Loc-Aware: Integrating Geographic Metadata into Contrastive Learning

Juliana Carvalho[1[0009−0001−3784−6320]⋆], Clément Rambour[2[0000−0002−9899−3201]], Arnaud Breloy[1[0000−0002−3802−9015]], and Javiera Castillo Navarro[1[0000−0003−4917−5103]]

[1] CEDRIC Laboratory, Conservatoire National des Arts et Métiers (CNAM)
[2] ISIR Laboratory, Sorbonne University

Abstract. Self-supervised learning (SSL) has emerged as a scalable paradigm for Earth Observation (EO), addressing the scarcity of labeled data by leveraging abundant unlabeled satellite imagery. However, standard contrastive approaches predominantly rely on visual content, often overlooking the rich auxiliary metadata associated with EO data. In this work, we investigate the integration of spatial and auxiliary metadata into the SSL pipeline to enhance representation learning. We propose Loc-Aware, a novel contrastive loss that structures the latent space by incorporating geographic proximity through a geodesic distance kernel defined on the Earth's manifold. This implicitly yields location-aware representations that do not require metadata as input during inference. Beyond our base spatial formulation, we introduce three methodological extensions to investigate adaptive kernels and incorporate other types of metadata, such as temporal and resolution information. We evaluate the transfer capabilities of these visual representations across six EO datasets and different encoder architectures. Our results demonstrate that the proposed method yields highly semantic representations, consistently outperforming both the image-image contrastive baseline (SimCLR) and state-of-the-art multimodal approaches like SatMIPS.

Keywords: Self-supervised learning · Earth observation · Geographic metadata.

1 Introduction

The expansion of satellite missions has resulted in a nearly continuous stream of Earth Observation (EO) imagery. However, supervised learning performance is hindered by the prohibitive cost and domain expertise required for ground-truth labeling [30]. While self-supervised learning (SSL) offers a robust solution for learning generalizable representations from unlabeled data [6,15,4,1], traditional contrastive frameworks typically treat all negative samples as equally dissimilar, ignoring potential semantic correlations. To address this in the geospatial

⋆ juliana.carvalho-de-souza@lecnam.net

domain, methods like Tile2Vec [19] explicitly exploit these semantic correlations by assuming that images acquired from geographically proximate locations naturally share underlying features (e.g., similar land cover or vegetation patterns), using spatial distance to guide the learning process. Existing methods that ignore this auxiliary metadata may not fully capture the continuous nature of geospatial semantics. Although recent multimodal approaches have attempted to align images with metadata [20,3], they often require complex alignment strategies and may be limited by the nature of cross-modal contrastive objectives, which primarily capture redundant information across modalities [11].

To address these limitations, we propose Loc-Aware, a novel contrastive formulation inspired by the y-Aware loss [12], but fundamentally redesigned for the geospatial domain. Rather than merely adapting a continuous loss, we introduce a geodesic (Haversine) kernel that directly encodes the spherical geometry of two-dimensional geographic coordinates. This yields an image-only representation learning mechanism that exploits underlying spatial structures without the computational overhead or alignment difficulties of multimodal fusion techniques. Notably, our method generates representations that are implicitly guided by geographic metadata via the contrastive loss function, but do not require coordinate inputs at downstream inference. Our main contributions are:

- **Loc-Aware Loss.** We introduce a novel mathematical formulation incorporating geographic inductive biases into the InfoNCE objective. By modeling spatial relations through an exponential geodesic kernel, our objective induces a geographically structured latent space capturing the spherical topology of EO data.
- **Empirical Validation.** We evaluate the generalization of Loc-Aware via transfer learning across six remote sensing datasets. Our method significantly outperforms SimCLR in average k-NN accuracy and consistently surpasses specialized metadata-driven baselines, such as SatMIP and SatMIPS, across diverse downstream tasks.
- **Methodological Extensions.** To systematically evaluate the framework's robustness against batch stochasticity and multimodal heterogeneity, we expand Loc-Aware with three modular extensions that dynamically scale spatial bandwidth and integrate temporal and resolution metadata.

We release our code to support future research at https://github.com/JuCarv-bit/loc-aware.

2 Related Work

Self-Supervised Learning in Remote Sensing. Self-supervised learning (SSL) has emerged as a powerful paradigm for representation learning without the need for exhaustive human annotations [13,30]. In the computer vision domain, contrastive methods like SimCLR [6] and MoCo [15] have established highly scalable pre-training frameworks by enforcing representation invariance across augmented views. For EO, representation learning is particularly critical

due to the prohibitive costs of expert labeling. Prior works have demonstrated that pre-training on generic datasets (e.g., ImageNet) often yields sub-optimal transfer performance on EO tasks compared to domain-specific pre-training [28]. Consequently, remote sensing variants have adapted these frameworks, employing strategies such as spatial tile-to-tile contrastive learning [19], temporal contrastive pre-training like Seasonal Contrast [21], and masked autoencoders like SatMAE [9]. Subsequent adaptations have expanded into cross-sensor architectures like CROMA [14] and cross-scale invariance strategies such as Scale-MAE [24]. Although some methods implicitly use metadata for discrete pair sampling, their fundamentally vision-centric objectives fail to integrate continuous geographic coordinates into the loss function, leaving the rich spatial geometry unexploited.

