X-JEPA: A Novel Joint Learning Cross-Modal Predictive Alignment Framework for Remote Sensing Image Retrieval

Supplementary Material

Supplementary Overview

This supplementary material complements the main paper by providing additional results, ablations, and visualization-based analysis for our proposed X-JEPA model. Specifically, we include:

- **Section A**: Detailed descriptions of datasets used for RS-CMIR evaluation.
- **Section B**: Extensive ablation studies covering masking strategies, PSA loss sensitivity, normalization, and patch size effects.
- **Section C**: Additional qualitative results showcasing top-5 retrievals, attention maps, semantic generalization across modalities, and extended related work.

These findings offer deeper insight into the architecture's robustness, alignment strategies, and semantic reasoning capability under cross-modal conditions.

A. Datasets Description

We evaluate the performance of X-JEPA on two large-scale publicly available benchmarks designed for RS-CMIR, encompassing diverse sensor modalities, including SAR, multispectral, and optical imagery. These datasets enable robust assessment under both unimodal and cross-modal retrieval protocols.

BigEarthNet-14K (BEN-14K): This benchmark [6] comprises 14,832 image pairs from Sentinel-1 (SAR) and Sentinel-2 (multispectral) satellites, annotated with 19 land cover categories. The Sentinel-1 data includes VV and VH polarization bands, while Sentinel-2 provides multispectral imagery at 10m and 20m resolutions, harmonized via bicubic interpolation for consistency across modalities.

fMoW-RGB: The Functional Map of the World RGB dataset [2] contains high-resolution satellite images categorized into 62 semantic geospatial classes. Originally curated for classification tasks, it includes approximately 363,000 training samples and 53,000 testing samples, offering a diverse visual distribution for unimodal retrieval.

fMoW-Sentinel: As an extension of the fMoW dataset, fMoW-Sentinel [3] incorporates Sentinel-2 multispectral images corresponding to the RGB samples. It preserves the original 62-class taxonomy while substantially expanding the dataset, with 712,874 training samples, 84,939 validation images, and 84,966 test instances, enabling scalable cross-modal evaluations.

Table 1. Ablation on Masking strategy for both modalities.

Masking Strategy	S1→S1	S2→S2	S1→S2	S2→S1
Random	**72.98**	**82.65**	**61.23**	**63.73**
MultiBlock	71.77	77.45	54.05	60.12

Table 2. Ablation on Normalization strategy

Normalization Type	S1→S1	S2→S2	S1→S2	S2→S1
DyT [7]	69.74	81.43	55.56	58.89
LAYERNORM (Ours)	**72.98**	**82.65**	**61.23**	**63.73**

B. Extended Ablation Study

In this section, we present additional ablations referenced in Section 4.5 of the main paper. Specifically, we analyze: (i) the impact of masking pattern design (random vs. multiblock), (ii) the sensitivity of performance to the PSA loss weight λ_{PSA}, (iii) the effect of different normalization strategies (LayerNorm vs. DyT), and (iv) the influence of patch size selection in the encoder. These results offer deeper insights into the architectural robustness and training stability of X-JEPA.

B.1. Effect of Masking Pattern Design

We study the impact of masking pattern, specifically, random vs. multiblock strategies, in shaping the semantic learning behavior of X-JEPA. As shown in Table 1, random masking consistently outperforms multiblock across all tasks. This is attributed to the higher spatial entropy of random patterns, which compels the model to learn robust semantic priors from sparse and scattered contexts. Conversely, multiblock patterns concentrate missing tokens in local regions, limiting the generalization capacity, especially for cross-modal scenarios involving heterogeneous sensor shifts.

B.2. Impact of Normalization Strategy

Table 2 evaluates the role of normalization schemes in cross-modal alignment. We compare LAYERNORM, adopted in X-JEPA, against the dynamic token-aware normalization from DyT [7]. While DyT uses learnable statistics conditioned on token content, we find that it underperforms across all tasks. Specifically, LAYERNORM delivers average gains of +2.1% (unimodal) and +5.2% (cross-modal) F1-score. This suggests that global per-token normalization provides more stable gradients and alignment consistency under modality heterogeneity, unlike token-wise normalization, which amplify inter-sensor variance.

