Self-supervised Learning in Remote Sensing: A Survey

www.surveyx.cn

Abstract

Self-supervised learning (SSL) has emerged as a transformative approach in remote sensing, leveraging vast unlabeled datasets to autonomously extract meaningful features, thereby reducing the dependency on costly labeled data. This survey explores the integration of SSL with multi-modal data, emphasizing its role in enhancing scene classification, object detection, change detection, and semantic segmentation. The incorporation of frameworks like TransFuse illustrates SSL's potential in image fusion, achieving state-of-the-art performance in diverse applications. The survey underscores the importance of label-efficient learning in agricultural productivity and highlights the critical role of architectural choices in SSL performance. Despite its promise, challenges remain in data quality, computational demands, and multi-modal integration, necessitating further research into more efficient SSL frameworks. Future directions include optimizing hyperparameter tuning, extending frameworks to other SSL methods, and exploring benchmarks across diverse geographical regions. By addressing these challenges, SSL can significantly advance environmental monitoring and geospatial analysis, offering innovative solutions to longstanding data analysis challenges in remote sensing.

1 Introduction

1.1 Significance of Self-supervised Learning in Remote Sensing

Self-supervised learning (SSL) has emerged as a pivotal methodology in remote sensing, enabling the autonomous extraction of meaningful features from extensive unlabeled datasets, thereby alleviating the need for costly manual annotations [1, 2]. This capability is particularly beneficial in remote sensing, where obtaining labeled data is resource-intensive. SSL empowers models to develop robust visual representations essential for analyzing the complex, high-dimensional data characteristic of remote sensing applications [3].

The integration of SSL enhances model generalization and robustness, particularly in scenarios with limited labeled data. Its ability to utilize multi-view data, as seen in multiview object recognition, is crucial for improving representation learning in geospatial contexts [4]. Moreover, SSL techniques have proven effective in enhancing image quality and resolution, thus playing a vital role in remote sensing applications [5].

Despite the computational demands often associated with SSL methods, their scalability addresses dataset imbalances in large-scale unlabeled datasets, facilitating the extraction of general visual representations without manual labels [1]. Furthermore, innovative data augmentation strategies within SSL frameworks have been shown to improve performance in anomaly detection tasks, enhancing detection capabilities in remote sensing [6].

1.2 Motivation for Integrating Multi-modal Data

The drive to integrate multi-modal data in remote sensing stems from the need to leverage diverse data sources for enhanced interpretation and analysis, especially in contexts where labeled data is scarce and costly. This strategy effectively utilizes unlabeled data, reducing reliance on extensive

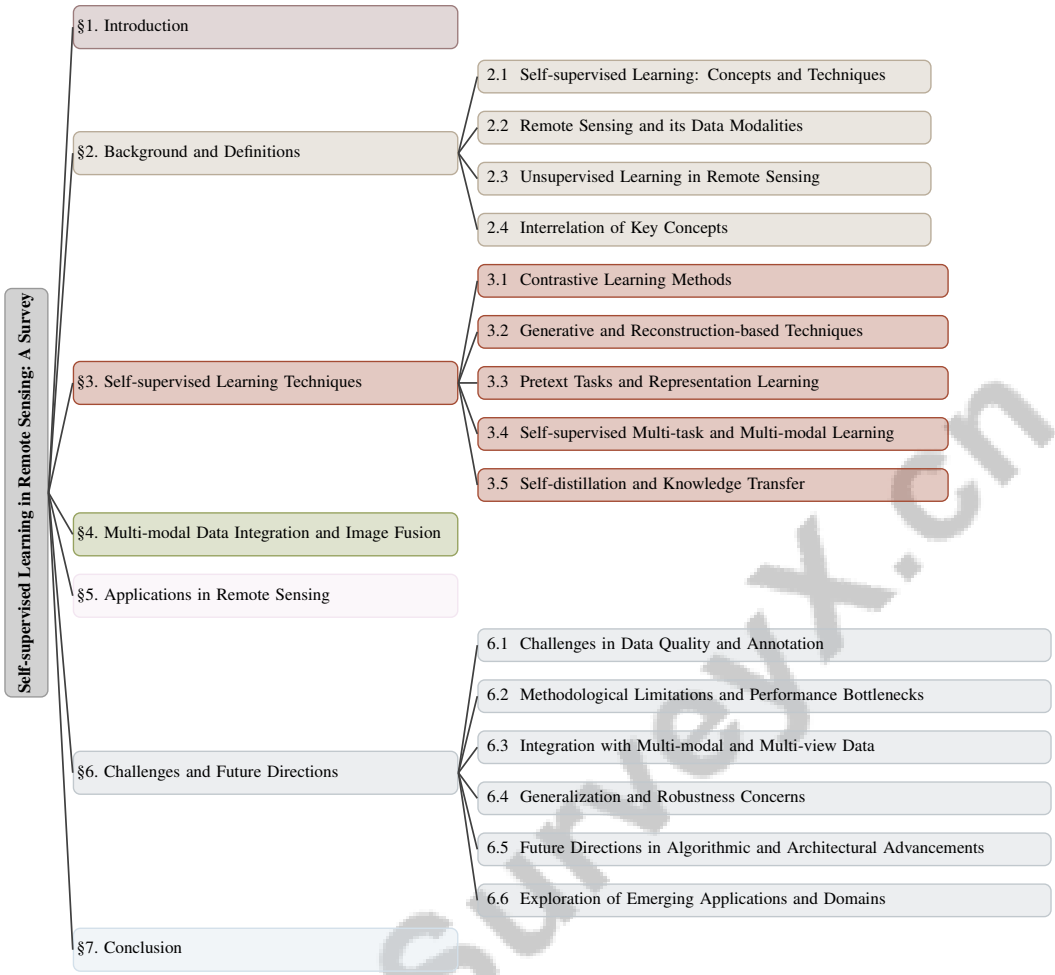


Figure 1: chapter structure

annotations and improving model performance [7]. By combining modalities like multispectral (MS) and synthetic aperture radar (SAR) data, implicit augmentation occurs, facilitating the learning of invariant features crucial for robust model performance [8].

Multi-modal data integration overcomes the limitations of traditional data augmentation methods, which may alter pixel values without preserving semantic integrity—essential in remote sensing tasks [9]. Additionally, combining modalities such as RGB images and predicted depth maps enhances the accuracy of ego-motion estimation under challenging conditions [10].

SSL significantly contributes to this integration by extracting features from unlabeled datasets, addressing the prevalent challenge of data scarcity in remote sensing [11]. Innovative methods leveraging prior knowledge and semantic cues further underscore the potential of multi-modal data integration in improving calibration accuracy [12]. The adaptability and efficiency of AI models are enhanced through this integration, showcasing broader applicability beyond remote sensing [13].

The proposed View Invariant Stochastic Prototype Embedding (VISPE) method exemplifies the advantages of multi-modal data by focusing on robust embeddings from limited unlabelled multiview data [4]. A unified framework for SSL methods, which addresses the lack of a principled theoretical understanding, further motivates the integration of various SSL techniques, enhancing downstream task performance by retaining relevant features affected by augmentations [14]. The potential to replace traditional supervised methods with semi-supervised approaches under limited labeled data conditions highlights the benefits of multi-modal integration in improving outcomes across diverse applications [9].

1.3 Benefits of Image Fusion and Data Integration

Image fusion and data integration are crucial for enhancing the analytical capabilities of remote sensing technologies by synthesizing complementary information from diverse sources. This integration enriches datasets, providing a comprehensive view of observed scenes and significantly improving analysis and interpretation accuracy [15]. The RS-BYOL framework exemplifies the advantages of leveraging multi-modal data, offering richer feature representations compared to single-modality approaches [16].

SSL plays a pivotal role in facilitating image fusion by reducing reliance on labeled data and enabling the use of extensive unlabeled datasets, especially beneficial in remote sensing where labeled data is often limited and costly. This approach enhances data efficiency and improves model performance under data-scarce conditions. The use of generative models within SSL frameworks further enhances representation learning by generating semantically consistent augmentations, increasing the diversity of training data and robustness of learned representations [17].

The integration of spatial and spectral features through SSL methods, such as the self-supervised spectral-spatial attention-based vision transformer (SSVT), has led to substantial improvements in data analysis, resulting in higher accuracy in applications like nitrogen status estimation [18]. Additionally, the Fusion Transformer method illustrates the benefits of combining features from different sensors to enhance activity recognition, showcasing the broader applicability of image fusion techniques [19].

Moreover, frameworks that allow for closed-form optimal representations and network parameters highlight the benefits of integrating different SSL approaches, enhancing the overall effectiveness of data fusion techniques [20]. The CELESTIAL pipeline further exemplifies the potential of SSL in remote sensing, achieving comparable accuracy to traditional methods while reducing dependence on labeled data [21].

1.4 Structure of the Survey

This survey is meticulously organized to provide a comprehensive understanding of SSL in remote sensing, exploring its significance, methodologies, and applications. The introduction outlines the motivation for integrating SSL with remote sensing data, emphasizing the benefits of utilizing multi-modal data, image fusion, and data integration. It highlights SSL's transformative effects on data representation and analysis and its advantages in leveraging unlabeled datasets for improved model performance and efficiency [22, 23, 24].

The second section offers background and definitions, elucidating key concepts related to SSL, remote sensing, multi-modal data, image fusion, and unsupervised learning, setting the stage for understanding their interrelation and relevance to remote sensing.

In the third section, various SSL techniques applicable to remote sensing are explored, including contrastive learning, generative learning, and pretext tasks, along with insights into their practical implementation.

The fourth section focuses on multi-modal data integration and image fusion, examining techniques and strategies for combining data from different sources and modalities while addressing associated challenges and solutions.

Applications of SSL in remote sensing are reviewed in the fifth section, covering tasks such as scene classification, object detection, change detection, and anomaly detection, demonstrating how SSL methods enhance performance in real-world scenarios.

The penultimate section identifies current challenges in applying SSL to remote sensing data and discusses potential future research directions, including data quality and annotation challenges, methodological limitations, and the integration of multi-modal and multi-view data.