Metadata in Self-Supervised Learning. Modern satellite imagery is intrinsically multimodal, frequently accompanied by rich acquisition metadata (e.g., coordinates, sensor types, timestamps). Early approaches, such as [2], formulated geo-location as a pretext task, creating a joint objective that combines spatial prediction with temporal contrastive learning. More recently, frameworks have explicitly integrated metadata into the latent space. Inspired by CLIP [23], methods like SatCLIP [20] and SatMIP [3] employ multi-modal architectures to align image and metadata embeddings. Extensions such as SatMIPS [3] further enrich this by combining image-image and metadata-image contrastive objectives. While effective, these approaches generally require tailored projection heads and complex alignment strategies, focusing primarily on shared information between modalities rather than complementary features [11].

Continuous Metadata and Distance-Aware Losses. An alternative to explicit multi-modal alignment is to inject metadata directly into the loss function as a structural prior. Grounded in Vicinal Risk Minimization (VRM) [5], [12] introduced the y-Aware loss for 3D Magnetic Resonance Imaging (MRI): a weighted InfoNCE objective where continuous proxy metadata (e.g., patient age) modulates pairwise attraction via a Radial Basis Function (RBF) kernel, defining a soft neighborhood in metadata space rather than a hard positive/negative split. We adopt this loss-level approach as our direct foundation and extend it to the geospatial domain via a geodesic kernel (Section 3).

Differences from Prior Methods. Unlike SatCLIP, SatMIP, and SatMIPS, which rely on dual-encoders for cross-modal alignment, our approach only requires an image encoder. Expanding on the y-Aware loss [12], Loc-Aware integrates coordinates directly into InfoNCE via a geodesic kernel, which structures the latent space without additional complex alignment networks.

3 Method

Contrastive Pre-training Objective. To learn visual representations that preserve real-world geographic topology in satellite imagery, we propose Loc-Aware: a self-supervised framework using sensor-provided coordinates as a soft supervisory signal. We assume access to an unlabeled dataset $\mathcal{D} = \{(x_i, m_i)\}_{i=1}^{N}$,

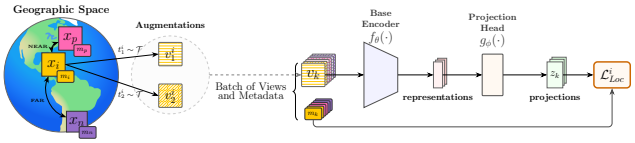


Fig. 1: Overview of the Loc-Aware training scheme. Two augmented views of each image are encoded and compared contrastively. Geographic metadata (in this case, location) modulates $\mathcal{L}^i_{Loc}$ (Eq. (3)) via a geodesic kernel (Eq. (2)): geographically close samples (x_p) act as soft positives, while distant ones (x_n) are treated as negatives.

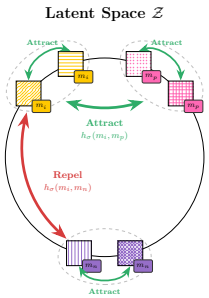


Fig. 2: Latent space dynamics of the Loc-Aware loss. The exponential geodesic kernel h_σ (Eq. (2)) induces a spatial prior that dynamically pulls the embeddings of geographically positive samples (e.g., m_p) closer to the anchor in $\mathbb{S}^{d-1}$ while pushing distant negative ones (e.g., m_n) apart, thereby preserving real-world topology within the representation space.

where $x_i \in \mathcal{X}$ is a satellite image and $m_i = (\varphi_i, \lambda_i) \in \mathbb{R}^2$ represents its associated geographic metadata (latitude and longitude in radians). Following SimCLR [6] (see Fig. 1), each training image x_i is augmented twice by independent draws $t^i_1, t^i_2 \sim \mathcal{T}$, producing two views $(v^i_1, v^i_2) \triangleq (t^i_1(x_i), t^i_2(x_i))$. Following standard contrastive learning notation, we define a base neural network encoder $f_\theta(\cdot)$ (e.g., a ViT-S backbone) and a non-linear projection head $g_\phi(\cdot)$ (e.g., a 2-layer MLP), parameterized by learnable weights θ and ϕ. For each view v^i_k ($k \in \{1, 2\}$), the projected embeddings $z^i_k = g_\phi(f_\theta(v^i_k)) \in \mathbb{S}^{d-1}$ are ℓ_2-normalized. Since outputs lie on the unit hypersphere, their inner products equal cosine similarities, denoted as $\mathrm{sim}(z_1, z_2) = z_1^\top z_2$.

