference-Parameter Sensitivity

Then, we study the sensitivity of the inference pipeline with respect to the retrieval depth and similarity threshold, keeping the rest of the configuration constant, and also the sensitivity of NMS IoU while keeping the rest of the configuration the same. The baseline retrieval depth is $K_r = 500$. When the depth of retrieval is increased from K=50 to K=1000, the number of final candidate detections goes from 6,015 to 14,803, with only a slight change in mean similarity from 0.999387 to 0.999378. Therefore, an increase in K fundamentally amounts to more candidate patches being fed into the spatial aggregation stage, not necessarily a significant change in the scale of similarity.

For the class-specific similarity thresholds, we apply a common offset

$$\tau_c' = \tau_c + \delta, \qquad \delta \in \{-0.10, -0.05, 0, 0.05, 0.10\}. \tag{11}$$

Across all five tested offsets, the final candidate count remains 13,291 and the mean score remains 0.999381, indicating no observable change in output count within the tested range. Finally, NMS IoU is varied over $\{0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$.

The final count is 12,815 at IoU values 0.2 and 0.3, and 13,291 for 0.4 and above. The effect is therefore limited over the evaluated range.

The three-panel analysis allows us to isolate the effects of retrieval depth from downstream post-processing effects. The number of candidates is clearly affected by retrieval depth, while the tested similarity-threshold offsets do not show any noticeable effect and NMS only affects the number of candidates at low IoU thresholds.

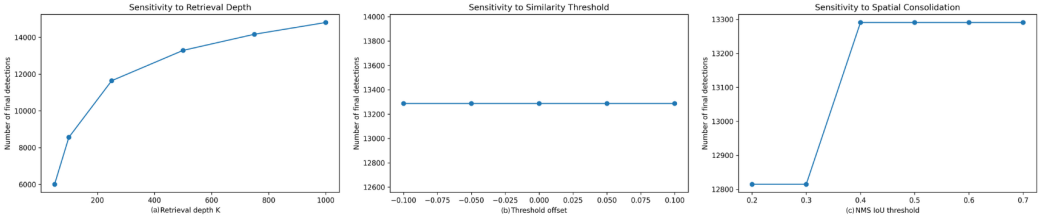

Figure 5: **Sensitivity to inference parameters.** Candidate counts under changes in retrieval depth, similarity-threshold offset, and NMS IoU.

7.4 Multi-Query Consistency and Failure Modes

We also examine the similarity of retrieval neighborhoods using different exemplars of the same query class presented in the same semantic class. The mean Jaccard@100 over 5,149 pairs of queries

in the same class is 0.0560, while the median is zero, showing that there are often significant differences in the top-100 retrieval neighborhoods between exemplars. The mean overlap is higher in Pond-1 (0.2412) and Pond-2 (0.2435) than in Brick Kiln (0.0029) and Play Ground (0.0026) within the classes. The detailed class-wise outcomes are given in Appendix A. These observations reveal an important restriction of exemplar-driven retrieval: the existence of a semantic class does not necessarily imply a similar retrieval neighborhood, and the similarity of the regions may mean that they refer to different spatial instances. The two techniques of spatial filtering and NMS may eliminate redundant candidates, but cannot distinguish for themselves whether a retrieved region is the desired target object. The framework should thus be taken as a similarity-based candidate localization system, and not as a ground-truth-verified detector.

8 Discussion and Limitations

The experiments demonstrate the feasibility of a self-supervised, four-channel representation for retrieval-based spatial inference, where a separate bounding-box prediction head is absent. Similarity space is defined by encoder, retrieved patches are converted to candidate regions by gallery metadata, by similarity filtering, and by spatial suppression. But embedding similarity does not necessarily mean spatial concentration or stability of the retrieval neighborhood. Retrieval depth increases the number of candidate regions as well as spatial support and a query perturbation experiment demonstrates that similarity and retrieval agreement are two different properties. Finally, sensitivity analysis results indicate that the number of candidates retrieved influences the number of candidates retrieved, while the tested similarity-threshold offsets do not cause any obvious changes, and NMS has only a small influence over the range considered. It appears that even for the galaxy sizes tested, embedding storage and constructing galleries are not particularly expensive, but exact similarity search is.

There are several limitations to the conclusions of this study. Absolute localization accuracy is not available because there are no independent target ground-truth annotations, thus it is not possible to measure the accuracy of the localization results using target-box IoU, AP, mAP, precision, or recall, which instead characterize retrieval behavior and repeatability. The evaluation is limited to challenge imagery and query-exemplar setting available, and does not provide generalization across sensors, geographic regions, seasons, spatial resolution, or public remote-sensing benchmarks. There is a low overlap between many pairs of queries which belong to the same class, meaning that retrieval is exemplar-dependent. Last but not least, this paper proposes and tests a single 4-channel ResNet-50 momentum-contrastive representation, leaving out the evaluation of matched RGB-only retraining, alternative self-supervised objectives, alternative architectures, and publicly annotated benchmarks.

9 Conclusion

We proposed an instance of retrieval-based methodology for candidate localization from 4-channel multi-spectral data based on the self-supervised representation learning. A modified ResNet-50 produces embeddings of 128-D space, and the overlapping gallery search and exact search convert query exemplars into ranked spatial candidates. Research has revealed that high embedding similarity does not necessarily imply spatial concentration or stability of retrieval neighborhoods. The larger the retrieval depth, the greater the number of candidates and the spatial support, and the perturbation experiments show different level of stability between embedding similarity and retrieval agreement. The results from sensitivity analysis indicate that the number of candidates influenced by the value of retrieval depth, while the tested similarity-threshold offsets do not present any significant changes and NMS only has a limited influence. The results reported here corroborate the idea of retrieval-driven candidate localization in the test scenario where the target annotations were absent, but cannot determine absolute localization accuracy or generalization beyond this specific scenario. It is still crucial to have annotated benchmarks and matched baselines for additional assessment.

References

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1597–1607, 2020a.

X. Chen, H. Fan, R. Girshick, and K. He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Y. Cong, S. Khanna, C. Meng, P. Liu, E. Rozi, Y. He, M. Burke, D. B. Lobell, and S. Ermon. SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. In *Advances in Neural Information Processing Systems*, volume 35, pp. 197–211, 2022.

M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou. The faiss library. *arXiv preprint arXiv:2401.08281*, 2024.

C. Fasana, S. Pasini, F. Milani, and P. Fraternali. Weakly supervised object detection for remote sensing images: A survey. *Remote Sensing*, 14(21):5362, 2022.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

O. Mañas, A. Lacoste, X. Giró-i Nieto, D. Vázquez, and P. Rodríguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9414–9423, 2021.

A. Neubeck and L. Van Gool. Efficient non-maximum suppression. In *Proceedings of the 18th International Conference on Pattern Recognition*, volume 3, pp. 850–855, 2006.

X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3024–3033, 2021.

10 Reproducibility Statement

The main paper provides the information required to reproduce the reported experiments. Section 4 describes the multispectral preprocessing, modified ResNet-50 encoder, momentum-contrastive training, gallery construction, and retrieval-based localization pipeline, while Table 1 summarizes the principal training and inference configuration. Section 5 specifies the dataset composition, query construction, and evaluation protocol, and Sections 6–7 report the retrieval, stability, consistency, sensitivity, and computational experiments. The supplementary material will provide the implementation and additional processing details used in the experiments, subject to any restrictions on redistribution of the challenge data.