

Article

C²DSSL: Context-Consistency Enhanced Collaborative Self-Supervised Learning for Remote Sensing Image Understanding

Wu Wen ¹, Jinghui Luo ¹, Kailun Qiu ¹, Zhong Xiao ^{2,*} and Gen Lai ¹

¹ School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou 510006, China; wenwu@gzhu.edu.cn (W.W.); kong.ljh@e.gzhu.edu.cn (J.L.); csklqiu@gzhu.edu.cn (K.Q.); laigen781011@gzhu.edu.cn (G.L.)

² School of Mechanical and Electrical Engineering, Guangzhou University, Guangzhou 510006, China

* Correspondence: gzxiaozhong@gzhu.edu.cn

Abstract

Self-Supervised Learning (SSL) has attracted increasing attention in remote sensing image understanding because it can learn transferable representations from unlabeled images. However, two issues remain insufficiently examined in collaborative SSL for remote sensing. First, when high-ratio masking removes entire small objects or structurally informative regions, the remaining visible patches may provide insufficient evidence for semantically coherent reconstruction. Second, heterogeneous self-supervised objectives may exhibit different loss scales, gradient magnitudes, and convergence behaviors such that fixed coefficients can produce uneven branch contributions during training. To address these issues, this paper proposes Context-Consistency Enhanced Collaborative Self-Supervised Learning (C²DSSL) for remote sensing image understanding. C²DSSL introduces a teacher–student context-consistency constraint, in which multi-scale reconstruction features from the complete teacher observation serve as contextual targets for the student network when reconstructing the corresponding masked observation. In addition, gradient-sensitive dynamic weighting uses temporally smoothed loss-gradient magnitudes as empirical signals to adjust the relative contributions of heterogeneous self-supervised objectives. Under the evaluated settings, adding the context-consistency constraint improves KNN representation evaluation, UCMerced classification, Potsdam semantic segmentation, and DOTA oriented object detection, while maintaining the Baseline performance on the Million-AID subset. Dynamic weighting shows task-dependent effects, including a performance gain on the Million-AID subset classification task when combined with CCL.

Keywords: self-supervised learning; remote sensing image understanding; masked image modeling; context consistency; dynamic weighting; collaborative representation learning

MSC: 68T45; 68T07; 68U10



Academic Editor: Iraklis Varlamis

Received: 29 June 2026

Revised: 19 July 2026

Accepted: 21 July 2026

Published: 4 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Remote sensing imagery provides spatially explicit observations of the Earth's surface and supports a wide range of geoscientific applications, including environmental monitoring, land-cover mapping, agricultural assessment, urban analysis, and disaster management [1–4]. Unlike conventional natural images, such imagery is acquired from overhead platforms under diverse spatial resolutions, sensor characteristics, geographic

regions, and environmental conditions. A single scene may contain heterogeneous land-cover categories, densely distributed objects, repeated spatial patterns, arbitrary object orientations, and substantial variations in object scale. Moreover, the semantic interpretation of a local region often depends not only on its visual appearance but also on its spatial relationships with the surrounding geographic structures. Geoscientific analysis may additionally require the integration of multiple spatial variables and thematic layers to characterize a geographic region comprehensively [5]. These characteristics make contextual and multi-scale spatial modeling particularly important for remote sensing image understanding.

Deep learning has achieved considerable progress in scene classification [6], object detection [7], semantic segmentation [8], and change detection [9]. However, its performance commonly depends on sufficiently large and accurately annotated datasets. Producing scene labels, pixel-level segmentation masks, and oriented bounding boxes is costly and often requires specialized domain knowledge. Meanwhile, satellite, airborne, and unmanned-aerial-vehicle platforms continuously produce large quantities of imagery without task-specific annotations. Self-supervised representation learning therefore provides a promising strategy for exploiting these unlabeled data and reducing annotation dependence, particularly when only limited labeled samples are available [10]. Consequently, learning robust and transferable representations from unlabeled geospatial imagery remains an important problem in remote sensing image understanding.

Self-Supervised Learning (SSL) learns transferable visual representations by constructing supervisory signals from unlabeled data [11,12]. Among existing paradigms, contrastive and self-distillation methods primarily learn view-invariant semantic representations, whereas Masked Image Modeling (MIM) learns spatial dependencies by reconstructing partially removed image content. Contrastive methods distinguish or align representations across augmented views [13–17]. In remote sensing, temporal observations, geographic correspondence, and multi-sensor data can provide useful signals for constructing semantically related views [18–20]. These objectives are effective for learning globally discriminative scene representations. However, they primarily optimize image- or view-level agreement and do not explicitly supervise the recovery of object boundaries, fine-grained textures, or spatially localized structures. Moreover, two augmented views of a heterogeneous remote sensing scene may preserve similar global semantics while containing different local regions, which may limit the direct transfer of purely global alignment objectives to dense prediction tasks. In comparison, MIM provides stronger supervision for local structure and spatial dependency modeling through masked reconstruction. Therefore, contrastive learning and MIM offer complementary learning signals: the former emphasizes global semantic invariance, whereas the latter emphasizes local structural recovery.

In contrast, MIM learns visual representations by reconstructing masked pixels or visual tokens from the remaining visible regions [21–23]. This objective encourages the model to capture local textures, spatial dependencies, and fine-grained structural information. Accordingly, recent remote sensing MIM methods have incorporated temporal, multi-scale, cross-view, and contextual information into masked representation learning [24–27]. Rather than assuming that random masking necessarily causes contextual information loss, we identify a conditional limitation: when a high masking ratio removes an entire small object or a structurally informative region from a heterogeneous scene, the remaining visible patches may provide insufficient evidence for semantically coherent recovery. Moreover, minimizing a pixel- or token-level reconstruction error does not explicitly require the recovered region to remain semantically compatible with the complete scene. Although context-enhanced MIM has introduced reconstructive templates and context-consistency

constraints to address this problem [27], how multi-scale contextual reconstruction can be coordinated with global and local self-supervised objectives within a unified teacher–student framework remains insufficiently examined. This constitutes the first research gap addressed in this work.

A second research gap concerns the optimization of heterogeneous objectives in collaborative SSL. Reconstruction, global contrastive learning, and local semantic alignment impose different constraints on the shared representation and may exhibit different numerical scales, gradient magnitudes, and convergence behaviors. Studies in multi-task learning have shown that fixed loss coefficients can produce uneven contributions from different objectives and have proposed uncertainty-, gradient-, or training-rate-based weighting strategies [28–30]. However, these findings do not imply that adaptive weighting will universally improve collaborative self-supervised learning. In particular, the interaction between context-consistent reconstruction and gradient-sensitive loss allocation has received limited analysis in remote sensing SSL. Therefore, this work investigates whether smoothed branch-level gradient information can serve as a practical proxy for adjusting the contributions of heterogeneous self-supervised objectives, while treating its effectiveness as an empirical question rather than an assumed advantage.

To address the above research gaps, this paper proposes Context-Consistency Enhanced Collaborative Self-Supervised Learning (C^2DSSL) for remote sensing image understanding. Unlike collaborative SSL approaches that mainly combine reconstruction and representation-alignment objectives as parallel training signals, C^2DSSL establishes an explicit reconstruction-level context-transfer pathway between the complete teacher observation and the corresponding masked student observation.

Specifically, the teacher network processes the complete image view and generates multi-scale reconstruction features that serve as contextual targets for the student network. The student network processes the masked view and reconstructs the missing regions from the remaining visible patches. A context-consistency constraint aligns the corresponding teacher and student reconstruction features, encouraging the masked reconstruction to remain compatible with the contextual structure of the complete scene. This design is intended to reduce the risk of semantic degradation when the visible patches provide insufficient evidence for reconstructing small or structurally informative regions. Global contrastive learning and local semantic alignment are incorporated as complementary representation-learning objectives, while gradient-sensitive dynamic weighting is investigated as an auxiliary strategy for adjusting their relative contributions during training.

Furthermore, C^2DSSL introduces a gradient-sensitive dynamic weighting mechanism for collaborative self-supervised optimization. Instead of maintaining fixed coefficients throughout training, the proposed strategy adjusts the contribution of each branch according to its temporally smoothed loss-gradient magnitude. In this work, the gradient magnitude is treated as a practical proxy for branch-level optimization pressure rather than a direct measure of optimization difficulty. Under the proposed weighting heuristic, branches with larger persistent gradient magnitudes receive larger normalized weights, whereas branches with smaller magnitudes receive lower weights. Because gradient magnitude may also be affected by loss scale, parameterization, and gradient noise, the effectiveness of this mechanism is evaluated empirically rather than assumed to be universal.

Taken together, C^2DSSL combines complete-view contextual supervision with masked reconstruction and introduces adaptive loss allocation for heterogeneous self-supervised objectives. The framework is evaluated under controlled ablation settings to examine the individual and combined effects of context consistency and dynamic weighting on remote sensing representation learning.

The main contributions of this paper are summarized as follows:

1. We formulate contextual semantic compatibility as a complementary objective to pixel- or token-level masked reconstruction. This formulation distinguishes accurate local reconstruction from maintaining semantic consistency between a masked observation and its corresponding complete scene.
2. We propose C²DSSL, which establishes an explicit reconstruction-level context-transfer pathway from the complete teacher observation to the masked student observation. Multi-scale teacher features provide contextual targets for student reconstruction, while global contrastive learning and local semantic alignment supply complementary representation-learning signals.
3. We investigate temporally smoothed gradient magnitude as an empirical proxy for adjusting the relative contributions of heterogeneous self-supervised objectives. Controlled evaluations across classification, semantic segmentation, and oriented object detection show that context consistency is associated with performance gains across the evaluated tasks, whereas the effect of dynamic weighting is task-dependent.

2. Related Work

This section reviews the studies closely related to this work, including self-supervised learning in remote sensing, contrastive learning, masked image modeling, and collaborative optimization with dynamic weight learning. By analyzing the advantages and limitations of existing methods, we further clarify the necessity of the proposed C²DSSL framework in context-consistent modeling and dynamic multi-branch collaborative optimization.

2.1. Self-Supervised Learning in Remote Sensing

Self-Supervised Learning (SSL) aims to learn generalizable visual representations from large-scale unlabeled data by designing pretext tasks without requiring manual annotations [11,12]. With the rapid growth of Earth observation data, SSL has attracted increasing attention in remote sensing research. Compared with traditional supervised learning, SSL can effectively reduce the dependence on annotated samples, making it particularly suitable for remote sensing interpretation scenarios where data annotation is expensive and requires professional expertise. Currently, SSL has been widely applied to downstream remote sensing tasks, including scene classification, semantic segmentation, object detection, and change detection [20,31–33].