The conclusion highlights the pivotal findings of the survey, emphasizing SSL's transformative role in remote sensing. It suggests that SSL can significantly enhance data analysis methodologies, paving the way for more efficient use of large-scale unlabeled datasets and fostering innovative approaches to earth observation. This underscores the necessity for further exploration and development of SSL techniques tailored specifically for remote sensing applications [23, 22]. The main contributions of

the paper are highlighted, and areas for future research are suggested, paving the way for continued advancements in this domain.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Self-supervised Learning: Concepts and Techniques

Self-supervised learning (SSL) is a pivotal paradigm in target-agnostic learning, improving predictive accuracy by utilizing unlabeled data, which is especially beneficial in remote sensing contexts where labeled data is scarce and costly [25, 3]. SSL methodologies are categorized into generative, predictive, and contrastive approaches, each employing unique strategies to derive meaningful data representations [5]. Generative techniques enhance representation robustness by creating new data or features, aiding model generalization from limited labeled examples, crucial in data augmentation without semantic distortion [6]. Predictive methods design pretext tasks to capture intrinsic data structures [5]. Contrastive learning distinguishes between similar and dissimilar data instances, demonstrating superior performance over supervised pretraining in various tasks [1]. SSL's adaptability to dynamic environments and sensor conditions is vital in remote sensing for integrating multi-modal data and extracting meaningful representations [3]. Challenges such as representation collapse and transformation selection for pretext tasks can hinder representation learning, which SSL frameworks address through diverse augmentation strategies, enhancing robustness and effectiveness [6]. SSL represents a significant advancement in machine learning, offering scalable solutions to the challenges posed by large, complex datasets [5].

2.2 Remote Sensing and its Data Modalities

Remote sensing is crucial for environmental monitoring, agriculture, and urban planning, enabling data acquisition without direct contact. It employs diverse sensors to capture information about Earth's surface, atmosphere, and oceans, providing insights for numerous applications, including climate change, resource management, and disaster response [18]. Data modalities in remote sensing include multispectral, hyperspectral, radar, and LiDAR data. Multispectral and hyperspectral sensors capture a range of electromagnetic spectrum wavelengths, allowing detailed analysis of surface materials and vegetation health, essential for tasks like vegetation monitoring and water quality assessment [26]. Technologies like spectral data analysis, radar systems, and LiDAR enhance the ability to interpret complex environmental information, offering unique data types and resolutions for accurate mapping and monitoring [16, 27, 22, 28, 29]. Radar sensors, such as synthetic aperture radar (SAR), provide valuable surface structure information, performing well in all-weather conditions, while LiDAR offers precise 3D surface elevation data, indispensable for topographic mapping and forestry applications. The integration of multiple modalities enhances remote sensing capabilities; for instance, the Fusion Transformer model combines features from passive Wi-Fi radar and channel state information, demonstrating multimodal learning's potential in data interpretation [19]. MMEarth, a dataset encompassing 1.2 million locations with data from 12 modalities, illustrates the breadth of environmental contexts captured through remote sensing [30]. Despite advancements, challenges persist in adapting image fusion methods to various tasks due to task-specific training data reliance. Unified transformer-based image fusion techniques, like TransFuse, enhance adaptability and performance across applications [15]. Traditional pre-training methods, reliant on RGB images, often fail to capture remote sensing data's spectral and spatial nuances, underscoring the need for sophisticated approaches to leverage diverse modalities [16]. Remote sensing is a fundamental pillar of contemporary environmental and geospatial analysis, utilizing diverse data modalities, including satellite imagery and UAV data, to provide insights into Earth's dynamic systems. Recent SSL advancements, such as Multi-Pretext Masked Autoencoder (MP-MAE) and RemoteCLIP, enhance meaningful representation extraction from vast unlabeled Earth observation data, improving performance in applications like image classification, semantic segmentation, and zero-shot learning, contributing to efficient global environmental monitoring and resource management [30, 31, 32, 33, 34]. The ongoing integration of advanced machine learning techniques and multimodal data promises further enhancements in remote sensing precision and applicability.

AI-generated, for reference only.

2.3 Unsupervised Learning in Remote Sensing

Unsupervised learning is pivotal in remote sensing, offering a framework for analyzing extensive datasets without labeled data, which is often limited and costly [2]. This paradigm is vital for tasks like clustering, segmentation, and anomaly detection, where supervised learning is constrained by manual annotation costs and logistics [35]. Unsupervised learning's capacity to identify patterns within data is advantageous in scenarios where manual labeling is impractical. A significant challenge is effectively clustering data samples without extensive labeled datasets, as traditional methods often rely on handcrafted features that may not scale adequately or capture intricate structures in remote sensing data [35]. Recent SSL developments have shown potential in overcoming these limitations by utilizing large volumes of unlabeled data to learn robust feature representations, enhancing clustering performance. In agricultural remote sensing, high costs and time for collecting, processing, and labeling large datasets present a challenge [2]. Unsupervised learning techniques, which do not rely on labeled data, offer a solution for extracting insights from agricultural data, aiding in crop monitoring and yield prediction. Unsupervised learning also facilitates multi-modal data integration in remote sensing, aligning and understanding relationships between data sources, such as images and LiDAR point clouds, vital in scenarios with limited labeled data. SSL methods leverage large amounts of unlabeled data to bridge gaps between sensor modalities, enhancing data interpretation and analysis, offering a promising alternative to supervised learning approaches that depend on extensive annotated datasets, which are challenging and costly [22, 33]. Developing effective pretext tasks in unsupervised learning is crucial for meaningful learning from large datasets, exploiting inherent data structures to learn useful representations enhancing performance across remote sensing applications. Despite annotation challenges and high-dimensional datasets, unsupervised learning, particularly SSL, has emerged as a vital element in remote sensing analysis, providing scalable and efficient methodologies for extracting insights from vast unlabeled data, reducing reliance on labor-intensive labeled datasets. Recent SSL advancements demonstrate potential to outperform traditional supervised approaches in tasks like scene classification, semantic segmentation, and object detection, leveraging remote sensing images' unique characteristics, positioning unsupervised learning as a foundational technique and promising avenue for future research and application in earth observation [36, 33, 22, 34].

2.4 Interrelation of Key Concepts

The interrelation of SSL, remote sensing, and data integration is foundational in advancing machine learning capabilities, particularly in data-scarce environments. SSL's ability to leverage unlabeled data is crucial in remote sensing, where acquiring labeled datasets poses significant logistical and financial challenges [37]. This is exemplified in tasks like segmenting building and road features from Digital Elevation Models (DEMs), where SSL techniques enhance feature extraction without extensive labeled datasets [38]. Data integration in remote sensing synthesizes information from various modalities into a cohesive analytical framework, improving learned representations' effectiveness in downstream tasks [12]. The integration of SSL techniques with multi-modal data underscores this interrelation, enhancing complex datasets' analysis typical of remote sensing applications [39]. However, different downstream tasks often require contradictory invariances, posing challenges in achieving high performance across all tasks using a single augmentation strategy [38]. The integration of SSL with multi-modal learning approaches remains an area with unanswered questions, particularly regarding optimal pretext task design [25]. The theoretical perspective based on identifying invariant content partitions through data augmentations illustrates the complexity of integrating SSL with remote sensing [3]. The interplay between SSL and remote sensing is also evident in the ability to classify sounds and images without extensive labeled datasets, highlighting SSL's versatility in enhancing data analysis [12]. From a theoretical standpoint, the framework based on information theory, focusing on mutual information and conditional entropy, provides insights into relationships between inputs, self-supervised signals, and task-relevant information [40]. This theoretical underpinning is crucial for understanding how SSL can be effectively integrated with remote sensing data to improve representation learning and generalization [37]. Moreover, the interrelation of these key concepts extends beyond traditional applications, offering significant advancements in fields like automated systems and robotic control [38]. By leveraging SSL, researchers can develop models capable of overcoming challenges posed by large, unlabeled datasets, paving the way for significant advancements in remote sensing and beyond. The combination of self-supervised tasks with few-shot learning serves as a data-dependent regularizer, enhancing the model's ability to learn richer representations [39].

3 Self-supervised Learning Techniques

Category	Feature	Method
Contrastive Learning Methods	Scale and Texture Handling	SAR[41], DZSR[42]
	Mask and Proxy Techniques	Odin[9]
	Generative Estimation	ATP[6]
Generative and Reconstruction-based Techniques	Reconstruction Techniques	LoMaR[43]
	Multi-Modal Integration	TF[15]
Pretext Tasks and Representation Learning	Contrastive-Based Learning	CSSL[37]
	Data Augmentation Techniques	ASSL[7]
	3D Spatial Tasks	RCR[5]
Self-supervised Multi-task and Multi-modal Learning	Modality-Focused Techniques	Self-MM[44], FT[19], MTSTN[45], MLF-VO[46]
	Consistency and Stability	VISPE[4]
	Scalability and Generalization	RE[47]
Self-distillation and Knowledge Transfer	Domain-Independent Techniques	SD-DAC[48]
	Non-Labeled Data Approaches	NSSL-WLD[49], SL+MLP[50], SCPNet[51]
	Discriminative Representation Learning	V2Net[3]

Table 1: This table provides a comprehensive summary of various self-supervised learning (SSL) techniques categorized into five distinct groups: contrastive learning methods, generative and reconstruction-based techniques, pretext tasks and representation learning, self-supervised multi-task and multi-modal learning, and self-distillation and knowledge transfer. Each category is further detailed with specific features and methods, highlighting the diverse approaches and applications within the SSL framework for enhancing feature extraction and representation learning from unlabeled data.

