International Journal of Remote Sensing

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tres20

Semi-supervised contrastive learning for remote sensing: identifying ancient urbanization in the south-central Andes

Jiachen Xu, Junlin Guo, James Zimmer-Dauphinee, Quan Liu, Yuxuan Shi, Zuhayr Asad, D. Mitchell Wilkes, Parker VanValkenburgh, Steven A. Wernke & Yuankai Huo

To cite this article: Jiachen Xu, Junlin Guo, James Zimmer-Dauphinee, Quan Liu, Yuxuan Shi, Zuhayr Asad, D. Mitchell Wilkes, Parker VanValkenburgh, Steven A. Wernke & Yuankai Huo (2023) Semi-supervised contrastive learning for remote sensing: identifying ancient urbanization in the south-central Andes, International Journal of Remote Sensing, 44:6, 1922-1938, DOI: 10.1080/01431161.2023.2192879

To link to this article: https://doi.org/10.1080/01431161.2023.2192879

Published online: 03 Apr 2023.

INTERNATIONAL JOURNAL OF REMOTE SENSING
2023, VOL. 44, NO. 6, 1922–1938
https://doi.org/10.1080/01431161.2023.2192879

Taylor & Francis
Taylor & Francis Group

Semi-supervised contrastive learning for remote sensing: identifying ancient urbanization in the south-central Andes

Jiachen Xu[a*], Junlin Guo[b*], James Zimmer-Dauphinee[c], Quan Liu[b], Yuxuan Shi[a], Zuhayr Asad[b], D. Mitchell Wilkes[b], Parker VanValkenburgh[d], Steven A. Wernke[c] and Yuankai Huo[b]

[a]School of Engineering, Vanderbilt University, Nashville, USA; [b]Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, USA; [c]Department of Anthropology, Vanderbilt University, Nashville, USA; [d]Department of Anthropology, Brown University, Providence, USA

ABSTRACT

Archaeology has long faced fundamental issues of sampling and scalar representation. Traditionally, the local-to-regional-scale views of settlement patterns are produced through systematic pedestrian surveys. Recently, systematic manual survey of satellite and aerial imagery has enabled continuous distributional views of archaeological phenomena at interregional scales. However, such 'brute force' manual imagery survey methods are both time- and labour-intensive, as well as prone to inter-observer differences in sensitivity and specificity. The development of self-supervised learning methods (e.g. contrastive learning) offers a scalable learning scheme for locating archaeological features using unlabelled satellite and historical aerial images. However, archaeological features are generally only visible in a very small proportion relative to the landscape, while the modern contrastive-supervised learning approach typically yields an inferior performance on highly imbalanced datasets. In this work, we propose a framework to address this long-tail problem. As opposed to the existing contrastive learning approaches that typically treat the labelled and unlabelled data separately, our proposed method reforms the learning paradigm under a semi-supervised setting in order to fully utilize the precious annotated data (<7% in our setting). Specifically, the highly unbalanced nature of the data is employed as the prior knowledge in order to form pseudo negative pairs by ranking the similarities between unannotated image patches and annotated anchor images. In this study, we used 95,358 unlabelled images and 5,830 labelled images in order to solve the issues associated with detecting ancient buildings from a long-tailed satellite image dataset. From the results, our semi-supervised contrastive learning model achieved a promising testing balanced accuracy of 79.0%, which is a 3.8% improvement as compared to other state-of-the-art approaches.

ARTICLE HISTORY
Received 23 July 2022
Accepted 8 March 2023

KEYWORDS
Machine Learning; Satellite Imagery; Semi-Supervised Learning; Contrastive Learning

CONTACT Yuankai Huo Huo.yuankai.huo@vanderbilt.edu Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, USA

*First Co-author. Contribute equally to this paper.