Remote sensing SSL has progressively developed from domain-aware contrastive learning to transformer-based self-supervised pretraining and large-scale foundation models. Seasonal, change-aware, and geography-aware SSL methods exploit temporal observations, geographic correspondence, and geolocation information to construct semantically related samples, thereby improving robustness to temporal, seasonal, and geographic variations [18,19,34,35]. Vision Transformer-based approaches have also been explored for patch-based and multimodal remote sensing representation learning. For example, joint SAR–optical pretraining combines complementary information from the two sensing modalities [36]. Beyond remote-sensing-specific methods, DINOv2 demonstrates that large-scale self-distillation with Vision Transformers can produce general-purpose visual representations, although it is not specifically designed to model the temporal, spectral, and spatial-resolution characteristics of remote sensing imagery [37].

Recent remote-sensing-specific methods incorporate different forms of domain knowledge. SatMAE introduces temporal and multispectral information into masked autoencoding, Scale-MAE explicitly models variations in spatial resolution and ground scale, and RingMo adapts large-scale masked pretraining to scenes containing dense and small objects [24,25,38]. SpectralGPT focuses on spatial–spectral representation learning, while geospatial foundation models explore continual pretraining on domain-specific

imagery [39,40]. In addition, RS³Lip combines self-supervised learning with part-level vision–language consistency to improve class and domain generalization in remote sensing classification [41]. These studies demonstrate the importance of geographic, temporal, spectral, scale-aware, and semantic information in remote sensing representation learning. However, they mainly emphasize view-level representation alignment, domain-specific masked prediction, vision–language consistency, or large-scale pretraining. Relatively limited attention has been paid to transferring reconstruction-oriented contextual information from a complete teacher observation to the corresponding masked student observation while coordinating global, local, and reconstruction objectives within a unified framework. This distinction defines the specific focus of C²DSSL.

2.2. Contrastive Learning

Contrastive Learning (CL) is one of the most important paradigms in SSL. Its core idea is to pull representations of similar samples closer in the feature space while pushing representations of dissimilar samples apart. Representative methods include SimCLR, MoCo, BYOL, and DINO. SimCLR constructs different views of the same image through strong data augmentation and learns image representations using a contrastive loss [13]. MoCo improves the stability and efficiency of contrastive learning by introducing a momentum encoder and a dynamic queue to maintain a large number of negative samples [14,15]. BYOL and DINO further reduce the dependence on negative samples through asymmetric architectures or self-distillation mechanisms, leading to stronger self-supervised representations [16,17]. SimSiam and Barlow Twins also explore alternative approaches to prevent representation collapse without negative samples [42,43]. SwAV introduces online clustering to contrastive learning [44].

In the field of remote sensing, CL has strong applicability. Remote sensing data often contain multi-temporal, multi-sensor, and geospatial attributes, which provide natural conditions for constructing positive pairs. For instance, images of the same geographic region acquired at different times or by different sensors can be regarded as semantically related samples [18,34]. This enables the model to learn representations that are robust to seasonal changes, viewing conditions, and sensor variations [19,20]. Therefore, CL has achieved promising performance in remote sensing scene classification, image retrieval, and large-scale land-cover recognition [45,46].

Nevertheless, CL still has several limitations. First, these methods usually focus on global feature alignment at the image or instance level, while their perception of pixel-level details and local spatial structures is relatively weak. For dense prediction tasks such as object detection and semantic segmentation, the model needs not only to understand global image semantics but also to accurately capture object boundaries, texture details, and local spatial relationships. Second, remote sensing images usually contain multiple land-cover objects, and different augmented views may lead to local semantic inconsistency, introducing noise into positive-pair construction. These problems limit the performance of purely contrastive methods in complex remote sensing image understanding.

2.3. Masked Image Modeling

Masked Image Modeling (MIM) borrows the idea of masked prediction from natural language processing [47]. It randomly masks part of the input image and requires the model to reconstruct the missing content from visible regions, thereby learning local dependencies and spatial structures within images. Representative methods include BEiT, MAE, and SimMIM. BEiT discretizes image patches into visual tokens and predicts the corresponding token IDs of masked patches [23]. MAE adopts an asymmetric encoder–decoder architecture, where only visible patches are encoded and a lightweight decoder

reconstructs the full image [21]. SimMIM further simplifies the reconstruction process by directly predicting the raw pixel values of masked patches with a lightweight prediction head [22].

MIM is particularly relevant to remote sensing because reconstruction-based objectives encourage the modeling of local textures, spatial dependencies, and object structures. Recent studies have extended masked pretraining beyond standard RGB patch reconstruction by incorporating domain-specific information. SatMAE integrates temporal and multispectral structures into the masking and encoding processes, whereas Scale-MAE introduces scale-aware representations to account for variations in spatial resolution and ground sample distance [24,25]. SatMAE++ further investigates transformer pretraining for multi-spectral satellite imagery, and SpectralGPT focuses on learning spatial–spectral foundation representations [39,48]. RingMo employs large-scale masked pretraining for complex scenes containing dense and small objects, while GFM explores continual domain-specific pretraining for geospatial foundation models [38,40]. Cross-view and context-enhanced methods additionally introduce complementary observations or reconstruction-level contextual constraints into masked representation learning [26,27]. Collectively, these approaches improve temporal, spectral, scale-aware, and contextual modeling, although their learning targets, input modalities, and pretraining scales differ substantially. C²DSSL instead focuses on reconstruction-level contextual transfer from a complete teacher observation to a masked student observation within a collaborative multi-objective framework.

Although MIM is effective for local structure modeling, random masking may present a conditional limitation in heterogeneous remote sensing scenes. When a high masking ratio removes an entire small object or a structurally informative region, the remaining visible patches may provide insufficient evidence for semantically coherent reconstruction. Context-enhanced MIM methods have therefore introduced complete-view reconstruction templates and contextual-consistency constraints [27]. These studies suggest that pixel-level reconstruction does not explicitly enforce semantic compatibility between the recovered regions and the complete scene, motivating the incorporation of complementary contextual supervision.

2.4. Collaborative Optimization and Dynamic Weight Learning

With the development of SSL, an increasing number of methods have attempted to combine multiple pretext tasks or learning branches to enhance both global discriminability and local perceptual ability [49,50]. For example, in teacher–student frameworks, the teacher network can provide stable supervision signals based on complete images or augmented views, while the student network learns robust representations through masked inputs, contrastive constraints, or reconstruction tasks [17]. Such collaborative learning frameworks can exploit the complementary information among different tasks, but they also introduce new optimization challenges.

In multi-branch or multi-task learning, different loss terms usually have different numerical scales, gradient magnitudes, and convergence speeds. Simply adding multiple loss terms with fixed weights may cause some tasks to dominate the training process while others are insufficiently optimized, thereby degrading the overall model performance [51,52]. Traditional methods often rely on manually assigned loss weights, but this strategy requires extensive hyperparameter tuning and is difficult to adapt to the dynamic training status of different tasks. Therefore, automatically adjusting the weights of different tasks or branches during training is an important problem in collaborative learning.

Several adaptive weight allocation methods have been proposed in multi-task learning. For example, uncertainty-based weighting methods adjust the importance of different loss terms by modeling task uncertainty [30]; dynamic weight average methods assign weights

according to the descending rates of task losses [53]; gradient-based adaptive weighting methods explore this problem from the perspective of loss gradients [28]. These methods use loss trends or gradient statistics as empirical signals for adapting task weights. However, gradient magnitude is not a direct measure of optimization difficulty because it may also depend on loss scale, parameterization, and stochastic gradient noise. Gradient-based weighting should therefore be regarded as an optimization heuristic whose effectiveness requires empirical evaluation. GradNorm normalizes the gradients of different tasks to regularize multi-task training [29].

For the proposed collaborative SSL framework, adaptive loss allocation is worth investigating because context consistency, masked reconstruction, and representation-alignment objectives may exhibit different numerical scales and optimization behaviors. Fixed coefficients may produce uneven contributions from these objectives during particular training stages, although adaptive weighting does not necessarily eliminate gradient conflicts. Therefore, this paper introduces a gradient-sensitive weighting mechanism to examine whether temporally smoothed loss-gradient information can improve the coordination among local reconstruction, context consistency, and global representation learning. Its effectiveness is evaluated through controlled ablation analysis rather than assumed to guarantee balanced optimization.

3. Materials and Methods

3.1. Overview

This section describes the proposed Context-Consistency Enhanced Collaborative Self-Supervised Learning framework, termed C²DSSL, for remote sensing image understanding. As shown in Figure 1, the proposed framework consists of a multi-view data pipeline, a teacher–student architecture, four self-supervised objectives—masked reconstruction, context consistency, global contrastive learning, and local semantic alignment—and a gradient-sensitive dynamic weighting module. These components are used to investigate complementary reconstruction and representation-learning signals and their joint optimization within a unified framework.

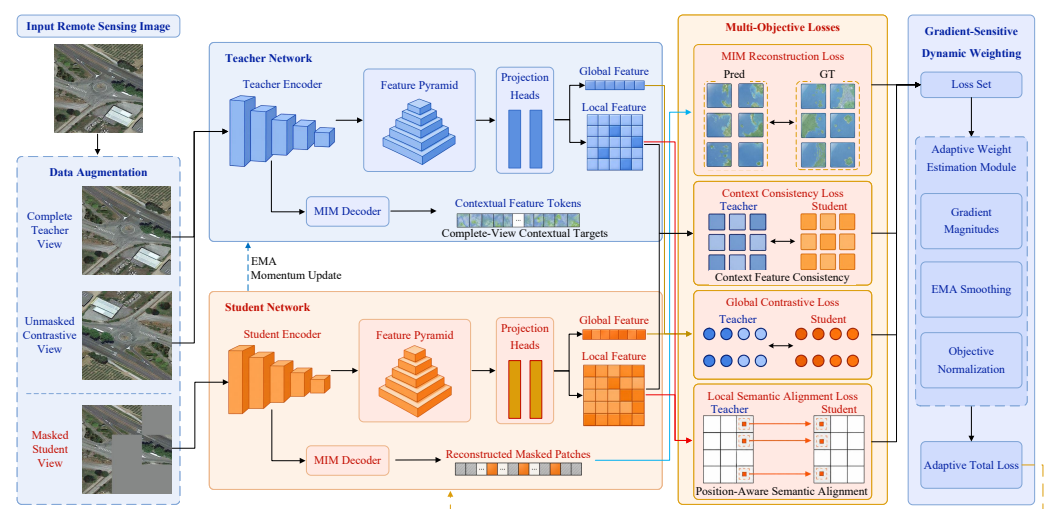


Figure 1. Overview of the proposed C²DSSL framework. The data augmentation module constructs a complete teacher view, a masked student view, and an unmasked contrastive view from the same source image. The unmasked contrastive view is encoded by the EMA teacher network and provides the target representation for image-level global contrastive learning.