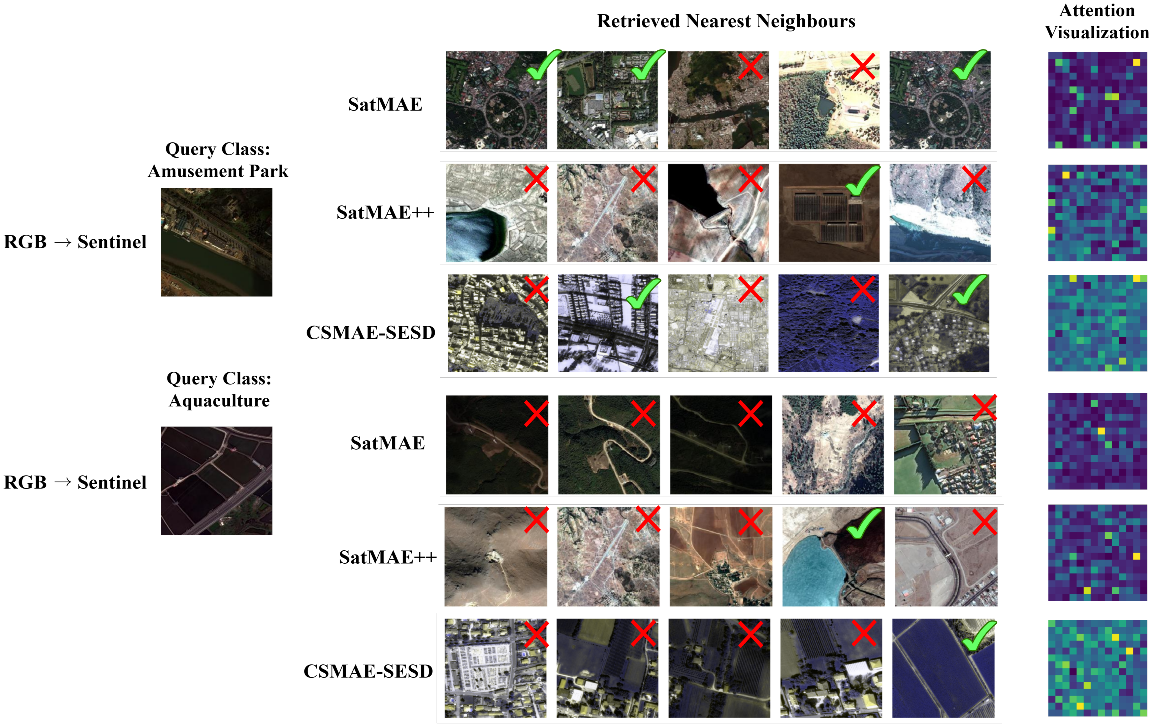

Figure 1. Qualitative retrieval comparison on fMoW for query *amusementpark* and *aquacuilture* across SatMAE, SatMAE++, CSMAE-SESD (Disjoint), and X-JEPA, including attention maps and cross-modal retrieval outputs.