In the realm of self-supervised learning (SSL), various techniques have been developed to enhance the extraction of meaningful representations from unlabeled data. Among these, contrastive learning methods stand out for their ability to leverage the relationships between data samples to improve feature learning. This subsection delves into the intricacies of contrastive learning techniques, highlighting their applications, adaptations, and effectiveness within the context of remote sensing. Table 1 presents a detailed classification and overview of self-supervised learning techniques, illustrating their diverse methodologies and applications in enhancing data representation and feature extraction. Additionally, Table 3 offers a comprehensive comparison of various self-supervised learning techniques, illustrating their methodologies and applications within the context of remote sensing and data representation. To provide a clearer understanding of the landscape of SSL techniques, **??** illustrates the hierarchical structure of these methods, with a particular focus on contrastive learning approaches. The figure categorizes SSL techniques into three main groups: contrastive learning methods, generative and reconstruction-based techniques, and pretext tasks. Each category is further subdivided to showcase specific applications, integrations, and advanced techniques, thereby underscoring the versatility and effectiveness of these approaches in enhancing feature extraction and representation learning from unlabeled data.

3.1 Contrastive Learning Methods

Contrastive learning (CL) has emerged as a pivotal approach in self-supervised learning (SSL), especially for applications in remote sensing where labeled data is often scarce and costly to obtain. By leveraging the contrast between positive and negative sample pairs, CL facilitates the extraction of meaningful representations from complex datasets, enhancing the quality of feature learning [35]. This method is particularly beneficial in remote sensing, where data often suffers from noise and variability across different modalities [11].

A notable adaptation of CL in remote sensing is the use of multi-scale inputs and adversarial learning to improve modality-invariant representation learning, as demonstrated in SAR applications. This approach distinguishes itself from traditional SSL methods by effectively capturing the intricate patterns inherent in remote sensing data [41]. Additionally, the SelfDZSR++ framework exemplifies the application of CL in enhancing the resolution of ultra-wide images using telephoto images as references, showcasing the adaptability of CL techniques to real-world image scenarios [42].

Odin, a specific CL technique, utilizes generated masks to learn object representations, highlighting the versatility of CL in discovering and representing objects within remote sensing imagery [9]. This method underscores the potential of CL to advance object detection and classification tasks in geospatial analysis.

The integration of CL with other learning paradigms, such as the hybrid texture model combining fixed-filter decompositions with learned representations, further enhances texture classification in

remote sensing applications. This hybrid approach allows for better generalization from limited data, addressing a common challenge in remote sensing [3].

Moreover, the Rubik's Cube Recovery technique exemplifies a novel proxy task in CL, where 3D neural networks are pre-trained by solving a task that simulates the mechanics of a Rubik's cube. This innovative approach aids in learning robust 3D representations, which are crucial for analyzing three-dimensional remote sensing data [5].

Incorporating generative adversarial networks (GANs) within the CL framework, as proposed by Yang et al., facilitates the estimation of transformation distributions and the construction of complementary distributions for effective pretext tasks. This integration enhances the robustness and effectiveness of the learned representations, particularly in the context of remote sensing [6].

Contrastive learning methods offer a robust and versatile framework for enhancing representation learning in remote sensing applications, particularly through self-supervised approaches that effectively generate meaningful representations from hyperspectral images and other datasets without the need for semantic labels. These methods, such as those utilizing cross-domain CNN architectures, cluster similar spectral vectors within a unified representation space, facilitating improved classification and transfer learning across diverse hyperspectral images. Additionally, recent advancements demonstrate that contrastive learning outperforms traditional supervised methods in various downstream tasks, including semantic segmentation and object detection, by learning richer and more structured representations. This comprehensive framework not only addresses the challenges of annotation in remote sensing but also bridges theoretical gaps between contrastive and non-contrastive methods, providing valuable insights for practitioners aiming to optimize representation learning in this field. [35, 20, 52]. By addressing challenges related to data scarcity, noise, and multi-modal integration, CL continues to enhance the capabilities of remote sensing technologies across various environmental and geospatial applications.

3.2 Generative and Reconstruction-based Techniques

Generative and reconstruction-based techniques in self-supervised learning (SSL) have gained significant traction in the domain of remote sensing, providing robust frameworks for feature extraction from unlabeled data. These techniques primarily focus on generating new data instances or reconstructing existing ones to enhance the learning of meaningful representations. Generative models, such as those discussed by Jin et al., are particularly relevant in this context, offering adaptable solutions for remote sensing applications through the generation of synthetic data that mimics real-world conditions [25].

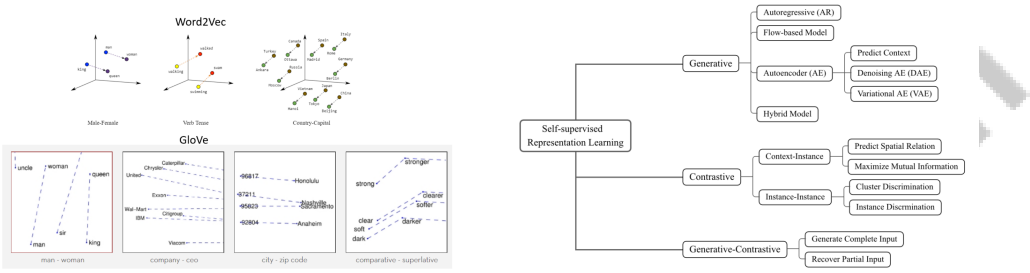
A notable approach in this category is the use of masked reconstruction strategies, exemplified by the LoMaR method, which performs masked reconstruction on small windows of image patches. This technique significantly enhances both efficiency and accuracy in feature extraction, making it well-suited for the high-dimensional data typical of remote sensing [43]. By focusing on reconstructing occluded or corrupted parts of an image, these methods enable models to learn spatial dependencies and contextual information, which are crucial for various remote sensing tasks.

Generative learning techniques also extend to video frame generation based on learned representations, as discussed by Schiappa et al. These methods leverage the temporal dynamics captured in remote sensing video data, thereby improving the model's ability to predict and analyze changes over time [53]. This capability is particularly beneficial for monitoring environmental changes and assessing the impact of natural disasters.

The integration of multi-view representation fusion, as highlighted by Li, further enhances the capabilities of generative models by combining features from different views into a cohesive representation. This approach utilizes probabilistic graphical models and neural networks to integrate diverse data sources, thereby enriching the feature space and improving model performance in remote sensing applications [54].

TransFuse, a unified image fusion framework, exemplifies the application of generative techniques in remote sensing. By employing a Transformer and CNN combined encoder-decoder architecture, TransFuse leverages self-supervised learning to improve feature extraction and fusion across different modalities [15]. This framework demonstrates the potential of combining generative models with advanced neural architectures to enhance the interpretation and analysis of remote sensing data.

Generative and reconstruction-based techniques are emerging as highly effective methodologies for enhancing self-supervised learning (SSL) in remote sensing, as they enable the utilization of vast amounts of unlabeled data to create rich, semantically consistent representations. Recent advancements in SSL, particularly through the application of generative models and innovative pretext tasks, have demonstrated significant improvements in performance on remote sensing tasks, showcasing the potential of these approaches to bridge the gap between the extensive availability of remote sensing imagery and the need for high-quality feature extraction without the burden of extensive manual annotation. [16, 17, 22, 32, 33]. By enabling the generation and reconstruction of high-quality data representations, these methods significantly contribute to the development of more accurate and robust remote sensing models.



(a) Word2Vec and GloVe: Visualizing Word Embeddings[55]

(b) Self-supervised Representation Learning[56]

Figure 2: Examples of Generative and Reconstruction-based Techniques

As shown in Figure 2, The example provided delves into the realm of self-supervised learning techniques, specifically focusing on generative and reconstruction-based methods. Illustrated through two distinct images, the first highlights the comparative visualization of word embeddings using Word2Vec and GloVe. These techniques are pivotal in natural language processing, as they represent words in high-dimensional spaces where the proximity between points correlates with semantic similarity. Word2Vec, depicted in a 3D embedding space, captures semantic relationships effectively, while GloVe offers an alternative approach. The second image presents a hierarchical tree diagram that categorizes various self-supervised representation learning methods. At the apex is "Self-supervised Representation Learning," which branches into three primary categories: Generative, Contrastive, and Generative-Contrastive. Each category is further subdivided, showcasing the diverse methodologies employed in self-supervised learning to enable machines to learn representations without explicit external labels. This structured overview underscores the significance and complexity of generative techniques in advancing machine learning capabilities. [**?**]thapa2022surveyselfsupervisedmultimodalrepresentation,liu2021self)

3.3 Pretext Tasks and Representation Learning

Pretext tasks are a cornerstone of self-supervised learning (SSL), particularly in remote sensing, where they enable the extraction of meaningful representations from unlabeled data. These tasks are designed to exploit inherent data structures, facilitating the learning of features that are transferable to various downstream applications. In remote sensing, where labeled datasets are often scarce and expensive to obtain, pretext tasks such as multi-task pretraining and self-supervised learning offer scalable solutions to improve model performance by leveraging large amounts of unlabeled data. Techniques like Multi-Task Pretraining (MTP) utilize a shared encoder and task-specific decoders to enhance various image interpretation tasks, while self-supervised methods, including those based on local-global view alignment, generate pseudolabels to train models effectively. These approaches have demonstrated significant improvements in performance across multiple remote sensing applications, even when trained on limited labeled data, thus addressing the challenges posed by the lack of annotated datasets. [32, 57]

A notable pretext task approach involves the segmentation of continuous activity signals into discrete samples, followed by data augmentation to generate positive pairs. This method is effective in learning robust representations for clustering tasks, as it minimizes contrastive loss across positive and negative pairs [37]. The Spectral Analysis Network (SANet) exemplifies this by conducting

AI-generated, for reference only.

spectral analysis on image patches to derive deep representations, demonstrating efficacy in clustering applications.

Generative models play a crucial role in the augmentation pipeline, enriching the SSL framework by producing diverse and semantically relevant image variations. This enhances the robustness and diversity of learned representations, which is vital for remote sensing tasks characterized by data heterogeneity [7]. In the context of 3D data, the Rubik's Cube Recovery method leverages a unique pretext task involving the rearrangement and rotation of cubes within a 3D volume. This task trains the network to learn invariant features, thereby improving the quality of 3D representations [5].