Given an input remote sensing image, the data augmentation module generates three views for collaborative self-supervised learning: a complete teacher view, a corresponding

masked student view, and an unmasked contrastive view. The complete teacher view and the unmasked contrastive view are generated from the same source image through standard random augmentation pipelines, including random cropping and color jitter, whereas block-wise masking is applied only to the student view. The complete teacher view provides contextual supervision for reconstruction-level consistency and local semantic alignment, while the unmasked contrastive view is used as the teacher-side target for image-level contrastive learning. The teacher network processes both complete views and is updated by the exponential moving average (EMA) of the student network without direct backpropagation.

The student branch is optimized by multiple complementary objectives. The masked reconstruction branch encourages the student network to recover masked patches and learn fine-grained local structures. The context consistency branch aligns the student reconstruction features with the complete-view contextual features generated by the teacher network and is intended to reduce the risk of semantic degradation when the visible patches provide insufficient reconstruction evidence. The global contrastive learning branch enhances image-level semantic separability, while the local semantic alignment branch enforces spatial correspondence between teacher and student features. Finally, the gradient-sensitive dynamic weighting module uses the gradient magnitude as a practical proxy for branch-level optimization pressure, smooths temporal fluctuations with EMA, normalizes the resulting weights across objectives, and generates an adaptive multi-objective training loss for analyzing the effect of dynamic weighting.

3.2. Teacher–Student Self-Supervised Distillation

Given an input remote sensing image $x \in \mathbb{R}^{H \times W \times C}$, it is first divided into a sequence of non-overlapping patches:

$$X = \{x_i\}_{i=1}^N, \quad x_i \in \mathbb{R}^{P \times P \times C}, \quad (1)$$

where H , W , and C denote the height, width, and number of channels of the image, respectively. P denotes the patch size, and $N = HW/P^2$ is the number of image patches.

To ensure valid patch-level correspondence between the teacher and student branches, the teacher view and the student view share the same geometric transformation, while different photometric augmentations can be applied. The complete teacher view is defined as:

$$X^t = \mathcal{T}_t(X), \quad (2)$$

where $\mathcal{T}_t(\cdot)$ denotes the complete augmented view for the teacher branch. A binary mask $M \in \{0, 1\}^N$ is generated for the student branch, where $M_i = 1$ indicates that the i -th patch is masked, and $M_i = 0$ indicates that the patch remains visible. The masked student view is formulated as:

$$X^s = (1 - M) \odot \mathcal{T}_s(X) + M \odot X_{\text{mask}}, \quad (3)$$

where $\mathcal{T}_s(\cdot)$ denotes the student view aligned with the teacher view, X_{mask} denotes the mask token or masked-patch replacement, and $\odot$ denotes element-wise multiplication.

For global contrastive learning, an additional unmasked contrastive view is generated from the same source image:

$$X^c = \mathcal{T}_c(X), \quad (4)$$

where $\mathcal{T}_c(\cdot)$ denotes the standard random augmentation pipeline used for the contrastive branch, including random cropping and color jitter. Unlike the student view X^s , no patch mask is applied to X^c . The two teacher-side views have different roles: X^t provides

complete-view contextual targets for context consistency and local semantic alignment, whereas X^c provides the target representation for image-level contrastive learning.

The student and teacher encoders are denoted as $E_s(\cdot)$ and $E_t(\cdot)$, respectively. The masked student view, complete teacher view, and unmasked contrastive view are encoded as:

$$\begin{aligned}\{F_s^l\}_{l=1}^L &= E_s(X^s), \\ \{F_t^l\}_{l=1}^L &= E_t(X^t), \\ \{F_c^l\}_{l=1}^L &= E_t(X^c),\end{aligned}\quad (5)$$

where l denotes the feature level. Both complete views are encoded by the EMA teacher branch. The hierarchical features derived from X^t are used for reconstruction-level contextual supervision and local alignment, while those derived from X^c provide the teacher-side target for the global contrastive objective. To integrate multi-scale semantic information, feature pyramid modules aggregate the hierarchical representations:

$$\begin{aligned}P_s &= \Psi_s(\{F_s^l\}_{l=1}^L), \\ P_t &= \Psi_t(\{F_t^l\}_{l=1}^L), \\ P_c &= \Psi_t(\{F_c^l\}_{l=1}^L),\end{aligned}\quad (6)$$

where $\Psi_s(\cdot)$ and $\Psi_t(\cdot)$ denote the feature pyramid modules of the student and teacher branches, respectively. The resulting features are further fed into reconstruction decoders and projection heads for different self-supervised objectives.

The teacher parameters are updated by EMA:

$$\theta_t \leftarrow m\theta_t + (1 - m)\theta_s, \quad (7)$$

where θ_s and θ_t denote the parameters of the student and teacher encoders, respectively, and m is the momentum coefficient. The stop-gradient operation is applied to the teacher branch so that it provides stable supervision without being directly updated by the student losses.

3.3. Masked Reconstruction Branch

The masked reconstruction branch aims to learn local textures, spatial structures, and object boundaries by recovering masked image regions. Remote sensing images usually contain complex land-cover patterns, dense small objects, and large homogeneous areas. Therefore, reconstructing masked regions from incomplete observations encourages the student network to capture both local structural details and spatial dependencies.

In this work, block-wise random masking is adopted to construct the masked student view. The set of masked patches is defined as:

$$\Omega = \{i \mid M_i = 1\}, \quad (8)$$

where $M_i = 1$ indicates that the i -th patch is masked. The student reconstruction decoder $D_s(\cdot)$ reconstructs the masked patches from the student feature pyramid:

$$\hat{X}^s = D_s(P_s). \quad (9)$$

The reconstruction objective consists of a focal L1 loss and a frequency-domain FFT loss:

$$\mathcal{L}_{\text{mim}} = \mathcal{L}_{\text{focal-L1}} + \lambda_{\text{fft}}\mathcal{L}_{\text{fft}}, \quad (10)$$

where λ_{fft} controls the contribution of the FFT loss. For clarity, the reconstruction error of the i -th masked patch is first defined as:

$$e_i = \|\hat{x}_i^s - x_i\|_1. \quad (11)$$

Then, the focal L1 loss is formulated as:

$$\mathcal{L}_{\text{focal-L1}} = \frac{1}{|\Omega|} \sum_{i \in \Omega} (\alpha_f + e_i^{\gamma_f}) e_i, \quad (12)$$

where α_f and γ_f are focal reconstruction parameters. Compared with the standard L1 loss, this formulation assigns larger weights to patches with higher reconstruction errors, allowing the model to pay more attention to difficult masked regions.

The FFT loss constrains the frequency-domain consistency between the reconstructed patches and the original patches:

$$\mathcal{L}_{\text{fft}} = \frac{1}{|\Omega|} \sum_{i \in \Omega} \|\mathcal{F}(\hat{x}_i^s) - \mathcal{F}(x_i)\|_1, \quad (13)$$

where $\mathcal{F}(\cdot)$ denotes the Fourier transform. This reconstruction objective encourages the student network to recover both spatial-domain structures and frequency-domain details from masked remote sensing images.

3.4. Context Consistency Branch

Although masked reconstruction supports local structure modeling, a heavily masked input may provide insufficient contextual evidence when complete small objects or structurally informative regions are removed. In such cases, reconstruction based only on the remaining visible patches may not remain semantically compatible with the complete scene.

To examine this issue, C²DSSL introduces a context consistency branch into the masked reconstruction process. The teacher network receives a complete augmented view and therefore preserves complete-view contextual information that can serve as a reconstruction target for the masked student branch. A teacher reconstruction decoder $D_t(\cdot)$ generates context-aware reconstruction templates:

$$\hat{X}^t = D_t(P_t). \quad (14)$$

The teacher reconstruction output is used as a contextual template to guide the student reconstruction. For the multi-scale implementation, the student and teacher decoders generate stage-specific reconstruction outputs, denoted as $\{\hat{X}^{\text{stu},r}\}_{r=1}^R$ and $\{\hat{X}^{\text{tea},r}\}_{r=1}^R$, respectively. To provide multi-scale contextual supervision, the context consistency loss is computed across multiple feature stages:

$$\mathcal{L}_{\text{ccl}} = \frac{1}{R} \sum_{r=1}^R \frac{1}{|\Omega_r|} \sum_{i \in \Omega_r} \left\| \hat{x}_i^{\text{stu},r} - \text{sg}(\hat{x}_i^{\text{tea},r}) \right\|_2^2, \quad (15)$$

where R denotes the number of feature stages used for multi-scale context consistency, and $R = 4$ is used in this work. Ω_r denotes the masked patch set at the r -th feature stage, and $\text{sg}(\cdot)$ denotes the stop-gradient operation. Different from the basic reconstruction loss, $\mathcal{L}_{\text{ccl}}$ does not only force the student output to approach raw pixels. Instead, it encourages the student reconstruction to remain consistent with the teacher-provided contextual templates at different semantic levels, thereby improving semantic coherence between local masked regions and global scene structures.

3.5. Global Contrastive Learning Branch

The global contrastive learning branch operates on the masked student view X^s and the unmasked contrastive view X^c generated from the same source image. Their feature pyramids are globally pooled and mapped into a dedicated contrastive embedding space:

$$z_s = q_g(\text{GAP}(P_s)), \quad z_c = q_g(\text{GAP}(P_c)), \quad (16)$$

where $q_g(\cdot)$ denotes the global projection head, and $\text{GAP}(\cdot)$ denotes global average pooling. Because X^s and X^c originate from the same source image, z_s and z_c form an image-level positive pair. The teacher-encoded representation z_c is used as a stop-gradient target, while teacher representations from other images are maintained as negative samples in the dynamic queue $\mathcal{Q}$. The global contrastive loss is defined as:

$$\begin{aligned} \mathcal{L}_{\text{gcl}} &= -\log \frac{\exp(\text{sim}(z_s, \text{sg}(z_c))/\tau)}{\mathcal{D}}, \\ \mathcal{D} &= \exp(\text{sim}(z_s, \text{sg}(z_c))/\tau) \\ &\quad + \sum_{z_q \in \mathcal{Q}} \exp(\text{sim}(z_s, z_q)/\tau). \end{aligned} \quad (17)$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity, τ is the temperature parameter, and z_q denotes a teacher representation from another image stored in the queue. This objective aligns two differently processed observations of the same source image at the global representation level and provides a complementary image-level learning signal to the reconstruction and local alignment objectives.

3.6. Local Semantic Alignment Branch

Remote sensing interpretation tasks such as object detection and semantic segmentation require not only image-level discrimination but also local spatial perception. Therefore, C²DSSL introduces a local semantic alignment branch to enforce fine-grained consistency between spatially corresponding teacher and student features.