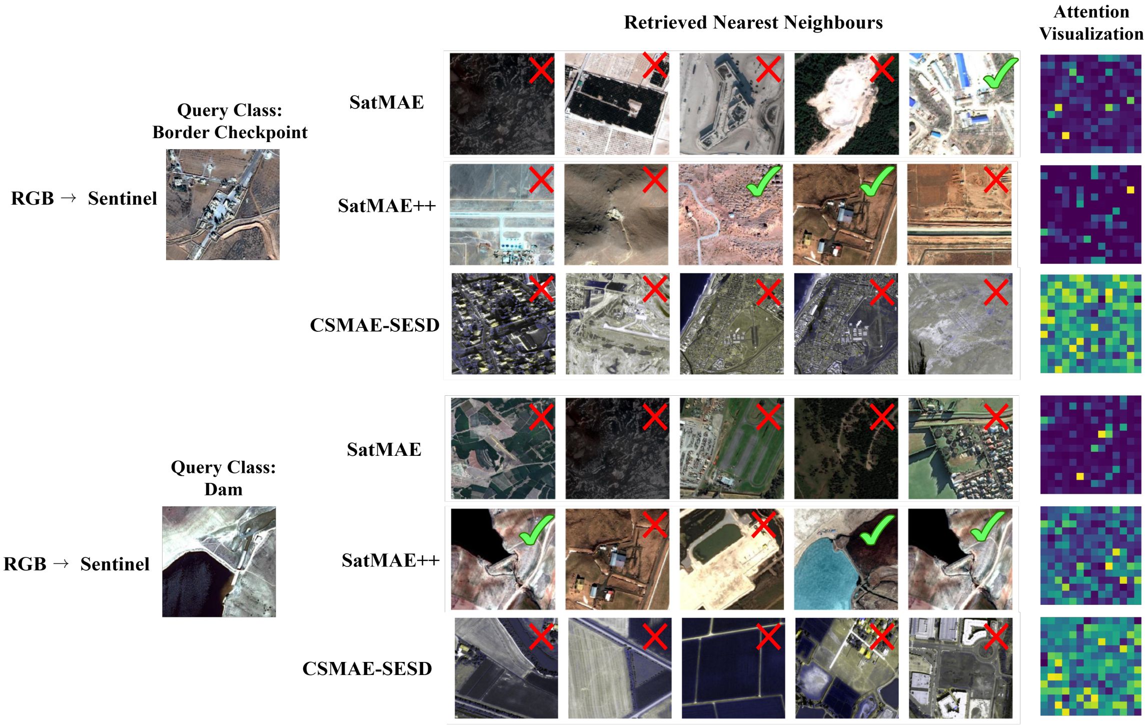

Figure 2. Qualitative retrieval comparison on fMoW for query *bordercheckpoints* and *dam* across SatMAE, SatMAE++, CSMAE-SESD (Disjoint), and X-JEPA, including attention maps and cross-modal retrieval outputs.