1. Introduction

Archaeological structures and settlements are essential sources of information that archaeologists use to study the economic, political, and social systems of ancient civilizations. Conventional approaches to mapping and recording settlement locations at local and regional scales have relied on field-based pedestrian survey methods, which require professionals to physically examine the landscape for evidence of ancient material culture (Banning 2002; Phillips and Willey 1953; Sanders 1961; Balkansky et al. 2000; Alcock and Cherry 2016). However, the scale of field-based surveys is ultimately limited by the physical impedances of fieldwork. Moreover, the distribution of both survey and excavation zones is often unsystematic, which further complicates efforts to synthesize findings across field projects. Since the early 2000s, archaeologists have made use of high-resolution satellite imagery to understand the spatial and structural patterns of the archaeological features at larger scales, including by step-wise visual identification of sites by trained specialists (Hanson and Oltean 2012; Fowler 2002; Lasaponara and Masini 2012; Bewley et al. 2016; Casana 2014; Lin et al. 2014; Parcak 2019). Such research has produced novel insights into macro- and inter-regional scale settlement patterns (Casana 2014; Casana and Cothren 2013; Wernke, VanValkenburgh, and Saito 2020). However, such 'brute force' manual imagery survey methods are very labour intensive, time-consuming, and prone to inter-observer differences in feature detection sensitivity and specificity (Casana 2014). In part, these issues are inherent and can be attributed to the nature of the data, as archaeological features are generally very sparsely distributed across the landscape making manual identification and the labelling of archaeological features in satellite imagery a very low yield endeavour. Observational fatigue and inter-observer differences in detection rates additionally pose unavoidable risks. The resulting datasets are thus generally quite large in aerial extent but come with few labels (Mnih and Hinton 2012; Casana and Cothren 2013). Developing an effective machine learning algorithm for automating information extraction procedures on such large-scale, sparsely annotated, and unbalanced data is a long-standing machine-learning challenge in remote sensing.

In recent years, the rapid development of self-supervised contrastive learning has shown promise towards the task of utilizing large-scale, sparsely annotated data. However, the proportion of images containing archaeological settlements is often relatively low (<7% in our setting). Such an unbalanced data distribution is problematic for modern contrastive learning algorithms (Chen and Kaiming 2021; Grill et al. 2020), consequently leading to excessive favouring representations of the majority classes. By reforming contrastive learning in a semi-supervised setting, we emphasize the critical role of the sparse but valuable annotated positive instances in both training and fine-tuning stages.

In this work, we propose a novel self-supervised contrastive learning framework to identify ancient settlements through relict architectural feature detection in the south-central Andes. As opposed to existing self-supervised learning approaches, which typically model labelled and unlabelled data separately, we introduce a holistic, end-to-end semi-supervised learning framework that utilizes the highly unbalanced nature of the data to form pseudo-negative pairs by ranking the similarities between unannotated

image patches and annotated anchor images. Specifically, pseudo-negative images are employed to calculate a supervised contrastive (SupCon) loss (Khosla et al. 2020), which is seamlessly integrated with the contrastive loss (Chen et al. 2020).

To test this approach, this project surveys an approximately 4,000 km^2 region of the western cordillera of the southern Peruvian Andes (Figure 1). Utilizing images taken by Worldview 2 and Worldview 3 Satellite platforms, our dataset consists of 95,358 unlabelled images and 5,830 labelled images, where the ratio between positive and negative instances is roughly 1:100. We show that our semi-supervised contrastive learning model outperforms its self-supervised and fully-supervised counterparts, along with traditional supervised networks such as ResNet50.

There also have been recent general computer vision studies for semi-supervised contrastive learning methods (Zhang et al. 2022; Yang et al. 2022) that allow for the optimal utilization of vast amounts of unlabelled data. These approaches tend to refine the quality of pseudo labels by continuously selecting positive samples. However, the primary challenge in archaeological field lies in the extremely imbalanced proportion of positive and negative samples in the archaeological dataset (1:100). The highly imbalanced proportion of negative classes can lead to ineffective training in conventional self- and semi-supervised contrastive learning. In comparison, our proposed model offers several advantages: **(1)** Leveraging the foreground image, aligning with the class-specific few-shot learning design

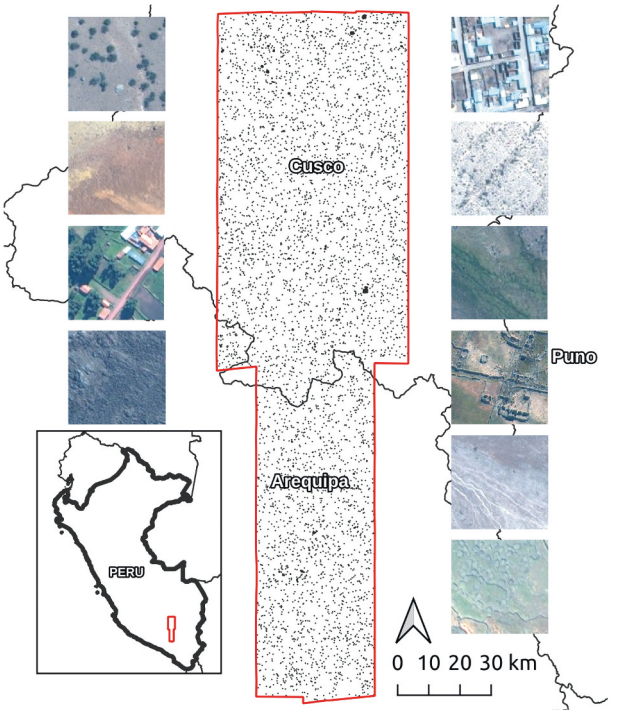

Figure 1. Survey Region. the study region encompasses approximately 4,000 km^2 of the western cordillera of the southern Peruvian highlands, including portions of the modern Cusco and Arequipa districts. Sample tiles represent the diversity of land formation in the region and the sample locations are shown in black.

for the self-supervised contrastive task. **(2)** The supervised contrastive task is balanced and further ensures the discriminative representation between the two classes in the latent space.