Our method builds upon the y-Aware loss [12], which extends InfoNCE [22] (originally designed to maximize a lower bound on the mutual information $I(v_1, v_2)$ between augmented views) to continuous proxy metadata $y_i \in \mathbb{R}$. Instead of treating negatives uniformly, it defines a soft neighborhood around anchors via an RBF kernel $w_\sigma(y_k, y_i) = \exp(-\|y_k - y_i\|^2/2\sigma^2)$ (where $\sigma > 0$ is the RBF hyperparameter), where samples with similar metadata act as soft positives:

$$\mathcal{L}^y_{NCE} = -\sum_{k=1}^{n} \frac{w_\sigma(y_k, y_i)}{\sum_{j=1}^{n} w_\sigma(y_j, y_i)} \log \frac{\exp(\mathrm{sim}(z^i_1, z^k_2)/\tau)}{\frac{1}{n}\sum_{j=1}^{n} \exp(\mathrm{sim}(z^i_1, z^j_2)/\tau)}, \tag{1}$$

where $\tau > 0$ is the temperature hyperparameter and n is the mini-batch size. In our case, the metadata consists of geographic metadata, so we adapt this formulation to the geospatial domain by replacing the Euclidean RBF $w_\sigma(y_k, y_i)$ with

an exponential geodesic kernel $h_\sigma(m_k, m_i)$ directly over geographic coordinates m_i, respecting Earth's spherical geometry:

$$h_\sigma(m_x, m_y) = \exp\left(-\frac{\tilde{d}(m_x, m_y)^2}{2\sigma^2}\right), \tag{2}$$

where $\tilde{d}(m_x, m_y) = \frac{d(m_x, m_y)}{D_{max}}$ represents the normalized spatial distance between two samples. The raw spatial distance $d(m_x, m_y)$ is computed using the standard Haversine formula [26], and $D_{max} = \pi R$ is the maximum possible great-circle distance (using Earth's radius $R \approx 6371$ km). As illustrated in Fig. 2, this scale-aware formulation dynamically pulls the embeddings of geographically co-located samples closer together in $\mathbb{S}^{d-1}$ while pushing distant ones apart. This yields the Loc-Aware InfoNCE loss for a single sample i:

$$\mathcal{L}^i_{Loc} = -\sum_{k=1}^{n} \frac{h_\sigma(m_k, m_i)}{\sum_{j=1}^{n} h_\sigma(m_j, m_i)} \log \frac{\exp(\text{sim}(z_1^i, z_2^k)/\tau)}{\frac{1}{n}\sum_{j=1}^{n} \exp(\text{sim}(z_1^i, z_2^j)/\tau)}, \tag{3}$$

Note that the sum over k includes $k = i$: since both views of sample i share the same geographic centroid, $h_\sigma(m_i, m_i) = 1$, ensuring the true positive always receives the maximum proximity weight. The overall objective averages this loss over all samples in the mini-batch: $\mathcal{L}_{Loc} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}^i_{Loc}$.

Downstream Transfer. To transfer the learned geographically-informed representations, we discard the projection head g_ϕ post-pretraining and retain only the base encoder f_θ. The resulting features are implicitly location-aware because the spatial information was incorporated through the contrastive objective. However, the coordinate metadata are not needed during supervised fine-tuning in geospatial downstream tasks.

Recovery of Standard InfoNCE. As $\sigma \to 0$, h_σ converges to a Kronecker delta, assuming no two distinct samples share the exact same centroid ($m_k \neq m_i$ for $k \neq i$): only the weight of the true positive ($k = i$) remains. This exactly recovers the canonical InfoNCE loss [22]:

$$\lim_{\sigma \to 0} \mathcal{L}^i_{Loc} = -\log \frac{\exp(\text{sim}(z_1^i, z_2^i)/\tau)}{\frac{1}{n}\sum_{j=1}^{n} \exp(\text{sim}(z_1^i, z_2^j)/\tau)} = \mathcal{L}_{NCE}. \tag{4}$$

Global Spatial Homogeneity. Conversely, as $\sigma \to \infty$, the kernel becomes uniform, treating all geographic locations as equally related to the anchor. By tuning σ between these limits, the loss interpolates between strict instance discrimination and broad geographic grouping.

4 Experiments and Results

4.1 Datasets

Pre-training dataset. We use the Functional Map of the World (fMoW) `rgb-baseline` dataset [8,3] for pre-training (363,571 `train` samples) and evaluation (53,042 `val` samples) across 62 classes [2]. In the version we used [3],

images are cropped to their Areas of Interest, resized to 224×224, and paired with JSON file containing the acquisition attributes (location, timestamp, GSD, among others) necessary for our objectives.

Evaluation datasets. Transfer evaluation spans six remote sensing RGB datasets: RESISC45 [7], Optimal31 [29], UC Merced (UCM) [31], FGSC23 [32], EuroSAT [17], and So2Sat [33].

4.2 Implementation Details & Evaluation Protocol

Mainly inspired by [3], we pretrained ResNet-50 [16] and ViT-S [10] backbones in PyTorch for 200 epochs using AdamW with a one-epoch linear warmup, cosine decay, global batch size 1024, contrastive temperature $\tau = 0.1$ [6], and a fixed random seed of 42 for reproducibility. Optimizer hyperparameters and the bandwidth σ are detailed in Section 4.6.