Object-aware cropping is another effective pretext task, generating two views from an image to ensure that the resulting representations capture richer semantic content. This technique is especially advantageous in remote sensing applications, where the integration of spatial context significantly enhances the accuracy of data interpretation, particularly when utilizing self-supervised learning methods that leverage diverse, unlabeled Earth observation datasets for improved feature representation. [30, 22, 32]. Additionally, depth inconsistency masks in depth estimation frameworks enhance robustness by filtering out dynamic regions during training, thereby improving depth representation quality.

Bootstrap Your Own Latent (BYOL) exemplifies a pretext task approach that refines representations by sampling an image, creating two augmented views, and passing them through an online and target network. This method operates without relying on negative pairs, showcasing its adaptability to various remote sensing applications. Self-supervised tasks, such as jigsaw puzzles and rotation predictions, function as effective regularizers that enhance representation learning by encouraging the model to develop more meaningful feature representations. Recent studies have shown that these tasks exhibit significant applicability in remote sensing data, where they help to unlock the potential of self-supervised learning techniques that have primarily been explored in larger image datasets like ImageNet. By leveraging these tasks, researchers can improve the performance of models on various remote sensing applications, addressing challenges related to data annotation and representation uniformity. [22, 58, 59, 60]

In object detection, advanced methods such as SSODP and S4OD enhance performance by integrating self-supervised pretext tasks, which facilitate the learning of robust representations from unlabeled data. These pretext tasks, designed to extract meaningful semantic information, serve as auxiliary learning objectives that complement the primary detection tasks, ultimately leading to improved accuracy and efficiency in detecting objects across diverse datasets. [61, 62, 63, 64, 65]. SSODP generates multiple augmented views of an image and matches box features across these views to learn robust representations, while S4OD uses a self-supervised loss to enhance object detection capabilities.

Overall, pretext tasks are integral to advancing representation learning in remote sensing, offering efficient solutions for extracting valuable insights from complex and high-dimensional data. By utilizing self-supervised learning (SSL) techniques, models can significantly enhance their performance across a variety of remote sensing applications. This approach not only reduces the reliance on extensive labeled datasets—which are often costly and labor-intensive to create—but also capitalizes on the vast amounts of unlabeled remote sensing data available. Recent studies have demonstrated that SSL can outperform traditional supervised learning methods, particularly in tasks like scene classification, by effectively leveraging the unique characteristics of remote sensing imagery. This underscores the transformative potential of SSL in advancing earth observation technologies and methodologies. [33, 22, 24]

3.4 Self-supervised Multi-task and Multi-modal Learning

Self-supervised learning (SSL) has made significant strides in enhancing multi-task and multi-modal learning, particularly within the field of remote sensing, where the ability to effectively integrate diverse data sources and tasks is essential for improving model performance and generalization. Recent studies have shown that SSL can notably reduce error rates in few-shot learning scenarios, even with small datasets, highlighting its potential for optimizing model performance in remote sensing applications. Furthermore, a review of current SSL methodologies reveals that while SSL has achieved considerable success in computer vision, its full potential in earth observation remains underexplored, suggesting promising avenues for future research in self-supervised learning for remote sensing.

[22, 66]. By leveraging the strengths of multiple modalities and tasks, SSL frameworks can effectively address the challenges of data scarcity and feature extraction in complex environments.

The Self-Supervised Multi-task Multimodal Sentiment Analysis Network (Self-MM) exemplifies the incorporation of multi-task and multi-modal learning approaches, demonstrating how modality-specific representations can be learned to enhance sentiment analysis tasks [44]. This approach underscores the potential of SSL to harness diverse data streams for improved representation learning.

In remote sensing, the Fusion Transformer employs SSL techniques to enhance activity recognition, showcasing the adaptability of multi-modal learning approaches in extracting meaningful features from diverse sensor data [19]. By fusing information from different modalities, such as RGB and depth data, the Fusion Transformer exemplifies the power of SSL in integrating multi-modal data for robust performance across various tasks.

The Multi-Task Spatio-Temporal Network (MTSTN) further illustrates the application of SSL in multi-task learning by utilizing self-supervised techniques to improve air quality inference accuracy [45]. Through the integration of spatio-temporal data, MTSTN demonstrates the efficacy of SSL in enhancing prediction accuracy across multiple environmental monitoring tasks.

Moreover, the MLF-VO framework highlights the benefits of incrementally fusing RGB and inferred depth information across multiple layers, allowing for better utilization of both modalities in estimating relative pose [46]. This approach exemplifies the integration of multi-modal data within SSL frameworks to improve the accuracy and robustness of pose estimation tasks.

The VISPE method employs multiview consistency regularization to achieve view-invariant stochastic prototype embeddings, showcasing a self-supervised multi-task approach that enhances representation learning through the integration of multiple views [4]. This method highlights the potential of SSL to exploit view consistency for improved feature extraction and generalization.

Additionally, the application of reachability embeddings across five geospatial tasks, as discussed by Ganguli et al., demonstrates the adaptability of self-supervised multi-task learning in multimodal computer vision [47]. By applying SSL techniques across diverse tasks, this approach underscores the scalability and versatility of multi-task learning in remote sensing applications.

3.5 Self-distillation and Knowledge Transfer

Method Name	Knowledge Transfer	Representation Learning	Architectural Choices
SD-DAC[48]	Knowledge Distillation	Enhance Feature Learning	Resnet Architecture
SL+MLP[50]	Feature Transferability	Robust Representations	Mlp Projector
SCPNet[51]	Self-supervised Learning	Feature Map Projection	Weight-shared Network
V2Net[3]	Small Datasets	Robust Representations	Two-stage Computational
NSSL-WLD[49]	-	Self-supervised Learning	Byol, Barlow Twins

Table 2: Table 1 presents a comparative analysis of various self-supervised learning methods with respect to their knowledge transfer mechanisms, representation learning strategies, and architectural choices. This table highlights the diversity in approaches and architectures used to enhance model performance and generalization in remote sensing applications.

Self-distillation and knowledge transfer are pivotal strategies in self-supervised learning (SSL), particularly within remote sensing, where they enhance model efficiency and generalization capabilities. Self-distillation involves transferring knowledge from a complex model to a simpler one, leveraging unlabeled data to boost performance without extensive labeled datasets [48]. This approach is especially advantageous in remote sensing, where labeled data is often scarce and costly to acquire.

Table 2 provides a detailed comparison of different self-supervised learning methods, emphasizing their knowledge transfer techniques, representation learning strategies, and architectural decisions, which are crucial for improving model efficiency in remote sensing.

The effectiveness of self-distillation is highlighted by its ability to improve feature transferability, reduce forgetting, and increase representation diversity, leading to enhanced performance in continual learning scenarios [50]. This is crucial in remote sensing applications, where models must adapt to diverse and evolving data landscapes.

In the realm of representation learning, self-distillation techniques have been shown to generalize knowledge effectively across different modalities, overcoming significant modality gaps and offsets.

This is exemplified by frameworks that integrate diverse data sources, ensuring robust performance across various remote sensing tasks [51]. Additionally, the use of convolutional filters optimized through self-supervised objectives demonstrates the potential of SSL to enhance texture and visual feature extraction, which is vital for accurate remote sensing analysis [3].

The systematic evaluation of SSL techniques across various CNN architectures underscores the importance of architectural choices in determining representation quality [1]. This highlights the necessity of selecting appropriate architectures to maximize the benefits of self-distillation and knowledge transfer in remote sensing contexts.

Furthermore, non-contrastive representation learning methods leverage data structure through augmentations, enabling models to learn meaningful representations without labeled data [49]. This approach is particularly relevant in remote sensing, where large-scale unlabeled datasets are prevalent and traditional supervised methods are limited by data availability.

Self-distillation and knowledge transfer in self-supervised learning (SSL) offer scalable and effective strategies to address the challenges of training deep learning models on large and complex datasets in remote sensing, particularly in scenarios where annotated data is limited. By leveraging vast amounts of unlabeled remote sensing images, SSL can enhance model performance without the need for extensive human annotation, thereby improving the applicability of deep learning techniques in real-world remote sensing applications. This approach not only minimizes the dependency on labeled samples but also facilitates the development of high-performance models tailored to the unique characteristics of remote sensing data, such as multispectral and hyperspectral imagery [22, 33, 59, 24, 67]. By leveraging these techniques, models can achieve superior performance and generalization, paving the way for significant advancements in remote sensing technologies.

Feature	Contrastive Learning Methods	Generative and Reconstruction-based Techniques	Pretext Tasks and Representation Learning
Data Handling	Multi-scale Inputs	Masked Reconstruction	Data Augmentation
Learning Approach	Adversarial Learning	Synthetic Data Generation	Multi-task Pretraining
Application Focus	Remote Sensing	Feature Extraction	Image Interpretation

Table 3: This table provides a comparative analysis of self-supervised learning techniques, categorizing them into contrastive learning methods, generative and reconstruction-based techniques, and pretext tasks. It highlights key features such as data handling, learning approach, and application focus, offering insights into their diverse methodologies and applications in remote sensing and feature extraction.

4 Multi-modal Data Integration and Image Fusion

4.1 Techniques for Multi-modal Data Integration

The integration of multi-modal data in remote sensing is crucial for enhancing analytical accuracy and model performance. Techniques such as combining ultra-wide and telephoto images significantly improve SSL in super-resolution tasks, enhancing spatial resolution and detail [42]. SSL facilitates effective data fusion by integrating multiview data without view labels, thereby improving representation extraction from diverse modalities [4].

Reachability embeddings further enrich feature representations by merging GPS trajectory data with spatial context, enhancing computer vision tasks [47]. The Automated Transformation Policy Framework (ATP) enhances feature learning across modalities by estimating transformation distributions and generating pretext tasks [6]. Insights from medical imaging, where multimodal data fusion has improved model generalizability, underscore the broad applicability of these techniques [68].