For each spatial position p in the aligned feature maps, local projection heads are used to obtain local semantic representations:

$$u_s^p = q_l(P_s^p), \quad u_t^p = q_l(P_t^p), \quad (18)$$

where $q_l(\cdot)$ denotes the local projection head, and P_s^p and P_t^p denote the student and teacher features at spatial position p , respectively. To model local semantic distributions, a set of K semantic prototypes $\mathcal{C} = \{c_k\}_{k=1}^K$ is introduced. The teacher and student local distributions are computed as:

$$p_t^p(k) = \frac{\exp(\text{sim}(u_t^p, c_k)/\tau_t)}{\sum_{j=1}^K \exp(\text{sim}(u_t^p, c_j)/\tau_t)}, \quad (19)$$

$$p_s^p(k) = \frac{\exp(\text{sim}(u_s^p, c_k)/\tau_s)}{\sum_{j=1}^K \exp(\text{sim}(u_s^p, c_j)/\tau_s)}, \quad (20)$$

where τ_t and τ_s denote the temperature parameters of the teacher and student branches, respectively. The student temperature τ_s is fixed at 0.1 throughout pretraining. The teacher temperature τ_t is linearly increased from 0.04 to 0.07 during the first 10 epochs and is then maintained at 0.07 for the remaining epochs. The lower initial teacher temperature produces a sharper semantic target distribution, while the gradual increase progressively softens the teacher targets and avoids an abrupt change in their distribution during the early training stage. The local semantic alignment loss is then defined as:

$$\mathcal{L}_{\text{icl}} = -\frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \sum_{k=1}^K \text{sg}(p_t^p(k)) \log p_s^p(k), \quad (21)$$

where $\mathcal{P}$ denotes the set of spatially aligned positions. Since the teacher and student views share the same geometric transformation, this branch can perform position-wise semantic matching between corresponding local regions. In this way, the model can learn local semantic consistency while preserving spatial correspondence.

3.7. Gradient-Sensitive Dynamic Weighting

In collaborative self-supervised learning, different loss branches may exhibit different numerical scales, convergence behaviors, and gradient magnitudes. When multiple losses are combined using fixed manually assigned coefficients, their relative contributions to the shared representation may vary substantially during training. To investigate whether these changing contributions can be adjusted adaptively, C²DSSL introduces a gradient-sensitive dynamic weighting mechanism.

Let the set of self-supervised losses be:

$$\mathcal{S} = \{\mathcal{L}_{\text{mim}}, \mathcal{L}_{\text{ccl}}, \mathcal{L}_{\text{gcl}}, \mathcal{L}_{\text{icl}}\}. \quad (22)$$

For the k -th loss branch, the gradient magnitude is computed with respect to the corresponding branch-related parameters:

$$g_k^{(t)} = \left\| \nabla_{\theta_k} \mathcal{L}_k^{(t)} \right\|_2, \quad (23)$$

where t denotes the training iteration. In this work, $g_k^{(t)}$ is used as a practical proxy for the instantaneous optimization pressure generated by the k -th objective. A persistently large gradient magnitude suggests that the objective continues to exert a strong influence on its associated parameters, which motivates assigning it a larger weight under the proposed heuristic. Conversely, an objective with a smaller smoothed gradient magnitude receives a lower weight. It should be noted that gradient magnitude is not a direct measure of optimization difficulty, because it may also be affected by the numerical scale of the loss, parameterization, and stochastic gradient noise. The proposed relationship is therefore treated as an empirical weighting heuristic and is evaluated through the subsequent ablation experiments.

To reduce unstable fluctuations caused by a single iteration, an exponential moving average is used to smooth the gradient magnitude:

$$\bar{g}_k^{(t)} = \rho \bar{g}_k^{(t-1)} + (1 - \rho) g_k^{(t)}, \quad (24)$$

where ρ is the smoothing coefficient and is set to 0.95 in all experiments. A larger ρ retains a longer history of gradient observations and produces smoother but less responsive weight updates, whereas a smaller ρ responds more rapidly to current gradients but is more sensitive to iteration-level noise. With $\rho = 0.95$, each update retains 95% of the historical estimate and incorporates 5% of the current observation. This value is selected as an empirical compromise that suppresses short-term fluctuations while preserving sufficient responsiveness during the 30-epoch pretraining schedule. It should not be interpreted as universally optimal, because the preferred degree of smoothing may vary with the training duration, objective combination, and gradient statistics. The dynamic weight of the k -th branch is then calculated as:

$$w_k^{(t)} = \frac{\bar{g}_k^{(t)} + \epsilon}{\sum_{j=1}^{K_{\text{loss}}} (\bar{g}_j^{(t)} + \epsilon)}, \quad (25)$$

where K_{loss} is the number of loss branches, and ϵ is a stabilization constant set to 10^{-8} in all experiments. It provides a small numerical floor and prevents the denominator from becoming zero when the smoothed gradient magnitudes are extremely small. When $\bar{g}_k^{(t)}$ is substantially larger than ϵ , changing ϵ within a similarly small range has a negligible influence on the relative weights. In contrast, an excessively large ϵ would reduce the differences among the normalized weights and move them toward a uniform distribution, thereby weakening the responsiveness of dynamic weighting. The value 10^{-8} is therefore selected to provide numerical protection while remaining sufficiently small to avoid materially altering the normalized weights under ordinary nonzero-gradient conditions. The weights form a convex combination:

$$\sum_{k=1}^{K_{\text{loss}}} w_k^{(t)} = 1. \quad (26)$$

During implementation, the dynamic weights are detached from the computational graph to avoid second-order gradient computation.

3.8. Overall Objective

The dynamically weighted training objective of $C^2\text{DSSL}$ is formulated as a combination of the self-supervised branches:

$$\mathcal{L}_{C^2\text{DSSL}} = w_{\text{mim}}\mathcal{L}_{\text{mim}} + w_{\text{ccl}}\mathcal{L}_{\text{ccl}} + w_{\text{gcl}}\mathcal{L}_{\text{gcl}} + w_{\text{lcl}}\mathcal{L}_{\text{lcl}}. \quad (27)$$

This objective jointly considers masked reconstruction, context consistency, global semantic discrimination, local semantic alignment, and adaptive multi-objective optimization, and it is mainly used to analyze the effect of dynamic weighting when combined with context consistency.

For ablation analysis, four configurations are considered. The baseline objective without DW and CCL is defined as:

$$\mathcal{L}_A = \alpha\mathcal{L}_{\text{mim}} + \beta\mathcal{L}_{\text{gcl}} + \gamma\mathcal{L}_{\text{lcl}}, \quad (28)$$

where α , β , and γ are fixed balancing coefficients.

When only the gradient-sensitive dynamic weighting mechanism is introduced, the objective is formulated as:

$$\mathcal{L}_B = \sum_{k \in \{\text{mim}, \text{gcl}, \text{lcl}\}} w_k \mathcal{L}_k. \quad (29)$$

When only the context-consistency constraint is introduced, the objective is formulated as:

$$\mathcal{L}_C = \alpha\mathcal{L}_{\text{mim}} + \beta\mathcal{L}_{\text{gcl}} + \gamma\mathcal{L}_{\text{lcl}} + \lambda_{\text{ccl}}\mathcal{L}_{\text{ccl}}, \quad (30)$$

where λ_{ccl} controls the contribution of the context consistency loss.

When both DW and CCL are introduced, the complete objective is formulated as:

$$\mathcal{L}_D = \sum_{k \in \{\text{mim}, \text{ccl}, \text{gcl}, \text{lcl}\}} w_k \mathcal{L}_k. \quad (31)$$

Here, $\mathcal{L}_A$ denotes the baseline objective, $\mathcal{L}_B$ introduces only the gradient-sensitive dynamic weighting mechanism, $\mathcal{L}_C$ introduces only the context-consistency constraint, and $\mathcal{L}_D$ denotes the DW+CCL variant. In the following ablation analysis, $\mathcal{L}_C$ is retained as a strong context-consistency variant, while $\mathcal{L}_D$ is used to examine whether dynamic weighting can further improve the context-consistency framework across different downstream tasks.

4. Results and Discussion

4.1. Experimental Settings

This section presents the experimental settings, quantitative results, and discussion of the proposed C²DSSL framework. The datasets used in this section include UCMerced [54], Million-AID [55], ISPRS Potsdam, and DOTA v1.0 [56]. UCMerced is used as a compact unlabeled pretraining corpus to evaluate C²DSSL under a controlled limited-data setting. During self-supervised pretraining, category labels are strictly excluded and are introduced only during downstream fine-tuning and evaluation. The purpose of this protocol is not to simulate large-scale foundation-model pretraining but to examine whether the proposed contextual supervision and objective-coordination strategies can learn useful representations when unlabeled pretraining data are limited.

The downstream evaluations contain both in-domain and cross-dataset settings. UCMerced KNN evaluation and classification fine-tuning provide controlled in-domain analyses, whereas the Million-AID subset, Potsdam, and DOTA evaluate transfer to different image distributions and task structures. In particular, the sampled Million-AID subset contains 20,000 images from 51 fine-grained categories and is used to examine adaptation to a more category-diverse classification setting. Therefore, the current experiments primarily evaluate data-efficient representation learning and cross-dataset transferability under constrained pretraining data, rather than large-scale pretraining scalability.

For Million-AID, this paper does not use the full dataset. Considering device limitations, computational cost, and experimental consistency, 20,000 images are sampled from 51 fine-grained categories to construct a Million-AID subset for supervised fine-tuning and transfer analysis. This subset is used to evaluate whether the learned representations can adapt to a larger and more fine-grained remote sensing scene distribution. ISPRS Potsdam is used for semantic segmentation, and DOTA v1.0 is used for oriented object detection. All internal configurations are trained and evaluated under the same experimental protocol and fixed random seed to maintain consistency across the controlled comparisons. Since repeated runs were not conducted, the reported results do not provide estimates of run-to-run variance.

For downstream evaluation, the self-supervised pretrained ResNet-50 backbone is used as initialization, while the task-specific classification, segmentation, and detection heads are randomly initialized. For UCMerced and the Million-AID subset, image classification is performed under the same fine-tuning protocol for all configurations. For ISPRS Potsdam semantic segmentation, the pretrained backbone is transferred to a segmentation framework, and the mean Intersection over Union (mIoU) is adopted as the evaluation metric. For DOTA v1.0 oriented object detection, the pretrained backbone is transferred to an oriented object detection framework, and the mean Average Precision (mAP) is used for evaluation. All downstream experiments use identical training settings across Configs A–D to ensure fair comparison. The experiments were implemented in Python 3.9 using the PyTorch 1.12.0 deep learning framework with CUDA 11.3 acceleration.

The key hyperparameter settings for self-supervised pretraining and multi-objective losses are summarized in Table 1.