Backbone and projection head. We append a standard SimCLR two-layer MLP projection head $g_\phi(\cdot)$ [6] to project backbone features into a 128-D embedding space. Concretely, for ViT-S the head is $\mathbb{R}^{384} \to \mathbb{R}^{384} \to \mathbb{R}^{128}$, and for ResNet-50 it is $\mathbb{R}^{2048} \to \mathbb{R}^{2048} \to \mathbb{R}^{128}$. The Loc-Aware loss operates on these projected embeddings $z = g_\phi(f_\theta(x))$, while all downstream evaluation and UMAP visualizations use the raw encoder representations $h = f_\theta(x)$ (384-dim for ViT-S), discarding g_ϕ entirely.

Image augmentations. We follow the satellite-adapted augmentation strategy of [3], including random resized cropping, flipping, discrete 90° rotations, color jittering, Gaussian blur, and per-channel fMoW normalization.

Metadata preprocessing. Geographic centroids (φ, λ) in radians are computed from the `raw_location` (*Well-Known Text* or WKT polygons) in the fMoW JSON metadata as average of polygon vertices and next they are converted from degrees to radians, yielding $(\varphi, \lambda) \in [-\pi/2, \pi/2] \times [-\pi, \pi]$. We precompute these to enable fast batch-wise Haversine distance calculations without I/O bottlenecks.

Evaluation protocol. Frozen encoder features are ℓ_2-normalized and classified via weighted k-NN ($k = 5$, $\tau_{knn} = 0.07$) using softmax-weighted neighbor labels. Performance is reported as top-1 accuracy.

4.3 Main Results

Main Results (ViT-S). The downstream results for the ViT-S backbone are displayed in Table 1. Loc-Aware yields consistent improvements, outperforming SimCLR (+2.94% average accuracy) and establishing a new state-of-the-art among location-aware methods (SatMIP/SatMIPS). Non-random models were pre-trained for 200 epochs. The "Avg. Acc." denotes the unweighted mean accuracy in percentage across all six downstream datasets. For a fair comparison, SatMIP and SatMIPS are re-executed using their original text-based location encoding (latitude and longitude in degrees).

Table 1: **Performance comparison on classification accuracy % (ViT-S).** Best results are in **bold**, second-best are underlined. Grey indicates random initialization. †Re-executed using raw location metadata. *Uses the theoretical anchor ($\sigma \approx 0.0212$) derived from pre-training statistics ** Is empirically tuned ($\sigma = 0.0220$, default). See Section 4.6 for discussions about hyperparameter tuning.

Model	R45	O31	UCM	F23	Euro	So2	Avg. Acc.
Random	32.85	29.35	45.71	30.09	71.48	33.33	40.47
SimCLR	87.30	83.01	94.76	55.83	94.96	57.17	78.84
SatMIP†	87.46	85.27	95.95	58.50	96.17	56.63	80.00
SatMIPS†	89.38	86.88	95.71	61.17	94.72	57.16	80.84
Loc-Aware*	89.49	86.67	**97.14**	59.83	97.13	57.43	81.28
Loc-Aware**	**89.56**	**86.99**	**97.14**	**62.01**	**97.22**	**57.77**	**81.78**

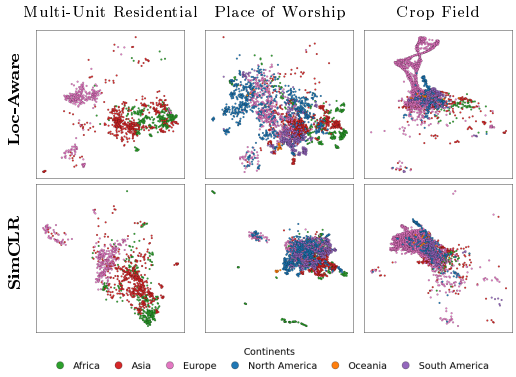


Fig. 3: UMAP projections of MoW embeddings for three classes, comparing Loc-Aware (top) with SimCLR (bottom), colored by continent.

4.4 Quality of Visual Representations

UMAP. We computed 2-D UMAP projections of ℓ_2-normalized fMoW validation embeddings, colored by continent (Fig. 3). While SimCLR suffers from continental overlap, Loc-Aware leverages spatial context to disentangle geographic regions, isolating European samples in *Crop Field*. Quantifying this, the mean Silhouette Coefficient [27] across all 62 fMoW classes (using continents as clusters) is 0.044 for Loc-Aware versus 0.024 for SimCLR (an $\approx$1.8$\times$ improvement), confirming stronger continental separation.

Top-k Image Retrieval. We evaluate the feature space via global top-3 retrieval (cosine similarity), matching validation queries against the training set with a 50m deduplication threshold to prevent trivial temporal matches. For a *Race Track* query (Fig. 4), SimCLR relies solely on texture, incorrectly retriev-