Data fusion methods utilizing multiple datasets for self-supervised pre-training highlight the potential of diverse data sources to enhance feature extraction and model performance [13]. Integrating multi-modal data improves predictive accuracy by automatically pairing diverse data types, such as optical satellite images and sensor outputs, based on geographic location and time. This integration enhances label efficiency and model performance in tasks like image classification and semantic segmentation. Techniques such as Multi-Task Pretraining (MTP) and frameworks like Contrastive Mask Image Distillation (CMID) leverage multi-modal datasets to learn robust representations adaptable to various remote sensing applications [69, 30, 57, 54, 34].

4.2 Image Fusion Strategies

Image fusion strategies are essential in remote sensing for synthesizing information from various modalities, such as multispectral, hyperspectral, and radar imagery, to enhance scene representation accuracy. Techniques like multi-view representation learning and spectral analysis capture complex data relationships, improving image clustering and analysis robustness [70, 54]. This integration enhances resolution, contrast, and overall image quality, facilitating precise environmental monitoring and decision-making.

Deep learning models using SSL frameworks, such as TransFuse with a unified transformer-based architecture, effectively integrate multi-modal data, enhancing adaptability across applications [15]. Generative models within SSL frameworks augment image fusion by generating semantically consistent augmentations, increasing representation diversity and robustness [17]. Attention mechanisms, exemplified by the self-supervised spectral-spatial attention-based vision transformer (SSVT), enhance data analysis by focusing on task-relevant features [18].

SSL techniques address challenges related to data heterogeneity and sensor variability by leveraging advanced multi-modal pretext tasks and innovative background augmentation techniques. These methods preserve critical information from diverse modalities, enhancing generalization and performance in remote sensing applications like image classification and change detection. This approach improves label and parameter efficiency, addressing the challenges of limited training data in Earth observation contexts [30, 71, 32, 57].

Image fusion strategies significantly enhance the quality and interpretability of remote sensing imagery through advanced techniques like multi-modal pretraining and SSL, improving environmental monitoring and geospatial analysis. These strategies enable robust models to efficiently learn from large volumes of unlabelled Earth observation data, facilitating applications such as accurate crop nitrogen status prediction and various downstream tasks like scene classification and semantic segmentation [30, 18, 32, 34].

4.3 Challenges in Multi-modal Data Integration

Integrating multi-modal data in remote sensing presents challenges that can hinder SSL methods' effectiveness. A primary concern is the absence of a mathematical framework to justify SSL performance, leading to uncertainty regarding its applicability across various scenarios [72]. This lack of theoretical grounding complicates predictions of SSL techniques' efficacy when applied to diverse data modalities.

High computational costs associated with dual-branch SSL methods pose another challenge, as these methods often require large-scale datasets for optimal performance, limiting their accessibility in resource-constrained environments [67]. The computational demands of processing and integrating multiple data sources necessitate developing more efficient algorithms to reduce resource consumption while maintaining accuracy.

Label-efficient learning studies often lack comprehensive evaluation metrics, limiting their generalizability, particularly in remote sensing applications where diverse environmental conditions require adaptable models [2]. The absence of standardized evaluation frameworks can lead to inconsistent results and hinder the development of universally applicable solutions.

To address these challenges, developing theoretical models that clarify the mechanisms driving SSL performance is essential. Probabilistic models in SSL can enhance understanding of the intricate relationships between data modalities and SSL outcomes, particularly through generative latent variable frameworks that capture semantic similarities across different representations. This approach facilitates the integration of prior knowledge into models, improving positive sample selection and leading to more robust learning processes, as demonstrated by methods like Guided Positive Sampling Self-Supervised Learning (GPS-SSL) and SimVAE [72, 23]. Enhancing computational efficiency through algorithmic innovations and utilizing smaller, targeted datasets can improve SSL methods' accessibility and scalability. Establishing comprehensive evaluation metrics that encompass diverse scenarios will also enhance the generalizability and applicability of multi-modal data integration techniques in remote sensing.

5 Applications in Remote Sensing

The application of self-supervised learning (SSL) in remote sensing is expanding due to its ability to leverage unlabeled data, addressing the frequent scarcity of labeled datasets. SSL has notably improved scene classification and object detection by exploiting the intrinsic structures in remote sensing imagery, enhancing task accuracy and efficiency. This section delves into the advancements and methodologies in these areas, underscoring SSL's transformative impact on remote sensing research.

5.1 Scene Classification and Object Detection

SSL has significantly advanced scene classification and object detection in remote sensing by enabling high-performance training with extensive unlabeled datasets, thus reducing dependence on expensive labeled data. Techniques involving multispectral and hyperspectral imaging have enhanced classification accuracy and efficiency, supporting global mapping applications [22, 30, 33, 32]. Mazumdar's work [73] exemplifies SSL's effectiveness in segmentation tasks, achieving notable results in building and road segmentation with limited labeled data.

Hierarchical self-supervised learning (HSSL) further demonstrates SSL's real-world applicability, as shown by Zheng [74]. Ganguli et al. [47] report substantial performance improvements in scene classification and object detection using SSL, while Hnaff's Odin framework [9] achieves state-of-the-art object detection results across various modalities.

The VISPE method, as discussed by Ho [4], showcases SSL's adaptability in remote sensing through strong generalization capabilities. Jin [25] provides insights into SSL's potential for analogous applications. Parthasarathy et al. [3] highlight V2Net's data efficiency over traditional deep learning, enhancing feature extraction in remote sensing tasks. Zhuang's approach [5] and Kolesnikov's experiments [1] further emphasize the importance of architectural choices in maximizing SSL benefits.

The integration of SSL into scene classification and object detection marks a significant advancement, enabling high-performance pre-training models from abundant unlabeled images. This approach reduces reliance on annotated datasets and outperforms traditional methods, particularly with multispectral and hyperspectral data. Techniques like Extreme-Multi-Patch Self-Supervised Learning (EMP-SSL) enhance SSL efficiency, achieving substantial performance improvements with fewer training epochs [33, 24].

5.2 Change Detection and Semantic Segmentation

SSL techniques significantly enhance change detection and semantic segmentation in remote sensing by utilizing large volumes of unlabeled data, thus reducing reliance on labor-intensive labeled datasets. Recent studies demonstrate SSL's effectiveness in leveraging temporal consistency and semantic information to improve model accuracy and generalization [33, 75, 22, 26].

Advancements in SSL frameworks have notably improved change detection tasks by exploiting the temporal dynamics of multi-temporal data. The Multi-Task Pretraining (MTP) paradigm, employing a shared encoder with task-specific decoders, enhances performance across tasks, including change detection. Models like RemoteCLIP integrate vision-language capabilities for robust feature extraction and semantic understanding, optimizing temporal data analysis [31, 24, 57]. SSL's integration with multi-modal data extracts robust features invariant to changes in lighting, seasonality, and atmospheric conditions, crucial for monitoring environmental changes and urban development.

In semantic segmentation, SSL provides a scalable approach for acquiring rich feature representations, significantly reducing reliance on extensive labeled datasets. Advancements like EMP-SSL demonstrate impressive convergence rates and out-of-domain dataset transferability, enabling robust performance with minimal labeled data. Pretext tasks, such as image inpainting and jigsaw puzzle solving, enhance spatial dependency learning, improving segmentation accuracy [76, 77, 59, 24].

Integrating generative models into SSL frameworks enhances semantic segmentation by producing semantically consistent image augmentations, improving training data diversity and quality. This approach increases Top-1 accuracy by up to 10

SSL techniques in change detection and semantic segmentation provide powerful tools for advancing remote sensing applications. By leveraging SSL frameworks, models achieve enhanced performance while minimizing dependence on labeled data. Methods like Guided Positive Sampling SSL (GPS-SSL) and EMP-SSL facilitate rapid convergence, paving the way for significant progress in environmental monitoring and geospatial analysis [23, 66, 24].

5.3 Anomaly Detection and Data Quality Improvement

SSL has become pivotal in enhancing anomaly detection and data quality in remote sensing by leveraging vast amounts of unlabeled data. This capability is crucial where labeled datasets are often scarce and costly [59]. SSL frameworks have advanced feature representation learning, essential for identifying anomalies and improving data quality across various tasks.

SSL's application in anomaly detection benefits from effective data augmentation strategies, enhancing detection performance [78]. By employing appropriate augmentations, SSL captures underlying data distributions more effectively, improving anomaly identification in applications like environmental monitoring and disaster response.

SSL enhances feature extraction from complex data structures, such as point clouds, underscoring its utility in anomaly detection and data quality improvement [39]. Robust feature representations derived from SSL frameworks facilitate outlier isolation and data quality enhancement, ensuring accurate remote sensing data analyses.

In improving data quality, SSL techniques learn invariant representations that enhance image quality and consistency. For instance, SSL applied in synthetic aperture radar (SAR) demonstrates superior segmentation performance, especially for small-scale features [41]. This capability is vital for maintaining remote sensing data integrity, where high-quality representations are crucial for precise analysis.

SSL provides robust solutions for anomaly detection and data quality improvement in remote sensing. Advanced frameworks like Guided Positive Sampling (GPS-SSL) and Extreme-Multi-Patch Self-Supervised Learning (EMP-SSL) enhance performance while minimizing labeled data needs. This approach facilitates effective positive sample generation based on prior knowledge and allows for rapid training convergence, achieving high accuracy with minimal augmentations. These advancements enable transformative improvements in environmental monitoring and geospatial analysis, where labeled data availability is limited [78, 66, 23, 24, 76].

5.4 Agricultural and Environmental Applications

SSL is transformative in agricultural and environmental monitoring applications, addressing the challenge of limited labeled data. The high cost of labeled datasets makes SSL attractive, enabling meaningful feature extraction from vast unlabeled data [79]. This capability is especially beneficial in precision agriculture, where SSL enhances crop monitoring and yield prediction accuracy.

One notable application in agriculture is improving nitrogen status estimation using the self-supervised spectral-spatial attention-based vision transformer (SSVT). This approach enhances hyperspectral data analysis, leading to accurate plant health and nutrient assessments, crucial for optimizing crop yield and environmental sustainability [18]. By integrating spectral and spatial information, SSVT provides a comprehensive understanding of plant conditions, facilitating better agricultural management decisions.