Table 1. Key hyperparameter settings for self-supervised pretraining and multi-objective losses.

Parameter Category	Parameter Name	Value
<i>Self-supervised Pretraining</i>		
Backbone	Network architecture	ResNet-50
Pretraining dataset	Dataset usage	UCMerced, all images used without labels
Input size	Image resolution	224×224
Mask setting	Patch size/mask ratio	$32 \times 32/60\%$
Training schedule	Epochs/batch size	30/64
Optimizer	Optimizer/weight decay	AdamW/0.02
Learning rate	Initial LR/scheduler	1×10^{-3} /cosine annealing
Warmup	Epochs/minimum LR	$5/1 \times 10^{-6}$
EMA teacher	Momentum schedule	$0.996 \rightarrow 1.0$
Random seed	Seed	322
<i>Multi-objective Losses</i>		
Global contrastive loss	Temperature/queue size	0.2/4,096
Local semantic alignment	Spatial matches	20
Local semantic alignment	Student/teacher temperature	0.1 (fixed)/0.04 $\rightarrow$ 0.07 (linear warm-up, 10 epochs)
Reconstruction loss	Loss type	Focal L1 + FFT loss
Focal L1 loss	α_f/γ_f	0.2/1.0
FFT loss	λ_{fft}	1.0
Context consistency loss	λ_{ccl} /multi-scale	2.0/4-stage average
Dynamic weighting	ρ/ϵ	$0.95/1 \times 10^{-8}$
Fixed loss coefficients	MIM/GCL/LCL/CCL	0.5/1.5/1.0/2.0

4.2. Experimental Configurations and Metrics

To analyze the contribution of each component, four experimental configurations are designed, as shown in Table 2. Config A denotes the baseline objective without dynamic weighting (DW) and context consistency loss (CCL). Config B only introduces the gradient-sensitive dynamic weighting mechanism. Config C only introduces the context consistency loss. Config D combines both DW and CCL to evaluate whether adaptive loss weighting can further improve the context-consistency variant under different downstream settings. In addition to these internal ablations, external SSL comparisons and a masking-ratio sensitivity analysis are conducted to evaluate cross-method transfer performance and the robustness of the default masking setting.

Table 2. Ablation configurations. Here, $\checkmark$ indicates that the corresponding component is included, whereas $\times$ indicates that it is excluded.

Config	DW	CCL	Description
A	$\times$	$\times$	Baseline
B	$\checkmark$	$\times$	Baseline with DW
C	$\times$	$\checkmark$	Baseline with CCL
D	$\checkmark$	$\checkmark$	DW+CCL variant

DW adaptively adjusts the weights of different loss terms according to their gradient magnitudes and is used to investigate whether branch-level optimization imbalance can be alleviated during collaborative self-supervised training. CCL aligns the student reconstruction features with the complete-view contextual features provided by the teacher network and is intended to reduce the risk of semantic degradation when visible patches provide insufficient reconstruction evidence.

For self-supervised representation evaluation, KNN Top-1 Accuracy is adopted. This metric evaluates the semantic separability of the learned representations without fine-tuning the pretrained model. For image classification, Overall Accuracy (OA) is used as the evaluation metric. For semantic segmentation, mean Intersection over Union (mIoU) is used to evaluate pixel-level dense prediction performance. For oriented object detection, mAP is adopted to evaluate the detection ability of the model for multi-category oriented objects in remote sensing images.

For external comparison, SeCo [18] and GASSL [19] are included as representative remote sensing self-supervised learning methods. Their publicly released ResNet-50 pretrained encoders are evaluated using the same downstream datasets, data splits, task-specific heads, and evaluation metrics adopted for the proposed framework. Because the external encoders are pretrained using different data corpora and training schedules, the results are interpreted as external transfer references rather than strictly controlled comparisons under identical pretraining conditions.

For masking-ratio sensitivity analysis, Config C is evaluated using masking ratios of 0.40, 0.60, and 0.80, while the backbone, pretraining duration, random seed, and other optimization settings are kept unchanged. Config C is used to isolate the influence of context consistency from the task-dependent behavior of dynamic weighting. The masking ratio of 0.60 corresponds to the default setting adopted in the main experiments.

4.3. Self-Supervised Representation Evaluation

To first evaluate the influence of different modules on self-supervised representation quality, KNN evaluation is conducted on UCMerced. This evaluation does not require additional fine-tuning and therefore provides a non-parametric indication of the semantic organization of the pretrained features. The results are shown in Table 3.

Table 3. KNN evaluation results on UCMerced. Bold values indicate the best result among Configs A–D.

Config	KNN Top-1 (%)	Change over Baseline
A	60.31	–
B	61.25	+0.94
C	65.16	+4.85
D	59.38	–0.93

As shown in Table 3, Config C achieves the best KNN Top-1 Accuracy of 65.16%, outperforming the Baseline by 4.85 percentage points. Under the current experimental setting, this result indicates that the context consistency loss improves the KNN representation evaluation performance of the self-supervised pretrained model. Remote sensing images usually contain complex backgrounds, dense small objects, and multi-scale spatial structures. A pure reconstruction objective may overemphasize local pixel recovery while ignoring the semantic relationship between local regions and the global scene. The observed KNN gain is consistent with improved semantic organization of the pretrained feature space under the current experimental setting.

Config B only improves the result by 0.94 percentage points, suggesting that DW alone can slightly improve representation quality but its effect is relatively limited. Config D performs lower than the Baseline in KNN evaluation, indicating that jointly introducing DW and CCL does not provide an additional benefit in this setting. The underlying cause requires further analysis, because the result may be affected by loss interactions, weighting dynamics, or training variability.

4.4. Image Classification Results

To further analyze the influence of different modules on classification tasks, fine-tuning experiments were conducted on UCMerced and Million-AID. To avoid confusion between different datasets, the UCMerced fine-tuning results and the Million-AID fine-tuning results are reported separately.

4.4.1. UCMerced Fine-Tuning Results

Table 4 gives the classification fine-tuning results on UCMerced.

Table 4. Classification fine-tuning results on UCMerced. Bold values indicate the best result among Configs A–D.

Config	OA (%)	Change over Baseline
A	98.96	–
B	99.48	+0.52
C	100.00	+1.04
D	99.48	+0.52

As shown in Table 4, Config C achieves the best OA of 100.00% on UCMerced, improving the Baseline by 1.04 percentage points. Config B and Config D both achieve 99.48% OA, outperforming the Baseline by 0.52 percentage points. This result suggests that CCL can benefit both unsupervised representation evaluation and downstream classification under the current experimental setting. Since UCMerced is a relatively small-scale scene classification dataset, contextual representations learned during pretraining may benefit classification with limited supervised samples.

4.4.2. Million-AID Fine-Grained Classification Results

Table 5 reports the results on the Million-AID subset fine-grained classification task.

Table 5. Fine-grained classification results on the Million-AID subset. Bold values indicate the best result among Configs A–D.

Config	OA (%)	Change over Baseline
A	93.85	–
B	93.65	–0.20
C	93.85	0.00
D	94.48	+0.63

As shown in Table 5, Config D achieves the highest OA of 94.48%, improving the Baseline by 0.63 percentage points. This result indicates that DW and CCL can produce a task-dependent complementary effect in fine-grained semantic discrimination tasks. Compared with UCMerced, Million-AID contains more categories and more subtle inter-class differences. Therefore, the model needs to capture both global scene semantics and fine-grained local structural cues. The improvement of Config D suggests that the combination of dynamic weighting and contextual consistency can be beneficial in the Million-AID fine-grained classification setting, although this advantage is not consistently observed across all downstream tasks.

Figure 2 shows the performance of different configurations under two UCMerced evaluation settings: direct UCMerced fine-tuning and UCMerced evaluation initialized with the Million-AID-finetuned backbone. It should be noted that the second setting in Figure 2 does not denote the Million-AID classification accuracy itself. Instead, it represents the UCMerced classification result initialized with the Million-AID-finetuned backbone,

which is different from the Million-AID subset fine-grained classification task reported in Table 5.

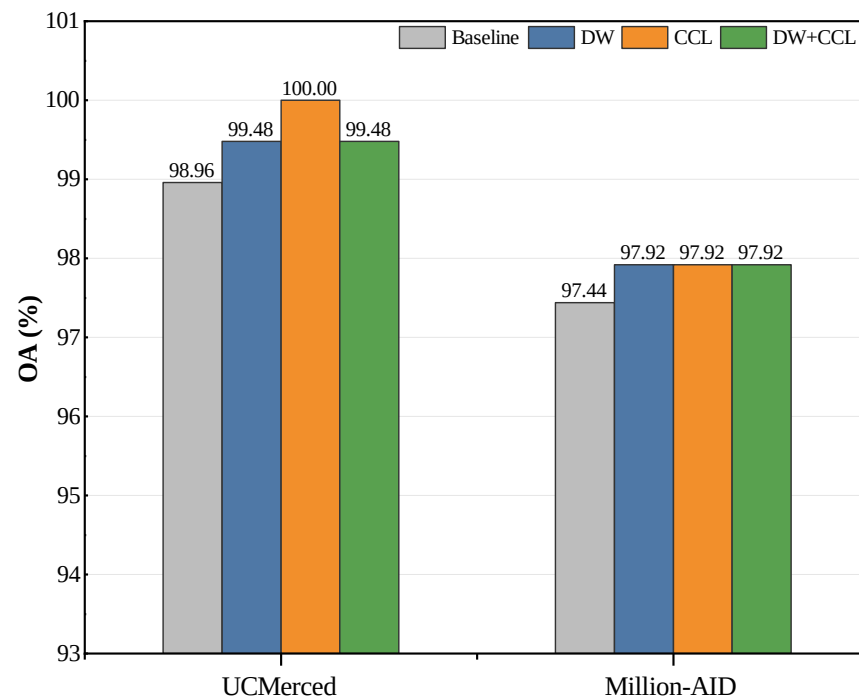


Figure 2. Performance comparison of different configurations under direct UCMerced fine-tuning and UCMerced evaluation using the Million-AID-finetuned backbone.

As shown in Figure 2, Config C achieves the best performance in the direct UCMerced fine-tuning setting. When the Million-AID-finetuned backbone is transferred back to UCMerced, Config B, Config C, and Config D all achieve 97.92% OA, which is higher than the Baseline of 97.44%. However, compared with direct UCMerced fine-tuning, the two-stage transfer results are generally lower, suggesting that supervised fine-tuning on the Million-AID subset may shift the feature space toward the sampled Million-AID category distribution and weaken its adaptability to the small-scale UCMerced classification task.

4.5. Semantic Segmentation and Oriented Object Detection

To evaluate the effectiveness of the proposed method on dense prediction tasks, experiments are further conducted on Potsdam semantic segmentation and DOTA oriented object detection. The results are shown in Table 6.