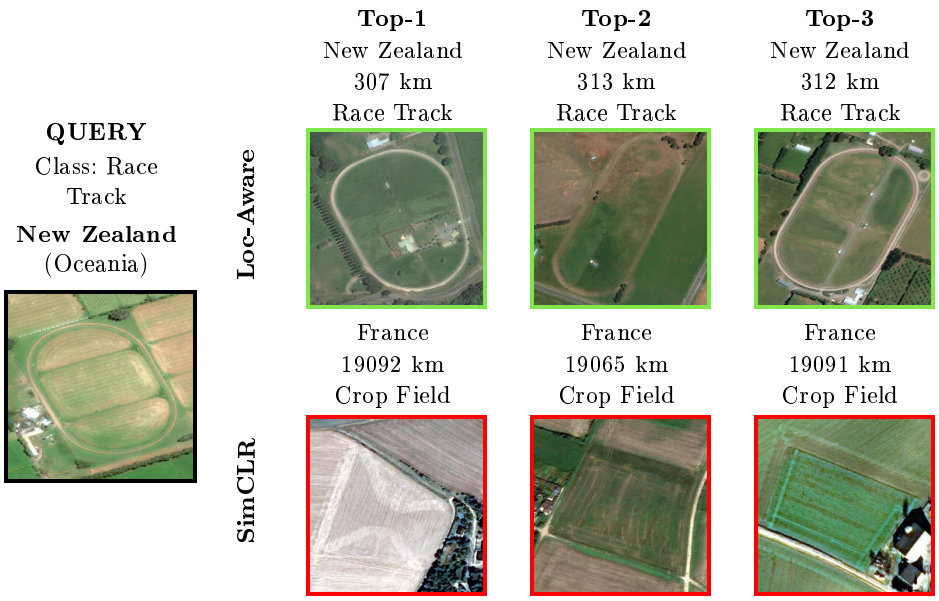


Fig. 4: Top-3 retrieval for a *Race Track* query. Misled by texture, SimCLR retrieves incorrect *Crop Fields* (red). Loc-Aware accurately finds regional *Race Tracks* (green).

ing morphologically similar *Crop Fields* 19,000 km away in France. In contrast, Loc-Aware restricts retrievals to a tight regional neighborhood (307–313 km) and correctly identifies *Race Tracks*. This confirms our geodesic kernel successfully couples visual semantics with geographic topology, mitigating texture-induced mismatches.

4.5 Ablation Study

Generalization Across Architectures (ResNet-50). To ensure that the performance gains are not architecture-specific, we evaluate the methods using a ResNet-50 encoder. As shown in Table 2, our continuous formulation effectively injects the spatial prior into the CNN. Loc-Aware outperforms SimCLR by a margin of 3.63% on average. We restrict this comparison to SimCLR to focus solely on how our method generalizes across backbones.

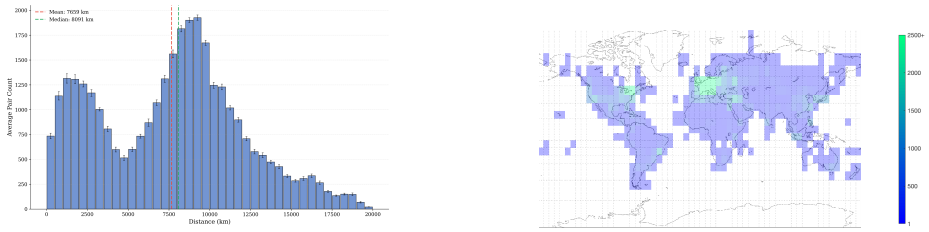
4.6 Hyperparameter Selection

Loc-Aware: Intuition and Choice of σ. As illustrated in Fig. 5, the fMoW training dataset exhibits severe clustering predominantly across North America and Europe [8]. This physical clustering influences training dynamics. By sampling empirical training batches (10 random batches of size 256 each), we

Table 2: **Performance comparison (%) with ResNet-50.** This ablation evaluates architectural generalization relative to the SimCLR baseline.

Model	R45	O31	UCM	F23	Euro	So2	Avg. Acc.
Random	24.36	22.25	42.61	27.18	60.24	36.95	35.60
SimCLR	85.65	83.01	94.29	55.22	93.57	53.78	77.59
Loc-Aware	**89.27**	**88.28**	**96.43**	**61.89**	**96.04**	**55.38**	**81.22**

observe a bimodal pairwise distance distribution (Fig. 5a) with intra-continental ($\mu_1 = 1909$ km) and inter-continental ($\mu_2 = 9375$ km) modes. Thus, we anchor the scale parameter at $\sigma \approx 2.12 \times 10^{-2}$ (a kernel half-life of 500 km). Representing merely ∼2.5% of Earth's maximum great-circle distance, this restrictive threshold isolates the lower tail of the intra-continental mode, ensuring most intra-continental samples act as spatial hard negatives.



(a) (b)

Fig. 5: **Dataset spatial density and pairwise distance distribution. (a)** Pairwise Haversine distances showing a bimodal distribution. **(b)** Geographic density highlighting sample concentration in North America and Europe [8].

Empirical Hyperparameter Search. A grid search centered around this anchor reveals that $\sigma = 2.20 \times 10^{-2}$ yields peak downstream performance (81.78%), as shown in Table 3.

Table 3: Grid Search of σ on average accuracy.

σ ($\times 10^{-2}$)	0.10	1.00	2.05	2.12	**2.20**	2.35	2.50	2.65	5.00	10.00
Avg. Acc. (%)	80.80	81.73	81.15	81.28	**81.78**	81.67	81.65	81.41	81.55	80.60

Optimizer Hyperparameters. Optimal learning rate (lr) and weight decay (wd) found via grid search [3] are detailed in Table 4.