In plant phenotyping, SSL frameworks are effective in non-invasive leaf segmentation, critical for assessing plant health and growth. The self-supervised leaf segmentation framework shows promise in extracting detailed phenotypic traits without extensive manual labeling, streamlining the phenotyping process for large-scale studies [14]. This application underscores SSL's potential to revolutionize plant science with scalable and efficient data analysis solutions.

SSL also enhances environmental monitoring, improving land cover change detection, vegetation dynamics, and water quality assessment. SSL's ability to independently extract knowledge from diverse unlabeled data sources makes it ideal for monitoring intricate environmental systems. This advantage is significant where traditional supervised methods struggle due to data availability and variability, as SSL leverages diverse datasets' rich structure without extensive manual annotation.

Advancements like EMP-SSL and GPS-SSL enhance SSL's effectiveness, achieving high performance in remote sensing and earth observation applications [23, 22, 24].

SSL offers robust solutions for agricultural and environmental applications, providing frameworks for extracting valuable insights from complex datasets. By employing techniques like EMP-SSL and GPS-SSL, researchers enhance model performance while minimizing labeled data dependence. This approach accelerates training, achieving convergence in just one epoch for various image datasets, and allows effective prior knowledge integration into the learning process. These innovations pave the way for substantial advancements in sustainable agriculture and environmental conservation, enabling robust machine learning models that efficiently analyze and interpret vast unlabeled data in applications like precision farming and plant disease diagnosis [66, 23, 2, 24, 76].

6 Challenges and Future Directions

6.1 Challenges in Data Quality and Annotation

The application of self-supervised learning (SSL) in remote sensing is impeded by data quality and annotation challenges, largely due to the scarcity and variability of labeled datasets. Traditional approaches reliant on extensive annotations face quality issues due to the high costs and time-intensive nature of large-scale labeling [74]. This is particularly problematic in remote sensing, where complex data structures demand robust solutions [25]. A major challenge is the limited labeled anomalies in training data, complicating the selection of augmentation strategies that accurately reflect anomaly-generating mechanisms [41]. Additionally, dataset imbalances with long-tailed class distributions exacerbate SSL's generalization challenges [3].

High computational and carbon costs further limit SSL's accessibility and practicality [47]. Excessive training times are prohibitive, especially in resource-constrained settings. Moreover, model adaptability to new contexts and dynamic environments remains a critical issue [25]. Despite progress, many studies lack comprehensive benchmarks and systematic evaluations, hindering a clear understanding of various approaches' strengths and weaknesses [5]. This absence of standardized evaluation frameworks can lead to inconsistent results and impede the development of universally applicable solutions in remote sensing.

Addressing these challenges requires advanced SSL frameworks that can handle noisy and imbalanced data, reduce reliance on detailed annotations, and enhance generalizability across diverse applications. Innovative approaches like Extreme-Multi-Patch Self-Supervised Learning (EMP-SSL) and Guided Positive Sampling Self-Supervised Learning (GPS-SSL) effectively leverage large volumes of unlabeled data, improving classification performance and representation transferability across datasets [66, 22, 23, 33, 24].

6.2 Methodological Limitations and Performance Bottlenecks

Methodological limitations and performance bottlenecks challenge SSL in remote sensing, primarily due to dataset complexity and diversity requiring robust approaches. SSL methods are sensitive to hyperparameter choices, especially in balancing objectives within composite loss functions, impacting performance [80]. Streamlined methods that simplify parameter tuning while maintaining high performance are needed. Existing frameworks, like Odin, face challenges in complex scenes where unsupervised discovery may not fully capture object relationships, indicating methodological constraints [9]. Similarly, VISPE may underperform compared to fully supervised methods, highlighting the need for methodological advancements [4].

Another issue is the reliance on estimated distribution quality affecting pretext task efficacy [6]. Performance bottlenecks also stem from balancing appearance and spatial invariances, as training on one may diminish effectiveness on the other [81]. Architectural choices often inadequately account for SSL performance in benchmarks, leading to suboptimal evaluations [1]. Establishing standardized benchmarks and evaluation protocols is crucial for understanding the strengths and weaknesses of various approaches.

To address these limitations and bottlenecks, developing frameworks that are efficient, adaptable, and capable of generalizing across diverse datasets and tasks is essential. Recent advancements, such as EMP-SSL and GPS-SSL, demonstrate significant reductions in training epochs and enhancements in

positive sample selection through principled use of prior knowledge [66, 23, 24, 67, 76]. Overcoming these challenges will enhance SSL's effectiveness and applicability in remote sensing, facilitating significant advancements in data analysis and interpretation.

6.3 Integration with Multi-modal and Multi-view Data

Integrating multi-modal and multi-view data in remote sensing presents substantial challenges due to diverse data sources and the complexity of aligning different modalities and views. A critical challenge is the absence of a unified framework capable of managing discrepancies between various data types—such as spectral, spatial, and temporal differences—which can obstruct seamless integration [4]. Sophisticated algorithms are needed to maintain consistency across diverse perspectives. Moreover, the computational demands of processing multi-modal and multi-view data require advanced techniques to manage increased data volume and complexity. Robust synchronization and calibration methods are essential for accurate data fusion, as misalignment can lead to erroneous interpretations [39]. This challenge intensifies in dynamic environments where data sources continually evolve, necessitating real-time integration capabilities to sustain data coherence and relevance.

SSL frameworks offer potential solutions by enabling the extraction of meaningful representations from unlabeled multi-modal and multi-view data, facilitating more effective integration. However, SSL methods must adapt to the unique characteristics of each modality and view, ensuring that learned representations are comprehensive and invariant [12]. Developing innovative pretext tasks tailored to multi-modal and multi-view data is essential for enhancing the robustness and accuracy of SSL frameworks in remote sensing applications. Despite the challenges of limited labeled training data in Earth observation, integrating multi-modal and multi-view data presents a transformative opportunity for advancing remote sensing technologies. By leveraging diverse data sources, researchers can develop advanced models like the Multi-Pretext Masked Autoencoder (MP-MAE), which have shown superior performance in tasks such as image classification and semantic segmentation. This approach enhances label and parameter efficiency and facilitates automatic data pairing from various sensors, driving innovation in global-scale remote sensing applications [30, 54]. By harnessing complementary information from different data sources, researchers can achieve more accurate insights into complex environmental systems, paving the way for improved decision-making and analysis. Addressing multi-modal and multi-view integration challenges is crucial to unlocking the full potential of SSL in remote sensing.

6.4 Generalization and Robustness Concerns

Generalization and robustness are paramount challenges in applying SSL to remote sensing, where dataset diversity and complexity necessitate models capable of adapting to new and unseen data. A primary concern is SSL frameworks' ability to generalize across different environmental conditions and sensor modalities, which often exhibit significant variability [4]. This variability can cause inconsistencies in model performance, particularly with datasets differing from those used in training. SSL model robustness is further challenged by noise and artifacts in remote sensing data, which can significantly affect the accuracy and reliability of learned representations [39]. Ensuring models maintain high performance amid such perturbations is essential for practical applications. Additionally, integrating multi-modal data introduces further complexity, as models must effectively manage inherent discrepancies between different data types [12].

To address these concerns, developing SSL techniques that enhance model adaptability and resilience to diverse data conditions is crucial. This includes creating robust pretext tasks that capture essential features of multi-modal data while minimizing sensitivity to noise and variability. Investigating innovative augmentation strategies and implementing ensemble methods can significantly enhance SSL models' generalization and robustness, particularly by enabling them to leverage abundant unlabeled data and improve performance across diverse datasets and imaging modalities. This strategy aligns with recent SSL advancements, which highlight its potential to overcome traditional supervised learning limitations that require extensive labeled datasets, facilitating more efficient earth observation tasks [33, 23, 22, 24]. To fully leverage SSL's capabilities in remote sensing, effectively addressing generalization and robustness challenges is essential. Advancements in methods like EMP-SSL and GPS-SSL demonstrate that enhanced efficiency and prior knowledge integration can significantly improve the quality of learned representations and their transferability across diverse datasets [23, 24]. Developing models that adapt effectively to real-world data complexities will yield

more accurate and reliable insights, paving the way for substantial advancements in environmental monitoring and geospatial analysis.

6.5 Future Directions in Algorithmic and Architectural Advancements

Future research in SSL for remote sensing is poised for advancement through several promising avenues. One critical direction involves exploring advanced label-efficient techniques that can enhance the scalability of existing methods and their applicability in real-world agricultural settings [2]. This includes developing adaptable models and enhancing target-agnostic learning frameworks, which are vital for advancing algorithmic capabilities in remote sensing [25]. Integrating additional domain knowledge into proxy tasks and developing more complex tasks can further enhance feature learning, particularly in 3D data contexts [5]. This approach is crucial for improving SSL frameworks' robustness and effectiveness, especially when applied to larger datasets and diverse data modalities [6]. Enhancing the VISPE method's performance in semi-supervised scenarios represents another promising research direction, suggesting potential algorithmic advancements in SSL [4]. Additionally, exploring the identifiability properties of methods maximizing the entropy reconstruction (ER) bound and refining the understanding of entropy dynamics in multi-view self-supervised learning (MVSSL) will contribute to more robust and adaptable SSL models [40].

Future research should also focus on adapting models to learn from diverse natural scenes by minimizing variability in local neighborhoods while maximizing global variability [3]. This approach will enhance SSL techniques' adaptability to various environmental conditions and data types, paving the way for significant advancements in remote sensing technologies. These future directions highlight the potential for significant SSL advancements in remote sensing, facilitating enhanced environmental monitoring and geospatial analysis. By refining algorithmic strategies and architectural designs, researchers can develop advanced models that are not only more resilient but also versatile enough to tackle the intricate challenges posed by real-world remote sensing data. Recent developments, such as RemoteCLIP—a vision-language foundation model—demonstrate the potential of integrating SSL and multi-task pretraining to improve model capabilities in tasks like zero-shot image classification and object counting. Furthermore, implementing unified data formats and extensive pretraining datasets allows these models to surpass existing benchmarks, highlighting their adaptability and robustness in diverse remote sensing applications [31, 57].