Table 6. Results on Potsdam semantic segmentation and DOTA oriented object detection. Bold values indicate the best performance in each metric column.

Config	Potsdam mIoU (%)	DOTA mAP (%)
A	33.39	43.62
B	33.60	44.15
C	35.82	45.24
D	34.25	44.78

As shown in Table 6, Config C achieves the best mIoU of 35.82% on Potsdam, improving the Baseline by 2.43 percentage points. For the DOTA oriented object detection task, Config C also obtains the highest mAP of 45.24%, improving the Baseline by

1.62 percentage points. These results show that CCL is more consistently effective than DW for dense prediction tasks in the current experimental setting.

Semantic segmentation and oriented detection both require strong spatial structural perception. Potsdam segmentation requires pixel-level classification of ground objects such as buildings, roads, and vegetation, while DOTA detection requires accurate recognition of multi-scale objects with arbitrary orientations. Compared with scene classification, dense prediction tasks rely more heavily on local structures, object boundaries, and contextual relationships. CCL encourages consistency between local reconstruction and complete-view contextual features during pretraining, which may contribute to the observed gains on spatially sensitive downstream tasks.

Figure 3 further visualizes the performance differences among different configurations on dense prediction tasks. Config C achieves the best results on both Potsdam and DOTA, and its improvement is more obvious than that of DW alone. Config D also outperforms the Baseline but does not surpass Config C, indicating that DW does not provide an additional benefit over CCL in these dense prediction settings. The underlying cause requires further analysis.

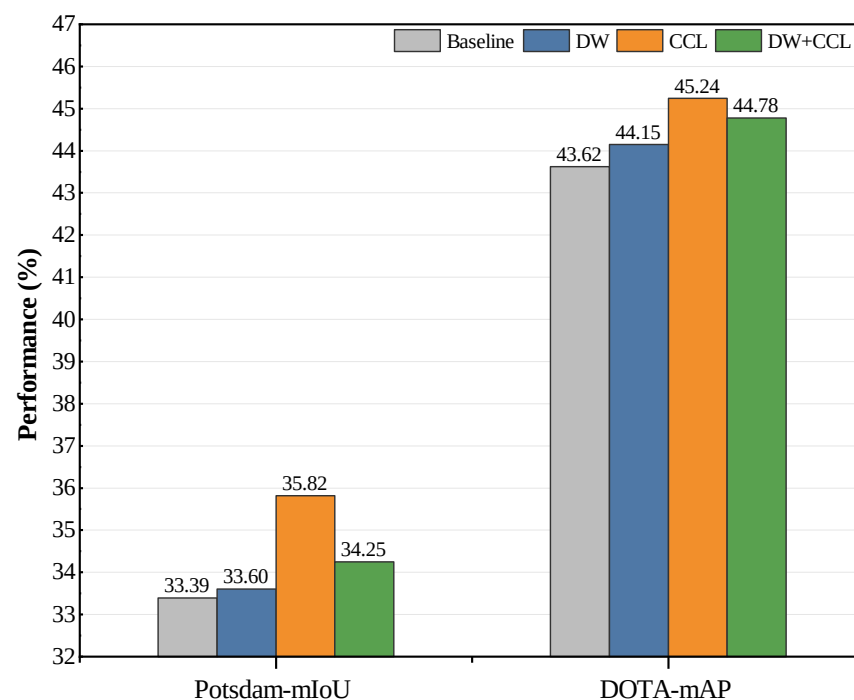


Figure 3. Performance comparison of different configurations on Potsdam semantic segmentation and DOTA oriented object detection.

From a task-specific perspective, the improvement on Potsdam suggests that context-consistent pretraining provides transferable representations for pixel-level land-cover discrimination, where spatial relationships and local structural cues are important. Similarly, the improvement on DOTA indicates that the learned representations can transfer to localization-sensitive recognition involving multi-scale and arbitrarily oriented objects. The consistent gains over the Baseline on both tasks provide task-level evidence that the proposed contextual supervision is not restricted to scene-level classification but can also benefit spatially sensitive dense prediction. Because matched prediction visualizations are not included, the current analysis focuses on task-level transfer performance rather than fine-grained pixel- or instance-level error attribution.

4.6. Performance Changes over the Baseline

To more intuitively compare the contribution of different modules, Figure 4 visualizes the performance changes of DW, CCL, and DW+CCL over the Baseline across different downstream tasks. The compared evaluation settings include direct UCMerced classification, UCMerced classification initialized with the Million-AID-finetuned backbone, Potsdam semantic segmentation, and DOTA oriented object detection.

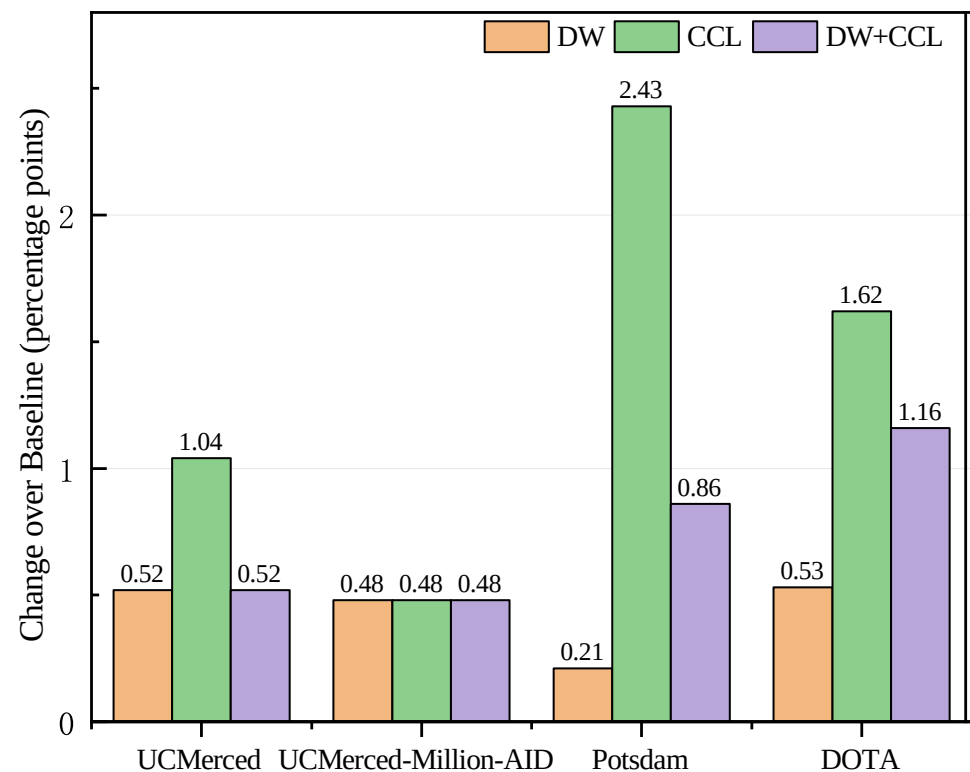


Figure 4. Performance changes of different modules over the Baseline across discrete downstream evaluation settings.

As shown in Figure 4, CCL produces performance gains across the examined downstream tasks under the current experimental settings. Specifically, CCL improves UCMerced classification OA by 1.04 percentage points, Potsdam mIoU by 2.43 percentage points, and DOTA mAP by 1.62 percentage points. These results suggest that context consistency may be beneficial for tasks that rely on spatial context and dense structural information.

In contrast, the effect of DW is not consistently positive. Config D decreases the UCMerced KNN Top-1 accuracy from 60.31% to 59.38%, and Config B decreases the Million-AID subset classification accuracy from 93.85% to 93.65%. Moreover, although Config D remains above the Baseline on the evaluated dense prediction tasks, it does not outperform the CCL-only configuration. These observations indicate that DW has a task-dependent interaction with the other self-supervised objectives rather than providing a uniformly beneficial optimization effect.

Under the current formulation, each dynamic weight is proportional to the corresponding temporally smoothed gradient magnitude. Because raw gradient magnitude may also be affected by loss scale, parameterization, and stochastic noise, an objective can receive a relatively large weight even when its gradient does not directly reflect unmet learning requirements. Introducing CCL additionally changes the number and distribution of the optimized objectives, which may further alter their relative weights. This mechanism provides a plausible explanation for the unfavorable results of Config D, but the current experiments

do not establish the existence of severe optimization conflicts or identify a single causal factor. Therefore, DW is retained as an auxiliary empirical strategy whose effectiveness depends on the objective combination and downstream task. Scale-normalized gradients, weight clipping, constrained weight ranges, or stage-wise activation may provide more robust alternatives in future work.

It should be noted that Figure 4 compares performance changes across discrete downstream tasks rather than showing a continuous trend. Therefore, this figure should be interpreted as a module-wise comparison of gains in percentage points under different evaluation settings.

4.7. Comparison with Representative Remote Sensing SSL Methods

To provide an external reference beyond the internal ablation configurations, the proposed framework is compared with SeCo and GASSL using a common downstream evaluation pipeline. The comparison covers non-parametric KNN evaluation, scene classification, semantic segmentation, and oriented object detection. Since the external methods use different pretraining corpora and training schedules, the results should be interpreted as a comparison of downstream transfer performance rather than a strictly controlled comparison under identical self-supervised pretraining conditions. The results are reported in Table 7.

Table 7. Comparison with representative remote sensing self-supervised learning methods. Bold values indicate the best performance in each column; tied best results are all shown in bold.

Method	KNN Top-1 (%)	UCMerced OA (%)	Million-AID OA (%)	Potsdam mIoU (%)	DOTA mAP (%)
Baseline (Config A)	60.31	98.96	93.85	33.39	43.62
SeCo [18]	66.00	99.48	94.20	34.80	44.90
GASSL [19]	67.00	100.00	94.40	35.20	45.28
C ² DSSL (Config C)	65.16	100.00	93.85	35.82	45.24
C ² DSSL (Config D)	59.38	99.48	94.48	34.25	44.78

As shown in Table 7, GASSL achieves the highest KNN Top-1 accuracy of 67.00%, followed by SeCo at 66.00% and C²DSSL Config C at 65.16%. On UCMerced classification, Config C and GASSL both achieve an OA of 100.00%, while SeCo reaches 99.48%. On Million-AID, C²DSSL Config D achieves the highest OA of 94.48%, marginally surpassing GASSL at 94.40%. For dense prediction, Config C achieves the highest Potsdam mIoU of 35.82%, and obtains a DOTA mAP of 45.24%, which is close to the highest result of 45.28% achieved by GASSL. Overall, C²DSSL achieves or ties the best result on UCMerced classification, Million-AID classification, and Potsdam semantic segmentation, while remaining competitive on KNN evaluation and DOTA oriented object detection. Nevertheless, these results should be interpreted together with the differences in pretraining data and training schedules among the compared methods.