Table 4: Selected learning rate (lr) and weight decay (wd).

	Loc-Aware		SimCLR	
	ViT-S	**ResNet-50**	**ViT-S**	**ResNet-50**
lr	3.75×10^{-4}	3.75×10^{-4}	3.75×10^{-4}	3.75×10^{-4}
wd	0.5	0.1	0.5	0.1

5 Loc-Aware Extensions

To evaluate the robustness of Loc-Aware to batch stochasticity and multimodal heterogeneity, we introduce three extensions addressing two key questions: (1) Can a fixed bandwidth capture spatial structures across random mini-batches? (2) Does incorporating auxiliary metadata[3] enrich representations or overly restrict them? To explore these questions, we dynamically adapt the bandwidth to local batch geometry (ALoc-Aware), and integrate the extended metadata either implicitly via sentence embeddings (Sent. Embed.) or explicitly via factorized kernels (Joint Kernel).

Adaptive Loc-Aware (ALoc-Aware). With stochastic mini-batches, a fixed global bandwidth σ rarely captures meaningful geographic neighbors consistently. To circumvent this, we introduce a sample-specific bandwidth σ_i, which replaces the fixed parameter in the geodesic kernel (Eq. (2)), yielding the adaptive formulation $h_{\sigma_i}(m_k, m_i) = \exp(-\tilde{d}(m_k, m_i)^2/2\sigma_i^2)$. Drawing inspiration from Stochastic Neighbor Embedding (SNE) [18], we model the proximity weights as a categorical distribution. The probability $P_{j|i}$ of sample i selecting neighbor j is defined as:

$$P_{j|i} = \frac{h_{\sigma_i}(m_j, m_i)}{\sum_{k \neq i} h_{\sigma_i}(m_k, m_i)}. \tag{5}$$

To calibrate the effective neighborhood size, we perform a binary search for σ_i such that the Shannon entropy of this distribution equals the logarithm of a target perplexity K (the effective number of neighbors):

$$H(P_i) = -\sum_{j \neq i} P_{j|i} \log P_{j|i} = \log(K). \tag{6}$$

Substituting the calibrated h_{σ_i} back into Eq. (3) guarantees a dense, balanced geographical neighborhood presence regardless of empirical batch variations.
We tuned ALoc-Aware via grid search, selecting $K = 256$ as the optimal configuration (Table 5). ALoc-Aware (81.26%) slightly underperforms Loc-Aware (81.78%), suggesting the fixed σ is already a stable prior across mini-batches.

[3] We extend location to timestamp and GSD because they drive seasonal variations and visual scale; this triplet (location, time, GSD) empirically yields the strongest representation learning in multimodal frameworks like SatMIP(S) [3].

Table 5: Grid Search of K Value on average accuracy

K Value	2^5	2^6	2^7	2^8	2^9	2^{10}	2^{11}	2^{12}
Avg. Acc. (%)	79.03	79.44	79.18	**81.26**	80.96	81.07	81.24	81.17

Implicit Metadata Fusion (Sent. Embed.). The first extended strategy collectively encodes all three metadata fields (location, timestamp, and ground sample distance) into a single embedding space using a *frozen* sentence transformer, unlike the trainable BERT in [3], simply to compute pairwise proximity distances. Following the key-value text serialization format of [3], we format the different metadata fields as key-value pairs of strings and concatenate them to form a composite sentence using the syntax "`location: value1, timestamp: value2, gsd: value3`". We then extract ℓ_2-normalized embeddings $\hat{\mathbf{e}}_i \in \mathbb{R}^{384}$ using a frozen `all-MiniLM-L6-v2` model [25]. To integrate this unified semantic space into our original formulation, we substitute the spatial distance $\tilde{d}$ in our similarity kernel (Eq. (2)) with the normalized cosine distance between metadata embeddings: $\tilde{d}_{\cos}(i,j) = (1 - \hat{\mathbf{e}}_i^\top \hat{\mathbf{e}}_j)/2 \in [0, 1]$. This yields an extended semantic kernel $h_\sigma^{\cos}$, which directly replaces h_σ in the Loc-Aware loss (Eq. (3)). Consequently, temporal, spatial, and resolution proximities jointly shape the contrastive objective without requiring any changes to the core visual architecture. We use a bandwidth of $\sigma = 0.0010$. As shown in Table 6, Sent. Embed. reaches 79.60%, outperforming SatMIP (79.20%) but falling short of Loc-Aware (81.78%), suggesting that implicit embedding might conflate distinct spatial signals.