1.1. Innovation of the work

The innovation of this study is four-fold:

- This study investigates a large, new survey region utilizing a cutting-edge representative deep learning approach. The study region encompasses approximately 4,000 km^2 of the western cordillera of the southern Peruvian highlands, including portions of the modern Cusco and Arequipa districts.
- We propose a novel contrastive learning scheme which is optimized for the unique challenges faced in remote sensing image analyses, such as **(1)** effectively learning design from limited annotated data with large-scale unannotated data, and **(2)** the highly imbalanced data distribution (e.g., the foreground objects of interests are much less than the background).
- A new semi-supervised contrastive learning method was introduced by aggregating the advantages from previous **(1)** self-supervised and **(2)** supervised contrastive learning strategies. Compared with traditional approaches, the proposed method maximizes utilization of large-scale unlabeled image data and small-scale labeled image data under a probabilistic learning model.
- A similarity-based down-sampling approach is proposed for pseudo-label synthesis in both the latent space learning section, as well as the supervised learning section, of the semi-supervised model.

2. Background and related research

This section provides an overview of the background and related research for contrastive representation learning and satellite remote sensing. The following brief literature survey includes a summary of recent contrastive learning methods and a discussion of the applications of machine learning in remote sensing research.

2.1. Contrastive representation learning

In contrast to supervised learning (Cunningham, Cord, and Jane Delany 2008), which requires the presence of labelled inputs to predict outputs, self-supervised learning (Le 2013) refers to the identification of the hidden patterns of a dataset without the usage of any labels. Comprising what is a relatively new family of self-supervised learning methods, contrastive representation learning has recently become a key approach in solving various computer vision tasks with state-of-the-art performance (Zhirong et al. 2018; Noroozi and Favaro 2016; Zhuang, Lin Zhai, and Yamins 2019; Hjelm et al. 2018; Chuang et al. 2020; Tian et al. 2020; Khosla et al. 2020; Cui et al. 2021). Designed to learn the general features of large datasets without labels, contrastive learning aims to pull similar sample pairs

together while pushing dissimilar pairs apart. As a result, the model is capable of learning the high-level features of a dataset even with few or no labels available.

In recent years, various contrastive representation learning methods have been proposed with different implementations. SimCLR (Chen et al. 2020) aims to pull the representations of different views of the same image closer while repulsing the views of different images in the latent space. SwAV (Caron et al. 2020) applies online clustering on different augmentations of the same image instead of performing explicit pairwise feature comparisons. Wu et al. (Zhirong et al. 2018) proposes the use of an offline memory bank to store all data representations, with training data randomly selected for negative-pair minimization. Instead of utilizing an offline dictionary, MoCo (Kaiming et al. 2020) utilizes a momentum design to build a dynamic dictionary that stores a negative sample pool, which demands a large batch size. To further alleviate the cost of storing negative pairs, BYOL (Grill et al. 2020) is proposed to incorporate an asynchronous momentum encoder into the model so that it can use only the positive pairs for training. Recently, SimSiam (Chen and Kaiming 2021) has been proposed to save GPU memory consumption by fully eliminating the momentum encoder. In addition, various efforts have been made to modify the contrastive learning approach within a fully-supervised setting; an example would be the SupCon loss proposed by Khosla (Khosla et al. 2020).

2.2. Remote sensing with machine learning

Satellite remote sensing has contributed to the execution of a variety of tasks, including climate change measurement, crop condition monitoring, natural disaster alerts, and archaeological site detection (Harris 1987). Satellites were first introduced to the field of archaeology in the late 1900s, with Landsat and SPOT imagery being used for archaeological predictive modelling and archaeological feature detection (Leisz 2013). Since then, the usage of satellite remote sensing in the detection of archaeological sites has picked up rapidly, leveraging all available technologies, from decommissioned CORONA imagery (Ur 2013) to the latest in multi-spectral imagery (Abrams and Comer 2013).