4.8. Masking-Ratio Sensitivity Analysis

To examine the influence of the amount of visible image information on context-consistent representation learning, Config C was evaluated using masking ratios of 0.40, 0.60, and 0.80. All other self-supervised pretraining and downstream evaluation settings were kept unchanged. The masking ratio of 0.60 corresponds to the default setting used in the main experiments. The results are reported in Table 8.

As shown in Table 8, the default masking ratio of 0.60 achieves the best KNN Top-1 accuracy of 65.16% and UCMerced OA of 100.00%. Reducing the masking ratio to 0.40 results in a KNN Top-1 accuracy of 63.80% and an OA of 99.48%. Although the lower masking ratio retains more visible image information, it also reduces the difficulty of the reconstruction objective and may provide a weaker representation-learning signal.

Increasing the masking ratio to 0.80 decreases the KNN Top-1 accuracy to 62.30% and the OA to 98.96%, suggesting that excessive masking leaves insufficient visible contextual evidence for reliable semantic reconstruction. These results indicate that a moderate masking ratio of 0.60 provides a favorable balance between reconstruction difficulty and available contextual information under the current limited-data pretraining setting.

Table 8. Sensitivity analysis of the masking ratio under Config C. Bold values indicate the best performance in each metric column.

Masking Ratio	KNN Top-1 (%)	UCMerced OA (%)
0.40	63.80	99.48
0.60	65.16	100.00
0.80	62.30	98.96

4.9. Two-Stage Transfer Analysis with the Million-AID-Finetuned Backbone

To further analyze the influence of supervised classification fine-tuning on downstream transfer, the backbone fine-tuned on the sampled Million-AID subset was used as initialization and further transferred to UCMerced classification, Potsdam semantic segmentation, and DOTA oriented object detection. Table 9 compares the results obtained using the UCMerced-pretrained backbone and the Million-AID-finetuned backbone.

Table 9. Two-stage transfer results using the Million-AID-finetuned backbone. Bold values indicate the best result in each numerical column within each downstream-task block; tied best results are all shown in bold.

Downstream Task	Config	UCMerced Pretrained	Million-AID Fine-Tuned	Change
UCMerced OA	A	98.96	97.44	−1.52
	B	99.48	97.92	−1.56
	C	100.00	97.92	−2.08
	D	99.48	97.92	−1.56
Potsdam mIoU	A	33.39	37.50	+4.11
	B	33.60	35.96	+2.36
	C	35.82	36.42	+0.60
	D	34.25	39.44	+5.19
DOTA mAP	A	43.62	40.04	−3.58
	B	44.15	40.22	−3.93
	C	45.24	41.33	−3.91
	D	44.78	40.84	−3.94

As shown in Table 9, the influence of the Million-AID-finetuned backbone varies across downstream tasks. For UCMerced classification, the Million-AID-finetuned backbone leads to lower accuracy for all configurations, suggesting that supervised fine-tuning on the Million-AID subset may shift the representation toward its own category space and reduce its adaptability to UCMerced.

For the Potsdam semantic segmentation task, the Million-AID-finetuned backbone generally improves the performance. Config D increases from 34.25% to 39.44%, with a gain of 5.19 percentage points, while the Baseline improves from 33.39% to 37.50%, with a gain of 4.11 percentage points. This indicates that supervised fine-tuning on the sampled Million-AID subset can enhance high-level semantic representation to some extent and benefit segmentation transfer.

However, for DOTA oriented object detection, the Million-AID-finetuned backbone decreases the mAP of all configurations by approximately 3.58 to 3.94 percentage points. This suggests that supervised classification fine-tuning may strengthen global category-discriminative features but weaken the sensitivity to object localization, orientation, and local structures, which are crucial for oriented detection.

4.10. Feature Clustering Visualization

In addition to quantitative experiments, feature clustering visualization is further used to qualitatively analyze the representation distributions learned by different configurations. Specifically, image features extracted by the pretrained backbone are projected into a two-dimensional space using a dimensionality reduction method, and samples from different categories are visualized with different colors.

Figure 5 shows the feature clustering comparison under different configurations. The first row denotes the feature distributions obtained by direct fine-tuning on UCMerced, while the second row denotes the results obtained after transferring the Million-AID-finetuned backbone to UCMerced. In the direct UCMerced fine-tuning setting, Config C produces relatively more compact and separable clusters, which is consistent with its best classification accuracy of 100.00%. This observation is consistent with improved feature separability under the evaluated classification setting.

In the two-stage transfer setting, the feature distributions become more scattered for some categories. Although the Million-AID-finetuned backbone can introduce fine-grained category information, it may also shift the feature space toward the Million-AID subset distribution. This observation is consistent with the quantitative results in Table 9, where transferring the Million-AID-finetuned backbone back to UCMerced does not further improve classification performance. Overall, the clustering visualization is consistent with improved feature separability for the CCL configuration under the evaluated classification setting, while supervised fine-tuning on the Million-AID subset may introduce a task-specific representation shift.

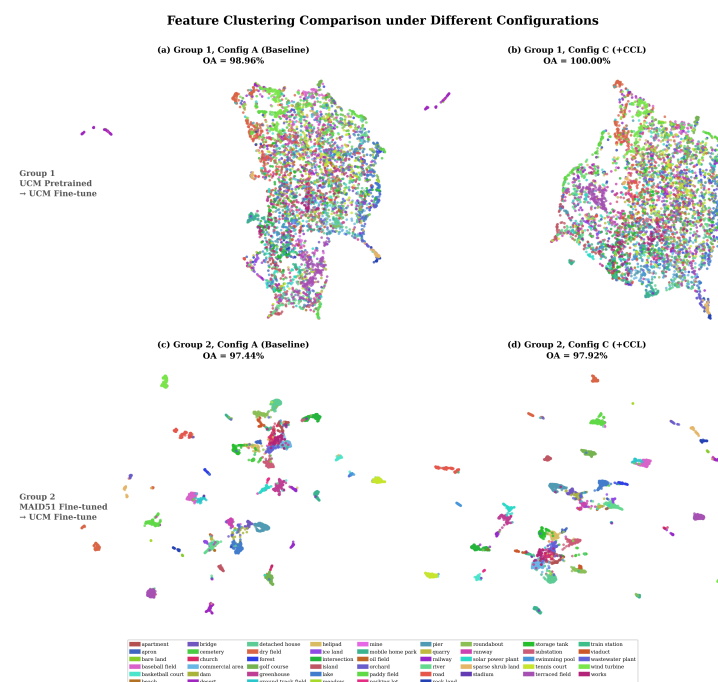


Figure 5. Feature clustering visualization under different configurations. Different colors denote different categories.

5. Conclusions

This paper investigates collaborative self-supervised representation learning for remote sensing imagery under conditions in which high-ratio masking may leave insufficient visible evidence for reconstructing small or structurally informative regions. To examine contextual reconstruction and heterogeneous objective coordination within a unified framework, we propose C²DSSL, a context-consistency enhanced collaborative self-supervised learning framework. The framework combines reconstruction-level contextual supervision with global and local representation-learning objectives and evaluates gradient-sensitive dynamic weighting as an auxiliary strategy for adjusting their relative contributions.

The context-consistency constraint uses the complete-view reconstruction features generated by the teacher network to guide the masked student reconstruction. Under the current experimental settings, Config C outperforms the Baseline by 4.85 percentage points in UCMerced KNN evaluation, 1.04 percentage points in UCMerced classification, 2.43 percentage points in Potsdam semantic segmentation, and 1.62 percentage points in DOTA oriented object detection. These results suggest that reconstruction-level contextual supervision can benefit representation transfer across the evaluated downstream tasks. However, the results should not be interpreted as establishing universal improvements across all datasets, architectures, or representation properties.

The external transfer comparison further shows that C²DSSL achieves or ties the best result on UCMerced classification, Million-AID classification, and Potsdam semantic segmentation under the adopted downstream evaluation pipeline, while remaining competitive on KNN evaluation and DOTA oriented object detection. Because the external encoders use different pretraining corpora and training schedules, these results are interpreted as transfer references rather than strictly controlled pretraining comparisons. In addition, the masking-ratio analysis shows that 0.60 provides the best performance among the tested ratios of 0.40, 0.60, and 0.80 under the current limited-data setting.

Regarding objective coordination, the gradient-sensitive dynamic weighting mechanism adjusts the relative contributions of different self-supervised objectives according to their smoothed gradient magnitudes. Its effect is task-dependent: a performance gain is observed on the Million-AID subset classification task, whereas the DW+CCL configuration does not consistently outperform the CCL-only configuration on the other evaluated tasks. Therefore, DW is treated as an auxiliary empirical strategy rather than a universally beneficial optimization mechanism. The two-stage transfer experiments further indicate that supervised fine-tuning on the Million-AID subset can improve Potsdam semantic segmentation to some extent but may reduce the performance on DOTA oriented object detection. This suggests that scene-level semantic fine-tuning and localization-sensitive dense prediction tasks may require different representation preferences.

Feature clustering visualization further supports the quantitative results. The learned representations with context consistency show better feature separability in the UCMerced classification setting, while the Million-AID-finetuned backbone introduces a certain task-specific representation shift when transferred back to UCMerced. These observations suggest that spatial contextual information learned during self-supervised pretraining may contribute to representation transfer across the evaluated tasks.

Limitations

Several limitations should be considered when interpreting the current results. First, the empirical evaluation uses UCMerced for self-supervised pretraining, a ResNet-50 backbone, and primarily RGB aerial imagery. Therefore, the transferability of C²DSSL to larger-scale pretraining datasets, transformer-based architectures, multispectral or SAR imagery, and multi-temporal observations remains to be examined. Second, the context

consistency and local semantic alignment branches rely on spatial correspondence between the complete teacher view and the masked student view. Their application to weakly aligned cross-view, cross-temporal, or crossmodal observations may require additional correspondence-estimation or feature-alignment mechanisms. Third, gradient-sensitive dynamic weighting uses temporally smoothed gradient magnitude as an empirical proxy rather than a direct measure of optimization difficulty. Its effect is task-dependent and may be influenced by loss scale, model parameterization, and stochastic gradient noise. Fourth, the current experiments are based on a single fixed random seed and do not quantify run-to-run variance. Therefore, relatively small performance differences should be interpreted cautiously. In addition, the current qualitative evaluation focuses mainly on feature-clustering visualization for classification and does not include matched prediction visualizations for segmentation and detection; therefore, fine-grained pixel- and instance-level error patterns remain to be examined. Finally, the teacher–student architecture, multiple input views, feature pyramid modules, reconstruction decoders, and projection heads introduce additional computational and memory overhead compared with a single-branch self-supervised framework. Future work will investigate broader data settings, stronger backbones, repeated evaluations under multiple random seeds, more robust objective-coordination strategies, and lighter implementations.