6.6 Exploration of Emerging Applications and Domains

SSL is set to transform a range of emerging applications and domains within remote sensing by leveraging its ability to autonomously extract meaningful features from unlabeled data. A promising area of research is integrating SSL into advanced environmental monitoring systems, significantly improving climate change impact detection, enhancing biodiversity conservation efforts, and optimizing sustainable resource management practices. By utilizing innovative approaches such as Guided Positive Sampling SSL (GPS-SSL) and Contrastive Mask Image Distillation (CMID), these systems can leverage vast amounts of unlabeled data to generate more accurate representations, facilitating effective decision-making in environmental conservation and resource management [22, 23, 64, 34, 24]. SSL techniques can enhance model precision and scalability in these applications, enabling more effective monitoring and decision-making processes.

In smart agriculture, SSL offers significant potential for optimizing crop management and improving yield predictions. By analyzing multi-modal remote sensing data, such as multispectral and hyperspectral imagery, SSL can provide detailed insights into crop health, soil conditions, and pest infestations. This capability is essential for implementing precision agriculture practices, which leverage advanced machine learning techniques to enhance crop productivity while reducing environmental impacts, such as nitrogen runoff, through accurate, real-time monitoring of crop nutrient status using innovative, label-efficient models and remote sensing technologies [18, 2].

Integrating SSL in urban planning and smart city initiatives is another emerging application area. By leveraging vast data from various sensors, SSL can facilitate analyzing urban dynamics, infrastructure development, and traffic management. By employing advanced self-supervised learning techniques and integrating auxiliary information, urban planners can enhance strategy efficiency, leading to more effective resource allocation, improved infrastructure development, and a higher quality of life for city residents [64, 82].

Furthermore, SSL can significantly enhance disaster response and management by enabling precise damage assessments and optimizing resource allocation during natural disasters. By leveraging advanced techniques such as Guided Positive Sampling (GPS-SSL) and Extreme-Multi-Patch Self-Supervised Learning (EMP-SSL), which improve data representation quality and reduce reliance on extensive training epochs, SSL can effectively utilize vast amounts of unlabeled data to provide timely insights in critical situations [23, 24]. By analyzing remote sensing data in real-time, SSL can offer critical insights into disaster impacts, enabling more effective response strategies and minimizing human and economic losses.

Moreover, SSL can play a pivotal role in advancing autonomous systems and robotics, enhancing the perception and decision-making capabilities of autonomous vehicles and drones. By extracting robust features from various data sources, SSL improves systems' abilities to navigate complex environments and execute tasks with enhanced accuracy and reliability. Recent advancements, such as GPS-SSL, demonstrate that injecting prior knowledge into the positive sample selection process significantly boosts representation quality, even with minimal data augmentations. Additionally, methods like EMP-SSL have shown that increasing the number of image crops can drastically reduce training time while achieving high performance across multiple datasets. These innovations underscore SSL's potential to leverage unlabeled data to develop versatile models that are robust to dataset imbalances and adaptable to varying domains, ultimately enhancing task performance in real-world applications [66, 38, 23, 64, 24].

The exploration of emerging applications and domains for SSL in remote sensing holds significant promise for advancing environmental, agricultural, urban, and autonomous systems technologies. By harnessing SSL frameworks' capabilities, researchers can devise cutting-edge solutions that effectively address the intricate challenges prevalent in remote sensing applications. This approach enhances the ability to generate high-quality representations from unlabeled data and facilitates integrating prior knowledge into the learning process. For instance, methods like GPS-SSL significantly improve performance by optimizing positive sample selection, while models such as RemoteCLIP leverage large-scale datasets to learn robust visual features and semantic text embeddings. These advancements pave the way for substantial progress in various remote sensing tasks, including zero-shot classification, image-text retrieval, and object counting, ultimately reducing reliance on extensive labeled datasets and enabling more efficient global mapping efforts [33, 23, 31].

7 Conclusion

Self-supervised learning (SSL) has demonstrated its transformative potential in remote sensing by effectively utilizing vast unlabeled datasets to enhance data analysis and interpretation. By minimizing reliance on labeled data, SSL enables advanced insights across a range of remote sensing applications. Its integration with multimodal approaches is pivotal for advancing research in this domain. Noteworthy contributions of SSL include its impact on scene classification, object detection, change detection, and semantic segmentation, where it has improved model performance and robustness. Frameworks such as TransFuse exemplify state-of-the-art results in image fusion, highlighting SSL's utility. The Toulouse Hyperspectral Data Set provides a valuable benchmark for evaluating SSL techniques, marking significant progress in urban land cover classification.

The survey emphasizes the importance of label-efficient learning in enhancing agricultural productivity, suggesting a need for further refinement of these methodologies. The architecture choice is crucial to SSL performance, indicating potential research directions. Future studies should focus on optimizing hyperparameter tuning and broadening the framework to include additional SSL methods not yet analyzed. Moreover, exploring benchmarks across diverse geographical regions and Earth Observation tasks is essential to fully understand SSL's impact and potential. Developing more efficient and adaptable SSL frameworks to manage the complexities of multimodal and multi-view data will be crucial for realizing SSL's full promise in remote sensing.

References

[1] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1920–1929, 2019.

[2] Jiajia Li, Dong Chen, Xinda Qi, Zhaojian Li, Yanbo Huang, Daniel Morris, and Xiaobo Tan. Label-efficient learning in agriculture: A comprehensive review, 2023.

[3] Nikhil Parthasarathy and Eero P. Simoncelli. Self-supervised learning of a biologically-inspired visual texture model, 2020.

[4] Chih-Hui Ho, Bo Liu, Tz-Ying Wu, and Nuno Vasconcelos. Exploit clues from views: Self-supervised and regularized learning for multiview object recognition, 2020.

[5] Xinrui Zhuang, Yuexiang Li, Yifan Hu, Kai Ma, Yujiu Yang, and Yefeng Zheng. Self-supervised feature learning for 3d medical images by playing a rubik's cube, 2019.

[6] Seunghan Yang, Debasmit Das, Simyung Chang, Sungrack Yun, and Fatih Porikli. Distribution estimation to automate transformation policies for self-supervision, 2021.

[7] Mustafa Taha Koçyiğit, Timothy M. Hospedales, and Hakan Bilen. Accelerating self-supervised learning via efficient training strategies, 2022.

[8] Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style, 2022.

[9] Olivier J. Hénaff, Skanda Koppula, Evan Shelhamer, Daniel Zoran, Andrew Jaegle, Andrew Zisserman, João Carreira, and Relja Arandjelović. Object discovery and representation networks, 2022.

[10] Ilyass Moummad, Romain Serizel, and Nicolas Farrugia. Self-supervised learning for few-shot bird sound classification, 2024.

[11] Qianwen Meng, Hangwei Qian, Yong Liu, Yonghui Xu, Zhiqi Shen, and Lizhen Cui. Unsupervised representation learning for time series: A review, 2023.

[12] Yixin Liu, Ming Jin, Shirui Pan, Chuan Zhou, Yu Zheng, Feng Xia, and S Yu Philip. Graph self-supervised learning: A survey. *IEEE transactions on knowledge and data engineering*, 35(6):5879–5900, 2022.

[13] Fatema-E-Jannat, Sina Gholami, Jennifer I. Lim, Theodore Leng, Minhaj Nur Alam, and Hamed Tabkhi. Multi-oct-selfnet: Integrating self-supervised learning with multi-source data fusion for enhanced multi-class retinal disease classification, 2024.

[14] Xufeng Lin, Chang-Tsun Li, Scott Adams, Abbas Kouzani, Richard Jiang, Ligang He, Yongjian Hu, Michael Vernon, Egan Doeven, Lawrence Webb, Todd Mcclellan, and Adam Guskic. Self-supervised leaf segmentation under complex lighting conditions, 2022.

[15] Linhao Qu, Shaolei Liu, Manning Wang, Shiman Li, Siqi Yin, Qin Qiao, and Zhijian Song. Transfuse: A unified transformer-based image fusion framework using self-supervised learning, 2022.

[16] Pallavi Jain, Bianca Schoen-Phelan, and Robert Ross. Self-supervised learning for invariant representations from multi-spectral and sar images, 2022.

[17] Sana Ayromlou, Vahid Reza Khazaie, Fereshteh Forghani, and Arash Afkanpour. Can generative models improve self-supervised representation learning?, 2024.

[18] Xin Zhang, Liangxiu Han, Tam Sobeih, Lewis Lappin, Mark Lee, Andew Howard, and Aron Kisdi. The self-supervised spectral-spatial attention-based transformer network for automated, accurate prediction of crop nitrogen status from uav imagery, 2022.

[19] Armand K. Koupai, Mohammud J. Bocus, Raul Santos-Rodriguez, Robert J. Piechocki, and Ryan McConville. Self-supervised multimodal fusion transformer for passive activity recognition, 2022.

[20] Randall Balestriero and Yann LeCun. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods, 2022.

[21] Suhas Kotha, Anirudh Koul, Siddha Ganju, and Meher Kasam. Celestial: Classification enabled via labelless embeddings with self-supervised telescope image analysis learning, 2022.

[22] Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, Lichao Mou, and Xiao Xiang Zhu. Self-supervised learning in remote sensing: A review. *IEEE Geoscience and Remote Sensing Magazine*, 10(4):213–247, 2022.

[23] Aarash Feizi, Randall Balestriero, Adriana Romero-Soriano, and Reihaneh Rabbany. Gps-ssl: Guided positive sampling to inject prior into self-supervised learning, 2024.

[24] Shengbang Tong, Yubei Chen, Yi Ma, and Yann Lecun. Emp-ssl: Towards self-supervised learning in one training epoch, 2023.