Starting in the 2000s, the development of machine learning (and more particularly, representation learning) offered major breakthroughs towards the analytical approaches employed on satellite images; this in turn lead to seminal insights and discoveries in the field of archaeology (Lary et al. 2016; Camps-Valls 2009; Ali et al. 2015; Cooner, Shao, and Campbell 2016; Comer and Harrower 2013; Parcak 2019). Depending on the specific problem, various types of machine learning algorithms have been employed, such as support vector machines (SVM), decision trees, random forests, etc. (Samui 2008; Azamathulla and Fu-Chun 2011; Friedl and Brodley 1997; Pal 2005). Recently, many deep learning (and more specifically, contrastive learning-based) methods have been utilized towards remote sensing applications (Hou et al. 2021; Yue et al. 2021; Wang et al. 2022; Liu et al. 2020; Xiang et al. 2021).

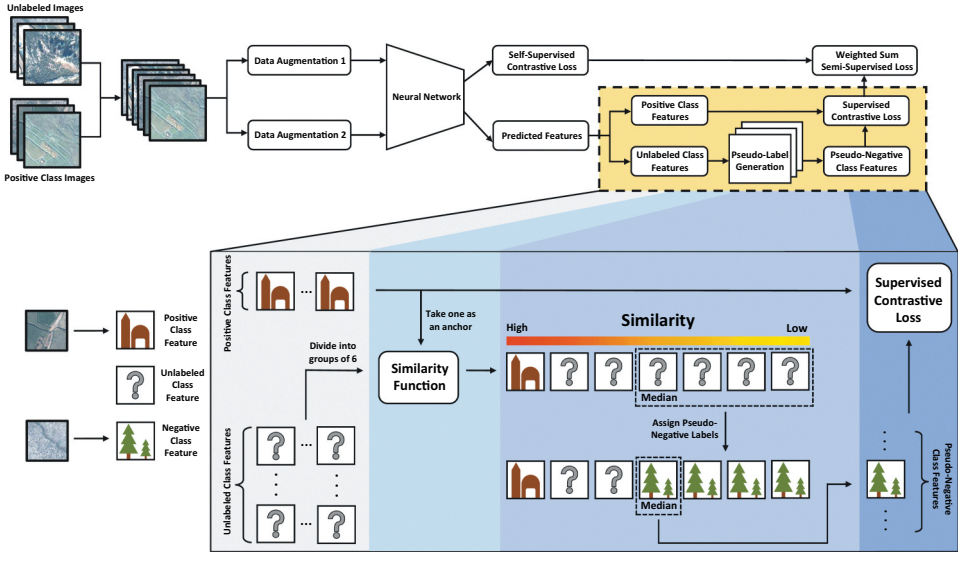

Figure 2. Overall framework. This figure demonstrates the general structure of our semi-supervised contrastive learning framework. The upper panel shows the general flow of our framework, which was adopted from the SimSiam network. The lower panel describes the process of obtaining the supervised contrastive loss from predicted features. The detailed discussions can be found in the Methods section.

3. Methods

The overall design of our framework is shown in Figure 2. An analysis of the backbone network and pseudo-label synthesis is presented below.

3.1. Overview of the SimSiam framework

In this work, SimSiam is chosen as our backbone network due to its simplicity and effectiveness. Compared to other widely-used self-supervised representation learning networks, SimSiam removes all additional structures such as negative samples (SimCLR), momentum encoder (BYOL), or clustering (SwAV) and still obtains great performance in its learning representations of unlabelled datasets. The overall loss function consists of two separate losses, namely (1) self-supervised contrastive loss (i.e. Cosine similarity) and (2) supervised contrastive loss. The rationale of employing both self-supervised and supervised loss is to form a new 'semi-supervised' learning scheme for remote sensing image learning. Compared with traditional self-supervised contrastive learning (Chen and Kaiming 2021) and supervised contrastive learning (Khosla et al. 2020) approaches, the proposed method maximizes utilization of large-scale unlabelled image data by incorporating the small-scale labelled image data. In addition, the recently introduced mixed-precision training feature is utilized to accelerate the training process.

Algorithm 1. Pseudo-Code for generating pseudo negative pairs.