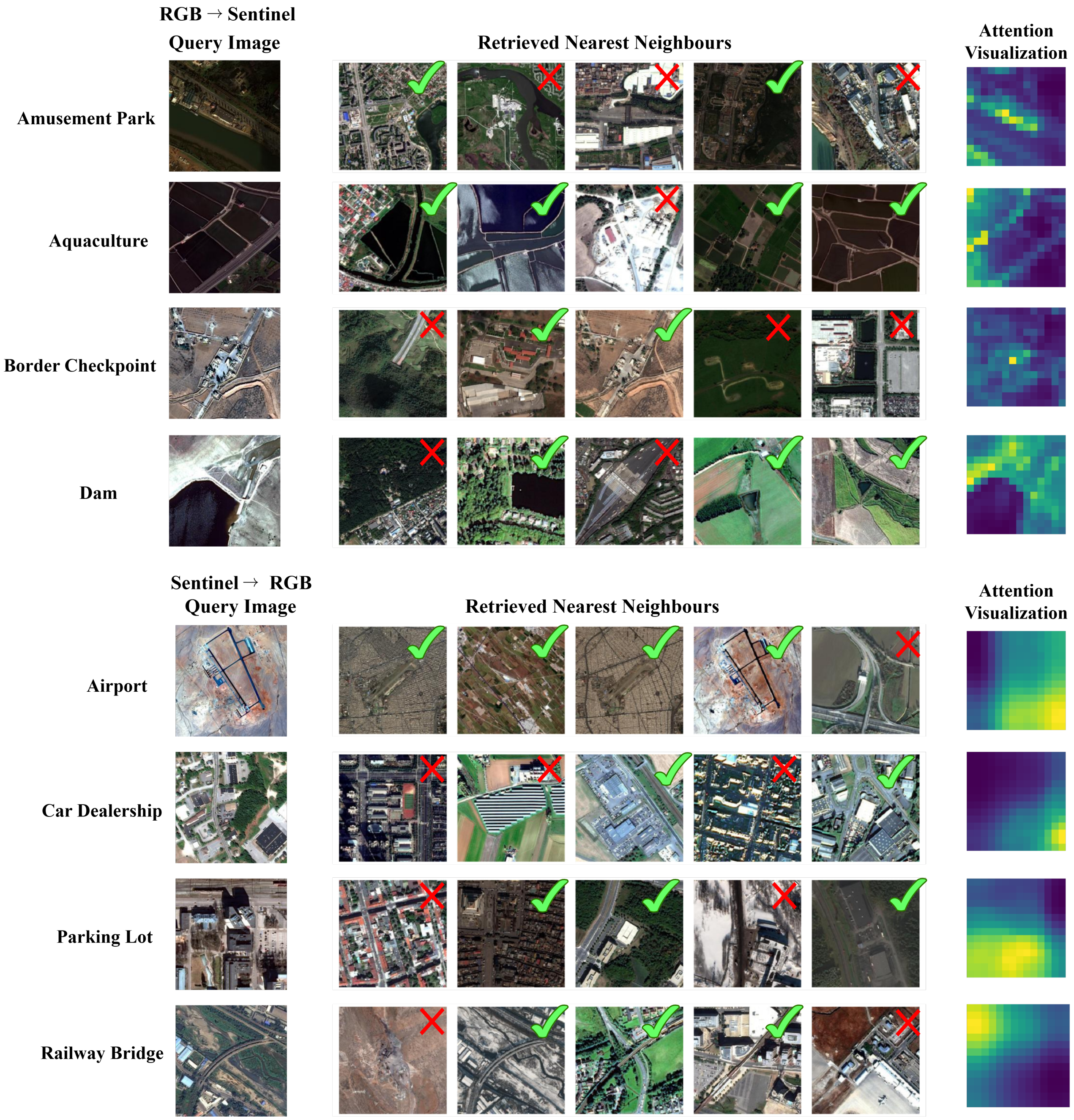

Figure 3. Retrieval results of X-JEPA for fMoW datasets on cross-modal tasks, RGB$rightarroe$sentinel and Sentinel→RGB

Table 3. Ablation on varying λ_{PSA} in training objective

λ_{PSA}	S1→S1	S2→S2	S1→S2	S2→S1
0.1	**72.98**	**82.65**	**61.23**	**63.73**
0.5	70.51	81.85	56.44	57.48
0.75	69.54	81.56	56.96	56.65

Table 4. Ablation on Patch size under multi-modal masking.

Patch Size	S1→S1	S2→S2	S1→S2	S2→S1
8×8	72,15	83.35	52.18	54.73
12×12	71.71	81.72	53.45	54.96
16×16	**72.98**	**82.65**	**61.23**	**63.73**
20×20	69.21	79.91	54.04	57.76

Table 5. Ablation on PSA vs. InfoNCE loss

Loss Variant	S1→S1	S2→S2	S1→S2	S2→S1
L2 + InfoNCE + VICReg	70.12	79.83	58.62	60.04
L2 + PSA + VICReg (Ours)	**72.98**	**82.65**	**61.23**	**63.73**

B.3. Loss Weight Sensitivity

We study the sensitivity of X-JEPA to the weighting coefficient λ_{PSA} that controls the strength of the Prediction Space Alignment (PSA) loss. As shown in Table 3, smaller values (e.g., $\lambda_{\text{PSA}} = 0.1$) yield consistently stronger performance across both unimodal and cross-modal retrieval tasks. Unlike fixed-distance losses, PSA adaptively captures the evolving latent structure, enabling dynamic and geometry-aware alignment across modalities.

Larger values tend to overemphasize second-order structure in the latent space, leading to instability and degraded alignment, especially in cross-modal scenarios. These findings highlight the importance of balancing geometric alignment (via PSA) with representational smoothness and diversity, which is maintained through the L2 and VICReg terms.

Mathematical Clarification of PSA Loss

Let $\Delta = \hat{z}_t - z_t$ denote the difference between the predicted and target embeddings. PSA measures this error using a learnable Mahalanobis metric:

$$\mathcal{L}_{\text{PSA}} = \sqrt{\Delta^\top M \Delta}, \qquad M = LL^\top \succ 0,$$

where $L \in \mathbb{R}^{d \times d}$ is a trainable parameter matrix. Parameterizing M as $LL^\top$ guarantees positive definiteness and allows the network to adaptively rescale error dimensions, emphasizing directions that capture modality-invariant semantics while suppressing noisy or unstable dimensions (e.g., those dominated by SAR speckle). Unlike a vanilla L2 loss, which assumes an isotropic error structure, PSA learns a task-specific metric that is more sensitive to meaningful feature discrepancies.