Joint Metadata Kernel. The Loc-Aware loss (Eq. (3)) weights contrastive pairs solely through the geodesic kernel $h_\sigma(m_k, m_i)$ (Eq. (2)), encoding geographic proximity alone. The second extended strategy employs a factored approach to keep each modality independent: rather than projecting location, GSD, and timestamp into a shared embedding space, we augment h_σ multiplicatively with two Gaussian RBF factors following the factored kernel design of [12]:

$$w_{\text{gsd}}(g_i, g_j) = \exp\left(-\frac{(\log g_i - \log g_j)^2}{2\,\sigma_{\text{gsd}}^2}\right), \quad w_{\text{time}}(t_i, t_j) = \exp\left(-\frac{(t_i - t_j)^2}{2\,\sigma_{\text{time}}^2}\right), \tag{7}$$

where i and j index samples in the mini-batch (as in Eq. (3)), $g_i > 0$ is the GSD of sample i in meters, and t_i is its acquisition timestamp in days since a fixed reference epoch. GSD is encoded in log-space so that equal multiplicative ratios receive equal kernel response regardless of absolute scale, since $|\log g_i - \log g_j| = |\log(g_i/g_j)|$. Because the joint weight is a product, a pair receives high contrastive weight only when simultaneously close in location, resolution, and acquisition time, imposing a stricter alignment objective than the single-variable baseline. Replacing $h_\sigma(m_k, m_i)$ in Eq. (3) with $h_\sigma(m_k, m_i) \cdot w_{\text{gsd}}(g_k, g_i) \cdot w_{\text{time}}(t_k, t_i)$ yields the joint-kernel loss $\mathcal{L}_{\text{JK}}^i$; setting $w_{\text{gsd}} \equiv w_{\text{time}} \equiv 1$ exactly

recovers Eq. (3). We implemented the Joint Metadata Kernel keeping $\sigma_{\text{loc}} = 2.20 \times 10^{-2}$ identical to the Loc-Aware baseline. The remaining two bandwidths use training split statistics: $\sigma_{\text{gsd}} = 0.206$ (std. dev. of $\log g_i$) and $\sigma_{\text{time}} = 426$ days (Median Absolute Deviation of t_i, robust to fMoW's skewed dates). As shown in Table 6, Joint Kernel achieves 81.50%, the best among all extensions and the strongest on fine-grained datasets (R45, O31, UCM), though it still falls below Loc-Aware (81.78%), indicating that the stricter multi-modal alignment slightly hurts generalization.

Table 6: **Performance comparison of Loc-Aware extensions: accuracies in %.** The table unifies the evaluation of adaptive bandwidth (ALoc-Aware), extended metadata encoding via Sentence Embeddings (*Sent. Embed.*), and the Joint Metadata Kernel against prior baselines. Best results across all configurations are highlighted in **bold**, and second-best are underlined. [1]Results directly reported from SatMIPS [3].

Model	Metadata	R45	O31	UCM	F23	Euro	So2	Average
SatMIP[1]	Extended	87.50	84.80	95.20	56.40	95.70	55.90	79.20
SatMIPS[1]	Extended	89.70	87.90	94.90	60.80	95.10	57.10	80.90
ALoc-Aware	Location	89.56	87.96	96.67	60.32	95.65	57.42	81.26
Sent. Embed.	Extended (Sentence)	87.32	85.59	96.67	56.07	95.06	56.90	79.60
Joint Kernel	Extended (Factored)	**89.97**	**88.06**	**98.33**	59.71	96.17	56.78	81.50
Loc-Aware	Location	89.56	86.99	97.14	**62.01**	**97.22**	**57.77**	**81.78**

Empirical Impact of the Extensions. Table 6 evaluates our extensions against the SatMIPS baseline. Analyzing the performance across the six datasets reveals three key findings:

- **Superiority over Baselines:** All our extensions, except Sentence Embedding, and the Loc-Aware baseline outperform the SatMIPS state-of-the-art (80.90% average). This validates the robustness of our spatial contrastive objective.
- **Geographic Prior Strength:** Notably, our base Loc-Aware model relying *solely* on location achieves the highest average accuracy (81.78%). Enforcing simultaneous constraints on location, time, and resolution can be overly restrictive, slightly degrading general-purpose feature learning.
- **Multi-Modal Integration:** When using all three metadata fields, independently factorizing the modalities (Joint Kernel) reaches the best accuracy within the extensions (81.50%), dominating fine-grained datasets (R45, O31, UCM). Conversely, compressing all modalities into a single text sentence via LLM embeddings (*Sent. Embed.*) achieves 79.60%. This suggests that implicit embedding might conflate distinct semantic signals.
- **Exploratory Proof of Concept:** Although these extensions do not surpass the overall performance of the proposed Loc-Aware baseline, they demon-

strate the framework's flexibility. They show that diverse multimodal constraints can easily be integrated into a single objective without the structural complexity of cross-modal mapping networks.

6 Conclusion

We introduced Loc-Aware, a contrastive objective that incorporates geographic proximity into self-supervised learning via a geodesic kernel. Extensive evaluations across six remote sensing benchmarks demonstrate its effectiveness, significantly outperforming SimCLR in average accuracy and consistently surpassing state-of-the-art metadata-driven methods (SatMIP, SatMIPS). These representations generalize exceptionally well across different visual backbones, including ViT-S and ResNet-50. Our exploration into methodological extensions, encompassing adaptive bandwidth calibration and extended metadata encoding, revealed that the purely geographic Loc-Aware baseline yields the strongest general-purpose representations. However, when integrating additional attributes like time and resolution, independently factorizing them via a Joint Metadata Kernel proved superior to a composite sentence embedding. Future work includes developing *learnable* weighting mechanisms to balance multimodal constraints, mitigating generalization bottlenecks. Furthermore, we aim to extend our geodesic objective to dense, patch-level contrastive learning for segmentation tasks, and apply Loc-Aware to multi-temporal sequences encoding seasonal invariances.