Input: An array of unlabelled class features: *f_un*
Input: An array of positive class features: *f_pos*
Output: An array of pseudo-negative class features: *f_neg*
1: $f\_un = Normalize(f\_un)$
2: $f\_pos = Normalize(f\_pos)$
3: Divide *f_un* into groups of 16 $\rightarrow$ *f_un_groups*
4: Randomly choose a feature from *f_pos* $\rightarrow$ *f_positive*
5: $f\_neg = [\,]$
6: **for** *f_un_group* in *f_un_groups* **do**
7: $sim\_array = Similarity\_Function(f\_un\_group + f\_positive)$
8: $f\_neg.append(Median(sim\_array))$
9: **end for**
10: **return** *f_neg*

3.2. Pseudo-label synthesis

For the synthesis of pseudo-labels, a mixed array of unlabelled and positive class features is used as the starting point. The first step is to normalize this array. Next, the array is decoupled into X features of unlabelled images (Group 1) and Y features of positively-labelled images (Group 2). Then, we divide the X unlabelled class features into subgroups of size k (16 in our case), and we end up with X/k subgroups. Meanwhile, a single feature is randomly selected from the Y positive class features for future use.

For each subgroup, we apply the cosine similarity function in order to compute the similarity between the previously selected positive feature and the features in the subgroup. Following this, the unlabelled image with the median similarity score is assigned a pseudo label (i.e. negative class) based on the hypothesis that negative images dominate the distribution of the entire cohort. The 1:100 ratio used in this paper follows the design from a previous publication (Yang and Zhi 2020). According to (Yang and Zhi 2020), a higher imbalance ratio can impose an additional challenge towards the classification tasks as compared to a scenario with moderately imbalanced data.

The pseudo-code for generating such pseudo-labels is presented in Algorithm 1. As an example, it assumes that there are N unlabelled images. The ratio of positive images (N_{pos}) to negative images ($N - N_{pos}$) is roughly 1:100 in this study. Thus, if the size of the batch is B ($B \ll N$) and the images are randomly selected, the probability of having exactly N positive image(s) in this batch is expressed as:

$$p(n) = \frac{\binom{N_{pos}}{n} \cdot \binom{N - N_{pos}}{B - n}}{\binom{N}{B}} \tag{1}$$

Following this, the probability of having one or less positive images (in other words, B or $B - 1$ negative images) among a randomly selected batch B, is:

$$p(n \leq 1) = \sum_{n=0}^{1} \frac{\binom{N_{pos}}{n} \cdot \binom{N - N_{pos}}{B - n}}{\binom{N}{B}} = \frac{\binom{N - N_{pos}}{B} + N_{pos} \cdot \binom{N - N_{pos}}{B - 1}}{\binom{N}{B}} \approx 1 \tag{2}$$

when $N \approx N - N_{pos}$. Therefore, every batch contains almost all negative images, which ensures that the pseudo-labels are negative.

3.3. Semi-supervised contrastive learning

The key innovation of our method is our proposal of a new semi-supervised contrastive learning strategy. In short, we aggregate the standard cosine similarity loss with the supervised contrastive (SupCon) loss.

3.3.1. Self-supervised contrastive task

In each iteration of the training, X unlabelled images and Y positively-labelled images are utilized as inputs for our SimSiam model, and the generated symmetric loss is named loss_cosine. The formulas are defined as (Chen and Kaiming 2021):

$$\mathcal{D}(p_1, z_2) = -\frac{p_1}{\|p_1\|_2} \cdot \frac{z_2}{\|z_2\|_2} \tag{3}$$

$$\mathcal{L} = \frac{1}{2}\mathcal{D}(p_1, z_2) + \frac{1}{2}\mathcal{D}(p_2, z_1) \tag{4}$$

The formulas above show the SimSiam symmetrized loss for a single data point (image) x. z_1 and z_2 denote the encoding vectors of the two augmented views x_1 and x_2 (generated from x). Following, p_1 and p_2 denote the projection views of the encoding vectors z_1 and z_2 by adding an MLP head on top of the shared encoder. Equation 3 represents the negative cosine similarity. Equation 4 shows how a contrastive pair is generated through two augmented views and used for computing the cosine similarity loss.

3.3.2. Supervised contrastive task

Following the self-supervised contrastive task, the encoding features, z (that are mentioned in the self-supervised task), are employed to calculate the supervised loss, which is named **loss_super**. The encoding features will go through the steps in the Pseudo-Label Synthesis subsection, and generate X/k pseudo-negative class features and Y positive class features (mentioned earlier in the subsection). Finally, these features are combined together as inputs to the SupCon (Khosla et al. 2020) loss function in order to calculate loss_super.

The formula for the SupCon loss is presented below (Khosla et al. 2020):

$$\begin{aligned} \mathcal{L}_{\text{out}}^{sup} &= \sum_{i \in I} \mathcal{L}_{\text{out},i}^{sup} \\ &= \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_p / \tau\right)}{\sum_{a \in A(i)} \exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_a / \tau)} \end{aligned} \tag{5}$$

Equation 5 presents a generalized form of the supervised contrastive loss within a multi-view batch, where $i \in I \equiv \{1 \ \ldots \ 2N\}$ is the index of an arbitrary augmented sample (view). The " $\cdot$ " symbol denotes the dot product. The index i represents the anchor image, while $P(i)$ represents all positive pairs of the anchor. $A(i) \equiv I \backslash \{i\}$, indicates the remaining $2N - 1$ views in the batch, excluding the anchor i. As shown in Equation 5, the contrastive

nominator maximizes the similarity between the anchor and its positive pairs. Meanwhile, the contrastive denominator differentiates the anchor from negative samples.