[25] Yuan Jin, Wray Buntine, Francois Petitjean, and Geoffrey I. Webb. Discriminative, generative and self-supervised approaches for target-agnostic learning, 2020.

[26] Marrit Leenstra, Diego Marcos, Francesca Bovolo, and Devis Tuia. Self-supervised pre-training enhances change detection in sentinel-2 imagery, 2021.

[27] Anthony Fuller, Koreen Millard, and James R. Green. Croma: Remote sensing representations with contrastive radar-optical masked autoencoders, 2023.

[28] Romain Thoreau, Laurent Risser, Véronique Achard, Béatrice Berthelot, and Xavier Briottet. Toulouse hyperspectral data set: a benchmark data set to assess semi-supervised spectral representation learning and pixel-wise classification techniques, 2024.

[29] Yidan Liu, Weiying Xie, Kai Jiang, Jiaqing Zhang, Yunsong Li, and Leyuan Fang. Hyperspectral anomaly detection with self-supervised anomaly prior, 2024.

[30] Mearth: Exploring multi-modal pr.

[31] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing, 2024.

[32] Xinye Wanyan, Sachith Seneviratne, Shuchang Shen, and Michael Kirley. Extending global-local view alignment for self-supervised learning with remote sensing imagery, 2024.

[33] Chao Tao, Ji Qi, Weipeng Lu, Hao Wang, and Haifeng Li. Remote sensing image scene classification with self-supervised paradigm under limited labeled samples, 2020.

[34] Dilxat Muhtar, Xueliang Zhang, Pengfeng Xiao, Zhenshi Li, and Feng Gu. Cmid: A unified self-supervised learning framework for remote sensing image understanding, 2023.

[35] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Revisiting contrastive methods for unsupervised learning of visual representations, 2021.

[36] Veenu Rani, Syed Tufael Nabi, Munish Kumar, Ajay Mittal, and Krishan Kumar. Self-supervised learning: A succinct review. *Archives of Computational Methods in Engineering*, 30(4):2761–2775, 2023.

[37] Li Chen, Jonathan Rubin, Jiahong Ouyang, Naveen Balaraju, Shubham Patil, Courosh Mehanian, Sourabh Kulhare, Rachel Millin, Kenton W Gregory, Cynthia R Gregory, Meihua Zhu, David O Kessler, Laurie Malia, Almaz Dessie, Joni Rabiner, Di Coneybeare, Bo Shopsin, Andrew Hersh, Cristian Madar, Jeffrey Shupp, Laura S Johnson, Jacob Avila, Kristin Dwyer, Peter Weimersheimer, Balasundar Raju, Jochen Kruecker, and Alvin Chen. Contrastive self-supervised learning for spatio-temporal analysis of lung ultrasound videos, 2023.

[38] Hong Liu, Jeff Z. HaoChen, Adrien Gaidon, and Tengyu Ma. Self-supervised learning is more robust to dataset imbalance, 2022.

[39] Changyu Zeng, Wei Wang, Anh Nguyen, and Yutao Yue. Self-supervised learning for point clouds data: A survey, 2023.

[40] Borja Rodríguez-Gálvez, Arno Blaas, Pau Rodríguez, Adam Goliński, Xavier Suau, Jason Ramapuram, Dan Busbridge, and Luca Zappella. The role of entropy and reconstruction in multi-view self-supervised learning, 2023.

[41] Xiaoman Zhang, Shixiang Feng, Yuhang Zhou, Ya Zhang, and Yanfeng Wang. Sar: Scale-aware restoration learning for 3d tumor segmentation, 2021.

[42] Zhilu Zhang, Ruohao Wang, Hongzhi Zhang, and Wangmeng Zuo. Self-supervised learning for real-world super-resolution from dual and multiple zoomed observations, 2024.

[43] Jun Chen, Ming Hu, Boyang Li, and Mohamed Elhoseiny. Efficient self-supervised vision pretraining with local masked reconstruction, 2022.

[44] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis, 2021.

[45] Meng Xu, Ke Han, Weijian Hu, and Wen Ji. Fine-gained air quality inference based on low-quality sensing data using self-supervised learning, 2024.

[46] Zijie Jiang, Hajime Taira, Naoyuki Miyashita, and Masatoshi Okutomi. Self-supervised ego-motion estimation based on multi-layer fusion of rgb and inferred depth, 2022.

[47] Swetava Ganguli, C. V. Krishnakumar Iyer, and Vipul Pandey. Scalable self-supervised representation learning from spatiotemporal motion trajectories for multimodal computer vision, 2022.

[48] Mohammed Adnan, Yani A. Ioannou, Chuan-Yung Tsai, and Graham W. Taylor. Domain-agnostic clustering with self-distillation, 2021.

[49] Alexander Marusov and Alexey Zaytsev. Non-contrastive representation learning for intervals from well logs, 2023.

[50] Daniel Marczak, Sebastian Cygert, Tomasz Trzciński, and Bartłomiej Twardowski. Revisiting supervision for continual representation learning, 2024.

[51] Runmin Zhang, Jun Ma, Si-Yuan Cao, Lun Luo, Beinan Yu, Shu-Jie Chen, Junwei Li, and Hui-Liang Shen. Scpnet: Unsupervised cross-modal homography estimation via intra-modal self-supervised learning, 2024.

[52] Hyungtae Lee and Heesung Kwon. Self-supervised contrastive learning for cross-domain hyperspectral image representation, 2022.

[53] Madeline C Schiappa, Yogesh S Rawat, and Mubarak Shah. Self-supervised learning for videos: A survey. *ACM Computing Surveys*, 55(13s):1–37, 2023.

[54] Yingming Li, Ming Yang, and Zhongfei Zhang. A survey of multi-view representation learning. *IEEE transactions on knowledge and data engineering*, 31(10):1863–1883, 2018.

[55] Sushil Thapa. Survey on self-supervised multimodal representation learning and foundation models, 2022.

[56] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1):857–876, 2021.

[57] Di Wang, Jing Zhang, Minqiang Xu, Lin Liu, Dongsheng Wang, Erzhong Gao, Chengxi Han, Haonan Guo, Bo Du, Dacheng Tao, and Liangpei Zhang. Mtp: Advancing remote sensing foundation model via multi-task pretraining, 2024.

[58] Bram Wallace and Bharath Hariharan. Extending and analyzing self-supervised learning across domains, 2020.

[59] Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[60] Aiden Durrant and Georgios Leontidis. Hyperspherically regularized networks for self-supervision, 2022.

[61] Nathaniel Simard and Guillaume Lagrange. Improving few-shot learning with auxiliary self-supervised pretext tasks, 2021.

[62] Yandong Li, Di Huang, Danfeng Qin, Liqiang Wang, and Boqing Gong. Improving object detection with selective self-supervised self-training, 2020.

[63] Trung Dang, Simon Kornblith, Huy Thong Nguyen, Peter Chin, and Maryam Khademi. A study on self-supervised object detection pretraining, 2022.

[64] Carlos Penarrubia, Jose J. Valero-Mas, and Jorge Calvo-Zaragoza. Self-supervised learning for text recognition: A critical survey, 2024.

[65] Xinnan Du, William Zhang, and Jose M. Alvarez. Boosting supervised learning performance with co-training, 2021.

[66] Jong-Chyi Su, Subhransu Maji, and Bharath Hariharan. When does self-supervision improve few-shot learning?, 2020.

[67] Yun-Hao Cao and Jianxin Wu. On improving the algorithm-, model-, and data- efficiency of self-supervised learning, 2024.

[68] Zelong Liu, Andrew Tieu, Nikhil Patel, Georgios Soultanidis, Louisa Deyer, Ying Wang, Sean Huver, Alexander Zhou, Yunhao Mei, Zahi A. Fayad, Timothy Deyer, and Xueyan Mei. Vis-mae: An efficient self-supervised learning approach on medical image segmentation and classification, 2025.

[69] Yongshuo Zong, Oisin Mac Aodha, and Timothy Hospedales. Self-supervised multimodal learning: A survey, 2024.

[70] Jinghua Wang, Adrian Hilton, and Jianmin Jiang. Spectral analysis network for deep representation learning and image clustering, 2020.

[71] Chaitanya K. Ryali, David J. Schwab, and Ari S. Morcos. Characterizing and improving the robustness of self-supervised learning through background augmentations, 2021.

[72] Alice Bizeul, Bernhard Schölkopf, and Carl Allen. A probabilistic model behind self-supervised learning, 2024.

[73] Priyam Mazumdar, Aiman Soliman, Volodymyr Kindratenko, Luigi Marini, and Kenton McHenry. Self-supervised masked digital elevation models encoding for low-resource downstream tasks, 2023.

[74] Hao Zheng, Jun Han, Hongxiao Wang, Lin Yang, Zhuo Zhao, Chaoli Wang, and Danny Z. Chen. Hierarchical self-supervised learning for medical image segmentation based on multi-domain data aggregation, 2021.

[75] Hao Chen, Wenyuan Li, Song Chen, and Zhenwei Shi. Semantic-aware dense representation learning for remote sensing image change detection, 2022.

[76] Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Semi-supervised learning with scarce annotations, 2020.

[77] Shervin Halat, Mohammad Rahmati, and Ehsan Nazerfard. Visual self-supervised learning scheme for dense prediction tasks on x-ray images, 2024.

[78] Jaemin Yoo, Tiancheng Zhao, and Leman Akoglu. Data augmentation is a hyperparameter: Cherry-picked self-supervision for unsupervised anomaly detection is creating the illusion of success, 2023.

[79] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020.

[80] Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective. *arXiv preprint arXiv:2006.05576*, 2020.

[81] Linus Ericsson, Henry Gouk, and Timothy M. Hospedales. Why do self-supervised models transfer? investigating the impact of invariance on downstream tasks, 2022.

[82] Yao-Hung Hubert Tsai, Tianqin Li, Weixin Liu, Peiyuan Liao, Ruslan Salakhutdinov, and Louis-Philippe Morency. Integrating auxiliary information in self-supervised learning, 2021.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.