Acknowledgments. The authors used an LLM as a writing aid and coding assistant for experimental scripts. All scientific content and conclusions are entirely the work of the authors. This work was performed using HPC resources from GENCI–IDRIS (Grant 2025-AD011017168) and was supported by the French National Research Agency (ANR) under the France 2030 program with the reference ANR-23-IACL-0007.

References

1. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-supervised learning from images with a joint-embedding predictive architecture. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15619–15629 (2023)
2. Ayush, K., Uzkent, B., Meng, C., Tanmay, K., Burke, M., Lobell, D., Ermon, S.: Geography-aware self-supervised learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10181–10190 (2021)
3. Bourcier, J., Dashyan, G., Alahari, K., Chanussot, J.: Learning representations of satellite images from metadata supervision. In: European Conference on Computer Vision. pp. 54–71. Springer (2024)
4. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)

5. Chapelle, O., Weston, J., Bottou, L., Vapnik, V.: Vicinal risk minimization. Advances in neural information processing systems **13** (2000)
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PmLR (2020)
7. Cheng, G., Han, J., Lu, X.: Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE **105**(10), 1865–1883 (2017)
8. Christie, G., Fendley, N., Wilson, J., Mukherjee, R.: Functional map of the world. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6172–6180 (2018)
9. Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D., Ermon, S.: Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. Advances in Neural Information Processing Systems **35**, 197–211 (2022)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
11. Dufumier, B., Castillo-Navarro, J., Tuia, D., Thiran, J.P.: What to align in multimodal contrastive learning? International Conference on Learning Representations (2025)
12. Dufumier, B., Gori, P., Victor, J., Grigis, A., Wessa, M., Brambilla, P., Favre, P., Polosan, M., Mcdonald, C., Piguet, C.M., et al.: Contrastive learning with continuous proxy meta-data for 3d mri classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 58–68. Springer (2021)
13. Ericsson, L., Gouk, H., Loy, C.C., Hospedales, T.M.: Self-supervised representation learning: Introduction, advances, and challenges. IEEE Signal Processing Magazine **39**(3), 42–62 (2022)
14. Fuller, A., Millard, K., Green, J.: Croma: Remote sensing representations with contrastive radar-optical masked autoencoders. Advances in Neural Information Processing Systems **36**, 5506–5538 (2023)
15. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
17. Helber, P., Bischke, B., Dengel, A., Borth, D.: Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **12**(7), 2217–2226 (2019)
18. Hinton, G.E., Roweis, S.: Stochastic neighbor embedding. Advances in neural information processing systems **15** (2002)
19. Jean, N., Wang, S., Samar, A., Azzari, G., Lobell, D., Ermon, S.: Tile2vec: Unsupervised representation learning for spatially distributed data. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 3967–3974 (2019)
20. Klemmer, K., Rolf, E., Robinson, C., Mackey, L., Rußwurm, M.: Satclip: Global, general-purpose location embeddings with satellite imagery. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 4347–4355 (2025)

21. Mañas, O., Lacoste, A., Giró-i Nieto, X., Vazquez, D., Rodríguez, P.: Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9414–9423 (October 2021)
22. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
23. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
24. Reed, C.J., Gupta, R., Li, S., Brockman, S., Funk, C., Clipp, B., Keutzer, K., Candido, S., Uyttendaele, M., Darrell, T.: Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4088–4099 (2023)
25. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks (2019), https://arxiv.org/abs/1908.10084
26. Robusto, C.C.: The cosine-haversine formula. The American Mathematical Monthly **64**(1), 38–40 (1957)
27. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics **20**, 53–65 (1987)
28. Stojnic, V., Risojevic, V.: Self-supervised learning of remote sensing scene representations using contrastive multiview coding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1182–1191 (2021)
29. Wang, Q., Liu, S., Chanussot, J., Li, X.: Scene classification with recurrent attention of vhr remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **57**(2), 1155–1167 (2018)
30. Wang, Y., Albrecht, C.M., Braham, N.A.A., Mou, L., Zhu, X.X.: Self-supervised learning in remote sensing: A review. IEEE Geoscience and Remote Sensing Magazine **10**(4), 213–247 (2022). https://doi.org/10.1109/MGRS.2022.3198244
31. Yang, Y., Newsam, S.: Bag-of-visual-words and spatial extensions for land-use classification. In: Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems. pp. 270–279 (2010)
32. Zhang, X., Lv, Y., Yao, L., Xiong, W., Fu, C.: A new benchmark and an attribute-guided multilevel feature representation network for fine-grained ship classification in optical remote sensing images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **13**, 1271–1285 (2020)
33. Zhu, X.X., Hu, J., Qiu, C., Shi, Y., Kang, J., Mou, L., Bagheri, H., Haberle, M., Hua, Y., Huang, R., et al.: So2sat lcz42: A benchmark data set for the classification of global local climate zones [software and data sets]. IEEE Geoscience and Remote Sensing Magazine **8**(3), 76–89 (2020)