3.3.3. Semi-supervised multi-task loss

In the final step, the total loss function is modelled as a multi-task loss design with different weighting parameters on the self-supervised (loss_cosine) and supervised (loss_-super) losses. The optimization of the weighting parameters of the loss function is inspired by Kendall (Kendall, Gal, and Cipolla 2018), in which the weighting parameters (v_1 and v_2 in Equation 6) are obtained in a data-driven manner based on training performance.

$$loss_{total} = (e^{-v_1} \cdot loss_1 + v_1) + (e^{-v_2} \cdot loss_2 + v_2) \tag{6}$$

Equation 6 is shown above, where the $loss_1$ and $loss_2$ represent the self-supervised contrastive loss and supervised contrastive loss, respectively. v_1 and v_2 denote the two weighting parameters that are automatically computed based on the training performance. Both two parameters are randomly initialized and included in the loss calculation per batch during the total loss optimization.

4. Data and experiments

The data used in the experiments are collected from WorldView 2 and WorldView 3 satellite constellations. The Data subsection below provides the descriptions of data acquisition and the preprocessing pipeline. The Experiment Design subsection describes the experimental design, including data setup, hyperparameters, and validation metrics.

4.1. Data

The satellite images used in this analysis were collected by the WorldView 2 and WorldView 3 satellite constellations and were provided by the Digital Globe Foundation following colour correction and orthographic correction using a coarse digital elevation model (DEM). The data was then pan-sharpened using the Bayesian fusion algorithm from the Orfeo-Toolbox (Grizonnet et al. 2017) so as to increase the spatial resolution of the multi-spectral imagery to 0.5 m for the WorldView 2 imagery, and 0.3 m for the WorldView 3 imagery. In this study, all spatial bands were dropped except for the Red, Green, and Blue spectral bands, and the imagery was re-sampled from 32 bits to 8 bits in order to reduce the storage size and computational requirements. In total, the images covered approximately 12,000 km^2. Finally, the study region was divided into approximately 1.6 million image tiles of size 76.8 $\times$ 76.8 metres (256 $\times$ 256 pixels at 0.3 m resolution).

Due to the semi-arid environment and limited vegetation coverage of the south-central Andes, satellites are able to capture clear and unobscured images of the ground and of the archaeological features of interest. Furthermore, ancient structures in this region were primarily constructed from stone, leading to relatively good preservation and consequently, a high visibility in satellite imagery. Of the 1.6 million image tiles produced, 5,000 of were randomly selected and manually coded for the presence/absence of archaeological buildings. To better balance sample size for the sparsely distributed modern and ancient settlements on the landscape, an additional set of 830 images

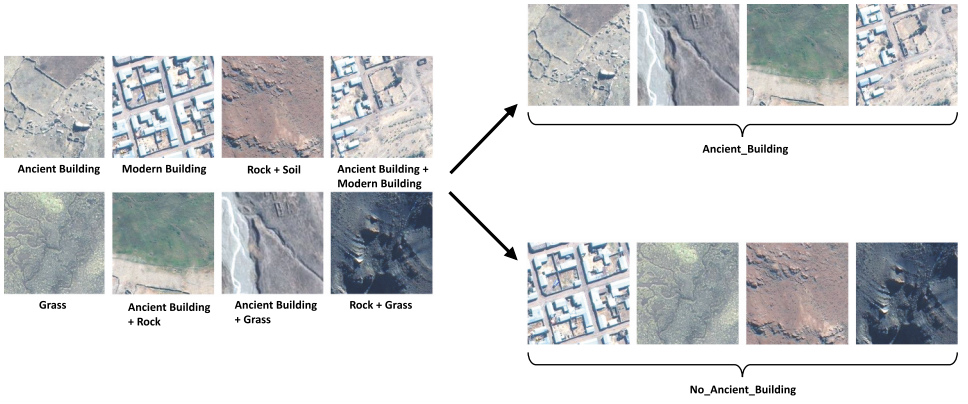

Figure 3. Example of annotated classes. This figure demonstrates example classes of the annotated data. The left panel shows the various types of unannotated images with a mixture of contents, including ancient/modern buildings, rock, soil, and grass. Due to the variety of potential objects, it's unrealistic to create a separate class for each unique combination of objects. Therefore, two classes were created based on the presence of ancient buildings, which is shown in the right panel.

known to contain examples of archaeological or modern structures was added to provide additional representation for the aforementioned categories.