Intuition and Complementarity with VICReg.

PSA acts as a *predictive-guided alignment loss*, shaping *where* in the latent space the predictor should align its outputs with the target. This allows the model to focus on semantically relevant components of the representation space, rather than minimizing error uniformly across all dimensions. VICReg plays a complementary role (refer Table 6 by ensuring that embeddings maintain sufficient variance and avoid collapse. Together, these objectives strike a balance between precise semantic alignment (PSA) and global representation diversity (VICReg).

Comparison with InfoNCE loss

As reported in Table 6 of the main paper, removing PSA reduces cross-modal F1 by 4–6%, confirming its importance in driving semantic alignment. We additionally compare PSA with a contrastive InfoNCE objective in Table 5, where $\mathcal{L}_{\text{PSA}}$ is replaced with a standard contrastive loss that requires large negative sets. InfoNCE underperforms PSA across all retrieval directions, showing a further 2–3% drop in cross-modal F1. This highlights PSA's advantage in avoiding false negatives (common in RS due to seasonal or sensor-induced similarity) and providing a smooth alignment signal without costly negative mining.

B.4. Effect of Patch Size

We analyze the impact of varying patch sizes on retrieval performance across both unimodal and cross-modal tasks. As shown in Table 4, a moderate patch size of 16×16 yields the best overall results, achieving the highest F1 scores in all evaluation directions. Smaller patches (e.g., 8×8) provide finer spatial resolution and perform competitively in unimodal settings but underperform in cross-modal scenarios where high-level semantic abstraction is essential. Conversely, larger patches (e.g., 20×20) reduce spatial granularity and lead to performance degradation due to the loss of local context.

Table 6. Component-wise ablation of VICReg loss.

VICReg Components	S1→S1	S2→S2	S1→S2	S2→S1
w/o $\mathcal{L}_{\text{var}}$	69.42	77.36	56.21	58.04
w/o $\mathcal{L}_{\text{cov}}$	71.03	80.54	59.14	61.02
$\mathcal{L}_{\text{var}} + \mathcal{L}_{\text{cov}} + \mathcal{L}_{\text{inv}}$	**72.98**	**82.65**	**61.23**	**63.73**

B.5. VICReg Loss ($\mathbf{L}_{\text{VICReg}}$) Analysis

Motivation

Cross-modal representation learning is highly susceptible to feature collapse, where the learned embeddings converge to a low-variance solution, resulting in degenerate representations that fail to capture meaningful semantic differences. To address this, we adopt the Variance–Invariance–Covariance Regularization (VICReg) objective [1], which stabilizes training by combining three complementary terms: invariance, variance, and covariance.

Component Ablation

Table 6 reports a fine-grained ablation by selectively disabling one component at a time. Removing the variance term leads to the largest performance degradation, confirming its critical role in avoiding collapse. The covariance term yields a smaller but consistent benefit, improving generalization by reducing feature redundancy. The complete VICReg formulation achieves the highest unimodal and cross-modal F1-scores.

C. Extended Qualitative Performance

We provide additional qualitative results to support the retrieval performance of X-JEPA. This section provides ex-

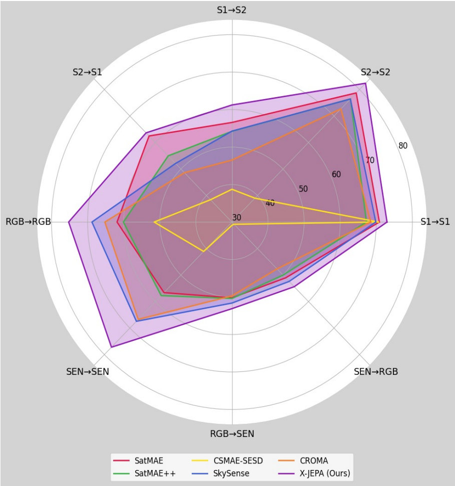

Figure 4. Comparision of F1-scores across all retrieval directions for various SOTA self-supervised models.

tended visualizations demonstrating the effectiveness under modality shift and structural variation. It includes comparative retrievals, top-5 retrieval results, and attention heatmaps.