Author Contributions: Conceptualization, W.W., J.L. and Z.X.; Methodology, W.W., J.L. and K.Q.; Software, J.L. and K.Q.; Validation, J.L., K.Q. and G.L.; Formal analysis, J.L. and W.W.; Investigation, J.L., K.Q. and G.L.; Resources, W.W. and Z.X.; Data curation, J.L. and K.Q.; Writing—original draft preparation, J.L. and W.W.; Writing—review and editing, W.W., Z.X. and G.L.; Visualization, J.L. and K.Q.; Supervision, W.W. and Z.X.; Project administration, W.W. and Z.X.; Funding acquisition, W.W. and Z.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 52477138; the Guangdong Basic and Applied Basic Research Foundation, Major Basic Research and Cultivation Project, grant number 2024A1515011976; and the Guangzhou Higher Education Teaching Quality and Teaching Reform Project, grant number 2023KCJJD002.

Data Availability Statement: The original contributions presented in this study are included in the article. The datasets used in this study are publicly available, including UCMerced, Million-AID, ISPRS Potsdam, and DOTA v1.0. Further inquiries can be directed to the corresponding author.

Acknowledgments: The authors would like to thank Guangzhou University for providing the research environment and computing support for this study.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Van Westen, C. Remote sensing for natural disaster management. *Int. Arch. Photogramm. Remote Sens.* **2000**, *33*, 1609–1617.
2. Bauer, M.E. Remote sensing of environment: History, philosophy, approach and contributions, 1969–2019. *Remote Sens. Environ.* **2020**, *237*, 111522. [[CrossRef](#)]
3. Xiao, Y.; Zhan, Q. A review of remote sensing applications in urban planning and management in China. In Proceedings of the 2009 Joint Urban Remote Sensing Event, Shanghai, China, 20–22 May 2009; pp. 1–5.
4. You, J.; Li, X.; Low, M.; Lobell, D.; Ermon, S. Deep gaussian process for crop yield prediction based on remote sensing data. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 31.
5. Gupta, D.; Biswas, A.; Ghosh, P.; Rawat, U.; Tripathi, S. Delineation of groundwater potential zones, groundwater estimation and recharge potentials from Mahoba district of Uttar Pradesh, India. *Int. J. Environ. Sci. Technol.* **2022**, *19*, 12145–12168.
6. He, N.; Fang, L.; Li, S.; Plaza, A.; Plaza, J. Remote sensing scene classification using multilayer stacked covariance pooling. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 6899–6910. [[CrossRef](#)]

7. Zhang, X.; Zhang, T.; Wang, G.; Zhu, P.; Tang, X.; Jia, X.; Jiao, L. Remote sensing object detection meets deep learning: A metareview of challenges and advances. *IEEE Geosci. Remote Sens. Mag.* **2023**, *11*, 8–44. [\[CrossRef\]](#)
8. Ma, X.; Wu, Q.; Zhao, X.; Zhang, X.; Pun, M.O.; Huang, B. SAM-assisted remote sensing imagery semantic segmentation with object and boundary constraints. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–16. [\[CrossRef\]](#)
9. Liu, C.; Chen, K.; Qi, Z.; Liu, Z.; Zhang, H.; Zou, Z.; Shi, Z. Pixel-level change detection pseudo-label learning for remote sensing change captioning. In Proceedings of the IGARSS 2024—2024 IEEE International Geoscience and Remote Sensing Symposium, Athens, Greece, 7–12 July 2024; pp. 8405–8408.
10. Alosaimi, N.; Alhichri, H.; Bazi, Y.; Ben Youssef, B.; Alajlan, N. Self-supervised learning for remote sensing scene classification under the few shot scenario. *Sci. Rep.* **2023**, *13*, 433. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Jing, L.; Tian, Y. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 4037–4058. [\[CrossRef\]](#)
12. Wang, Y.; Albrecht, C.M.; Braham, N.A.A.; Mou, L.; Zhu, X.X. Self-supervised learning in remote sensing: A review. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 213–247. [\[CrossRef\]](#)
13. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the 37th International Conference on Machine Learning, Virtual, 13–18 July 2020; pp. 1597–1607.
14. 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, Seattle, WA, USA, 13–19 June 2020; pp. 9729–9738.
15. Chen, X.; Fan, H.; Girshick, R.; He, K. Improved baselines with momentum contrastive learning. *arXiv* **2020**, arXiv:2003.04297.
16. Grill, J.B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. Bootstrap your own latent—a new approach to self-supervised learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21271–21284.
17. 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, Montreal, QC, Canada, 11–17 October 2021; pp. 9650–9660.
18. Manas, O.; Lacoste, A.; Giró-i Nieto, X.; Vazquez, D.; Rodriguez, P. Seasonal contrast: Unsupervised pre-training from uncured remote sensing data. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 9414–9423.
19. 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, Montreal, QC, Canada, 11–17 October 2021; pp. 10181–10190.
20. Kang, J.; Fernandez-Beltran, R.; Duan, P.; Liu, S.; Plaza, A.J. Deep unsupervised embedding for remotely sensed images based on spatially augmented momentum contrast. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 2598–2610. [\[CrossRef\]](#)
21. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16000–16009.
22. Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; Hu, H. Simmim: A simple framework for masked image modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 9653–9663.
23. Bao, H.; Dong, L.; Piao, S.; Wei, F. Beit: Bert pre-training of image transformers. *arXiv* **2021**, arXiv:2106.08254.
24. 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. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 197–211. [\[CrossRef\]](#)
25. 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, Paris, France, 1–6 October 2023; pp. 4088–4099.
26. Tang, M.; Cozma, A.; Georgiou, K.; Qi, H. Cross-scale mae: A tale of multiscale exploitation in remote sensing. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 20054–20066. [\[CrossRef\]](#)
27. Zhang, M.; Liu, Q.; Wang, Y. CtxMIM: Context-enhanced masked image modeling for remote sensing image understanding. *ACM Trans. Multimed. Comput. Commun. Appl.* **2025**, *21*, 1–22. [\[CrossRef\]](#)
28. Jha, A.; Kumar, A.; Banerjee, B.; Chaudhuri, S. Adamt-net: An adaptive weight learning based multi-task learning model for scene understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Virtual, 14–19 June 2020; pp. 706–707.
29. Chen, Z.; Badrinarayanan, V.; Lee, C.Y.; Rabinovich, A. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 794–803.
30. Kendall, A.; Gal, Y.; Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7482–7491.

31. Li, W.; Chen, H.; Shi, Z. Semantic segmentation of remote sensing images with self-supervised multitask representation learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 6438–6450. [[CrossRef](#)]
32. Li, H.; Li, Y.; Zhang, G.; Liu, R.; Huang, H.; Zhu, Q.; Tao, C. Global and local contrastive self-supervised learning for semantic segmentation of HR remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5618014. [[CrossRef](#)]
33. Scheibenreif, L.; Hanna, J.; Mommert, M.; Borth, D. Self-supervised vision transformers for land-cover segmentation and classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1422–1431.
34. Mall, U.; Hariharan, B.; Bala, K. Change-aware sampling and contrastive learning for satellite images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 5261–5270.
35. Li, W.; Chen, K.; Shi, Z. Geographical supervision correction for remote sensing representation learning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–20. [[CrossRef](#)]
36. Wang, Y.; Albrecht, C.M.; Zhu, X.X. Self-supervised vision transformers for joint SAR-optical representation learning. In Proceedings of the IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 139–142.
37. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. DINOv2: Learning robust visual features without supervision. *arXiv* **2023**, arXiv:2304.07193.
38. Sun, X.; Wang, P.; Lu, W.; Zhu, Z.; Lu, X.; He, Q.; Li, J.; Rong, X.; Yang, Z.; Chang, H.; et al. RingMo: A remote sensing foundation model with masked image modeling. *IEEE Trans. Geosci. Remote Sens.* **2022**, *61*, 1–22. [[CrossRef](#)]
39. Hong, D.; Zhang, B.; Li, X.; Li, Y.; Li, C.; Yao, J.; Yokoya, N.; Li, H.; Ghamisi, P.; Jia, X.; et al. SpectralGPT: Spectral remote sensing foundation model. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 5227–5244. [[CrossRef](#)] [[PubMed](#)]
40. Mendieta, M.; Han, B.; Shi, X.; Zhu, Y.; Chen, C. Towards geospatial foundation models via continual pretraining. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 16806–16816.
41. Jha, A.; Singha, M.; Bhattacharya, A.; Banerjee, B. RS3Lip: Consistency for remote sensing image classification on part embeddings using self-supervised learning and CLIP. *Comput. Vis. Image Underst.* **2025**, *251*, 104254. [[CrossRef](#)]
42. Chen, X.; He, K. Exploring simple siamese representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and PATTERN recognition, Virtual, 19–25 June 2021; pp. 15750–15758.
43. Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 12310–12320.
44. Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9912–9924.
45. 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, Nashville, TN, USA, 20–25 June 2021; pp. 1182–1191.
46. Jung, H.; Oh, Y.; Jeong, S.; Lee, C.; Jeon, T. Contrastive self-supervised learning with smoothed representation for remote sensing. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [[CrossRef](#)]
47. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; pp. 4171–4186.
48. Noman, M.; Naseer, M.; Cholakkal, H.; Anwer, R.M.; Khan, S.; Khan, F.S. Rethinking transformers pre-training for multi-spectral satellite imagery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 27811–27819.
49. Muhtar, D.; Zhang, X.; Xiao, P.; Li, Z.; Gu, F. CMID: A unified self-supervised learning framework for remote sensing image understanding. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–17. [[CrossRef](#)]
50. Wen, W.; Luo, J.; Tan, L.; Kostromitin, K.I.; Wang, J.; Zhang, P. S 2 EDL: Selective Semantic Efficient Distillation Learning for Large-Scale Remote Sensing Representation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *19*, 3790–3805. [[CrossRef](#)]
51. Caruana, R. Multitask learning. *Mach. Learn.* **1997**, *28*, 41–75. [[CrossRef](#)]
52. Zhang, Y.; Yang, Q. A survey on multi-task learning. *IEEE Trans. Knowl. Data Eng.* **2021**, *34*, 5586–5609. [[CrossRef](#)]
53. Liu, S.; Johns, E.; Davison, A.J. End-to-end multi-task learning with attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1871–1880.
54. 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, San Jose, CA, USA, 2–5 November 2010; pp. 270–279.

55. Long, Y.; Xia, G.S.; Li, S.; Yang, W.; Yang, M.Y.; Zhu, X.X.; Zhang, L.; Li, D. On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 4205–4230. [[CrossRef](#)]
56. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3974–3983.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.