Since ancient buildings were the objects of interest, all images were labelled into two classes: "ancient_building" (that is, "the presence of an archaeological structure", defined as a human-made structures less than 30 m in its largest dimension without evidence of modern roofs or maintenance) and "no_ancient_building" (that is, "no presence of an archaeological structure"). From the remaining unlabelled images, around 100,000 images were randomly selected to train the self-supervised deep learning models. Finally, those images were visually examined, and the defective ones (missing data) were abandoned, resulting in an unlabelled dataset of 95,358 images. Sample images are presented in Figure 3.

4.2. Experimental design

(1) For the Semi-supervised contrastive pre-training: the training dataset consists of 95,358 unlabelled images and 258 labelled foreground images. (2) For the supervised downstream classification fine-tuning task: The 5,830 labelled images were divided into training, validation, and testing splits, ensuring that images from the nearby physical space were placed into the same split in order to avoid the issue of data contamination. Additionally, in order to alleviate the unbalanced nature of our data source, the under-representative positive class (with ancient builds) in our training dataset was up-sampled to have the roughly an equal distribution to the negative class. The details of the data split are shown in Table 1. As a last step, all of the labelled and unlabelled images were resized to 128×128 so as to expedite the training process.

4.2.1. Semi-supervised contrastive training

The proposed semi-supervised contrastive learning model was adopted from the SimSiam network with major modifications on the loss function. The optimizer used was the SGD

Table 1. Datasets Setup.

Dataset	# of Ancient_Building	# of No_Ancient_Building
Training	193 (Original) 3,088 (Upsampled)	4,272
Validation	65	610
Testing	71	619
Unannotated	95,358	

optimizer which was initialized with a learning rate = 0.1, weight decay = 0.0001, and momentum = 0.9. The training dataset of the unlabelled images had a batch size = 512, while the labelled dataset had a batch size = 16. The Mixed-precision training features were integrated into our network in order to boost the training process.

The model was trained for 200 epochs with approximately 100 training hours. The 100 training hours were computed by a workstation with Intel Xeon Gold 5118 2.30 GHz CPU, 383 GB memory, and two NVIDIA GeForce RTX 2080 Ti GPU (11.0 GB dedicated GPU memory).

4.2.2. Supervised fine-tuning and testing

After pre-training the model using the unannotated data, an additional single linear layer was fined tuned with the labelled data. The F1 score and the balanced accuracy (Wegier and Ksieniewicz 2020; Feng, Zhou, and Tong 2021) on the validation set were the metrics used to select the best performance epoch as well as the optimal hyper-parameters.

4.2.3. Evaluation metrics

According to (Wegier and Ksieniewicz 2020), the F1 score aggregates the sensitivity and precision,

$$F1\ score = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity} \tag{7}$$

where the sensitivity (or recall) determines the accuracy of the minority class classification and precision indicates the probability of correct detection.

Balanced accuracy is the arithmetic mean of the sensitivity and specificity,

$$Balanced\ accuracy = \frac{Specificity + Sensitivity}{2} \tag{8}$$

The specificity, in a binary case, indicates the accuracy of recognizing the negative (majority) class.

5. Results

Extensive experiments are then designed to verify the effectiveness of our proposed model. Thus, a performance comparison with other state-of-art methods is presented for an ablation study.

Table 2. Quantitative results of different learning methods. SL and SSL correspond to Supervised Learning and Self-Supervised Learning, respectively. CE is short for Cross Entropy.

	Model	Loss Function	Balanced Accuracy	F1 Score
SL	ResNet 50[a]	Supervised CE Loss (Zhang and Sabuncu 2018)	0.701	0.612
	ResNet 50[b]	Supervised CE Loss (Zhang and Sabuncu 2018)	0.790	0.718
SSL	SimCLR	Self-Supervised Loss (Chen and Kaiming 2021)	0.766	0.554
		Semi-Supervised Loss (Ours)	0.769	0.613
	BYOL	Self-Supervised Loss (Grill et al. 2020)	0.731	0.681
		Semi-Supervised Loss (Ours)	0.757	0.696
	SimSiam	Self-Supervised Loss Only (Chen and He 2021)	0.752	0.744
		Supervised Loss Only (Khosla et al. 2020)	0.500	0.477
		Semi-Supervised Loss (Ours)	0.790	0.762

[a]This corresponds to model: ResNet 50 (from scratch).
[b]This corresponds to model: ResNet 50 (ImageNet pretrained).