C.1. Comparative Performance Visualization

Figure 4 offers an aggregate visual summary of F1-scores across all retrieval directions. The radial coverage effectively captures each model's consistency and retrieval robustness, with X-JEPA yielding the most balanced radial profile, reflecting consistent F1-scores across modalities and strong generalization to diverse sensor shifts.

To further contextualize these gains, Figure 1 and Figure 2 present qualitative retrieval comparisons across SatMAE, SatMAE++, CSMAE-SESD (Disjoint) models, and our X-JEPA on the fMoW dataset. We show queries such as airfields, ports, and coastal infrastructure retrieved across modalities. Competing methods often return visually similar but semantically inconsistent scenes, while X-JEPA reliably retrieves functionally aligned results, demonstrating its superior ability to abstract high-level semantic structure across modalities. The final column displays attention heatmaps from the predictor module. These maps serve as an interpretable proxy for semantic alignment, showcasing how X-JEPA grounds predictions by attending to domain-relevant regions. For instance, it highlights structured cues like runways, terminals, or port boundaries, which are functionally consistent across modalities but visually divergent. In contrast, baseline methods often exhibit diffuse or misaligned attention, which correlates with incorrect retrievals. While X-JEPA consistently retrieves semantically aligned results, occasional failure cases arise in dense urban zones with overlapping semantic categories (e.g., industrial vs. commercial), where attention can diffuse due to visual ambiguity.

To further probe the consistency of our framework, we analyze the bidirectional behavior of X-JEPA across all retrieval directions. Despite the inherent information imbalance between modalities (e.g., Sentinel-2/RGB providing richer spectral cues than Sentinel-1 SAR), our model is trained with *symmetric objectives*, jointly optimizing S1→S2 and S2→S1 predictions. Empirically, we observe only mild asymmetry, with F1-score gaps consistently remaining below 2.5% across datasets, underscoring robust generalization even when conditioned on the lower-information modality. Importantly, the consistently low gaps across datasets indicate that X-JEPA generalizes robustly even when conditioned on the lower-information modality. This finding highlights a key strength of our predictive approach: it learns cross-modal correspondences in a manner that is not overly sensitive to modality asymmetry, thereby enabling reliable retrieval performance in real-world scenarios where only one modality might be available or dominant.

In addition, we qualitatively inspected (see Figure 4 several retrieval pairs and found that failure cases under S1→S2 prediction are often attributable to extreme SAR noise (e.g., heavy speckle) or visually ambiguous scenes (e.g., dense urban areas with overlapping functional classes). However, even in these challenging cases, the retrieved images generally remain semantically plausible, reinforcing the model's ability to capture high-level functional alignment rather than overfitting to low-level visual similarity.

C.2. Retrieval Results with X-JEPA

Figure 3 presents extended qualitative examples of cross-modal retrieval using X-JEPA on the fMoW dataset, highlighting its semantic generalization capabilities under modality shifts. Each row depicts a query image (left), followed by the top-5 retrieved scenes from the opposite modality and the corresponding attention map (right) from the predictor module.

The queries span a range of functional geospatial categories, including *airport terminals*, *parking lots*, *railway bridges*, and *dams*, each characterized by variation in spatial geometry, environmental context, and sensor-induced distortions. Despite these challenges, X-JEPA consistently retrieves scenes with high semantic fidelity and structural alignment. For instance, it matches radial layouts in *airfields*, modular grids in *border checkpoints*, and hydrological boundaries in *aquaculture zones*—even when these appear visually dissimilar across modalities.

Crucially, the attention maps reveal that X-JEPA concentrates on modality-invariant and semantically salient regions, such as runway strips, overpass geometries, shorelines, and block-wise infrastructure layouts. These focused activations indicate that the model captures transferable, high-level features rather than relying on superficial texture similarity. Such behavior explains its robustness in challenging retrieval scenarios, including cluttered urban zones or topographically complex terrain.

These visualizations validate the model's ability to bridge semantic gaps between modalities without explicit supervision, and reinforce its practical utility in downstream tasks such as infrastructure monitoring, disaster response, and change detection across heterogeneous remote sensing platforms.

C.3. Attention Map Visualizations

To better understand X-JEPA's behavior during cross-modal retrieval, we visualize the attention maps generated by the final layer of the encoder for the query modality. These maps highlight spatial regions the model considers semantically important when forecasting latent embeddings of the target modality.

This interpretability mechanism serves two key purposes: (i) it reveals whether the model attends to semantically meaningful structures—such as road junctions, coastline boundaries, or gridded infrastructure—across modalities; and (ii) it confirms that the model effectively suppresses modality-specific noise and irrelevant textures. As seen in Figures 1, 2, and 3, X-JEPA consistently focuses on cross-modally aligned regions (e.g., runways, ports, and border structures) that are functionally diagnostic of the underlying scene class.

In contrast, baseline methods like SatMAE and CSMAE-SESD often exhibit diffuse or spatially inconsistent attention, gravitating toward texture-dominant but semantically weak regions. For instance, in the *airport* and *aquaculture* queries, X-JEPA accurately localizes class-specific visual anchors (e.g., radial symmetry, aquatic grid structures), validating its strong modality-invariant semantic reasoning. These attention patterns explain its robustness under cross-view shifts and further reinforce the model's reliability for real-world geospatial retrieval scenarios.

C.4. Semantic Feature Explanation

Beyond visual attention maps, we further analyze how X-JEPA captures modality-invariant semantic features that underpin its retrieval performance. The model does not rely on superficial textures or spectral signatures unique to a specific modality. Instead, it learns to abstract transferable concepts such as shape regularity, spatial layout, and object co-occurrence patterns that generalize across views.

For example, in the *airport* class, the model consistently retrieves scenes exhibiting radial symmetry and linear runways, regardless of whether the query is in RGB or SAR. Similarly, for *ports* and *aquaculture*, it focuses on repetitive dock structures or aquatic grid patterns, even when occluded or distorted in one modality. This behavior indicates that X-JEPA aligns modalities not at the pixel or texture level, but in a latent semantic space shaped by functional and spatial abstractions.

C.5. Extended Related work

While cross-modal alignment methods such as contrastive paradigms (e.g., CLIP-style models [4]) and query-shift adaptation techniques [5] have shown success in natural image domains, they generally rely on feature similarity maximization or query embedding adaptation and do not explicitly model predictive latent alignment between heterogeneous sensors.

Cross-modal retrieval in remote sensing (RS) introduces unique challenges due to large spectral, spatial, and geometric gaps between modalities such as Sentinel-1 and Sentinel-2 imagery. These gaps require approaches that capture modality-invariant semantics rather than relying solely on pairwise similarity. To address this, X-JEPA introduces a predictive framework where modality-specific predictors transform contextual embeddings into the latent space of the target modality under symmetric objectives. This design enforces semantic grounding, mitigates the false-negative problem inherent to contrastive learning, and produces interpretable attention maps that highlight functionally relevant structures such as runways or port boundaries.

References

[1] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021. 4

[2] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018. 1

[3] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022. 1

[4] Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. *Advances in Neural Information Processing Systems*, 34:29406–29419, 2021. 6

[5] Haobin Li, Peng Hu, Qianjun Zhang, Xi Peng, Xiting Liu, and Mouxing Yang. Test-time adaptation for cross-modal retrieval with query shift. *arXiv preprint arXiv:2410.15624*, 2024. 6

[6] Gencer Sumbul, Arne De Wall, Tristan Kreuziger, Filipe Marcelino, Hugo Costa, Pedro Benevides, Mario Caetano, Begüm Demir, and Volker Markl. Bigearthnet-mm: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 9(3): 174–180, 2021. 1

[7] Jiachen Zhu, Xinlei Chen, Kaiming He, Yann LeCun, and Zhuang Liu. Transformers without normalization. *arXiv preprint arXiv:2503.10622*, 2025. 1