5.1. Ablation study

This ablation study consisted of two alternative versions of our proposed pre-trained model. The first version conducted self-supervised contrastive learning using only loss_cosine as discussed in the Methods section, inspired by the SimSiam network (Chen et al. 2020). By contrast, the second version conducted supervised contrastive learning using only loss_super as discussed in the Methods section. The corresponding testing results on the downstream labelled data have been presented in Table 2 as SSL.

5.2. Comparison with Fully-Supervised Learning Benchmark

In addition to the contrastive learning framework discussed above, a fully supervised version of the experiment that was trained from scratch and used only the labelled images was also established, so as to further demonstrate the performance boost that is provided by the unlabelled data. Since our self-supervised framework employed ResNet50 as the backbone model, it's also used here for the canonical fully-supervised learning benchmarks. An SGD optimizer and cross entropy loss were employed to create a standard training environment. The model was trained for 16 epochs. The trained model with the best validation performance was selected to run the testing dataset. The results of investigating canonical fully-supervised benchmarks are demonstrated in Table 2 as SL.

5.3. Experiments on Additional Contrastive Learning Frameworks

In order to further illustrate the effectiveness of our network design, two additional contrastive learning frameworks – BYOL and SimCLR – were utilized and modified to also incorporate the semi-supervised loss. The corresponding F1 score and balanced accuracy are indicated in Table 2. We then evaluated the accuracy of positive and negative images in this highly unbalanced scenario. Our semi-supervised loss mechanism yields an accuracy of 0.803 for positive images and 0.734 for negative images when using the SimCLR backbone. It suggested that our semi-supervised loss mechanism had a balanced performance for both positive and negative cases.

6. Discussion

In this study, we developed a new, semi-supervised contrastive learning pseudo-label generation method based on the similarity matrix; in doing so, our ultimate goal was to enhance self-supervised training performance over a highly unbalanced dataset. By integrating self-supervised and supervised loss functions, we designed a new learning framework that simultaneously learns from both unannotated and annotated data.

The conducted experiments have yielded promising results. In the ablation study, the models trained using only loss_super were relatively ineffective in identifying ancient buildings, while the models trained using only **loss_cosine** produced competitive F1 scores and balanced accuracies; this indicated that having a model pre-trained on unlabelled data using self-supervised contrastive learning was essential in distinguishing ancient buildings from other objects. Nonetheless, our proposed semi-supervised model that was trained with loss_cosine + loss_super outperformed its self-supervised and fully-supervised counterparts by a decent margin. This result showed that the integration of fully-supervised and self-supervised networks was a complimentary aggregation. Furthermore, when comparing fully-supervised learning benchmarks (such as ResNet50) with our model, our solution exhibited superior performance on the downstream labelled dataset.

Figure 4 offered additional insights on the performance of our framework. The upper left and lower right corners indicate the correct classes prediction, while the upper right and lower left are the examples of false positives and false negatives. From the ablation study, the superior performance of our model can be attributed to the dynamic combination of fully-supervised and self-supervised information. The loss generated from the positive and pseudo-negative images serves as a complement to the other loss functions that were generated solely on unlabelled instances.

We believe there are several potential improvements for our semi-supervised contrastive learning framework. First, our proposed pseudo-labelling strategy would

Figure 4. Testing Sample Results. This figure presents representative samples from the testing results. The left panel indicates the true positive examples while the right panel indicates the true negative cases. Likewise, the upper row indicates the predicted positive ones while the lower row indicates the predicted negative ones.

only work for highly unbalanced data sets. Moreover, the current model has not been extended to multi-label classification scenarios. Meanwhile, the size of annotated training images in our study was still relatively small. To further facilitate the performance, we might need more training data especially with more positive images (ancient building) from our predicted positive class.

7. Conclusion

In this project, we proposed a new semi-supervised contrastive learning method for identifying relict architectural features in the south-central Andes from satellite imagery. As opposed to the existing solutions, we utilized the unbalanced nature of the large-scale unlabelled data to form pseudo-negative pairs. Using such negative pairs, we leveraged the contrastive learning method by introducing a holistic learning scheme with both cosine similarity loss and pseudo-supervision loss functions. According to the experimental results, our proposed framework yielded both superior accuracy and a higher F1 score as compared with its self-supervised and fully-supervised counterparts. The integrated model eventually outperformed traditional supervised networks (e.g. ResNet50) by 15% in F1 score and 8.9% in balanced accuracy.

These improved feature detection results show great promise for developing a machine-human teaming approach, in which human surveyors would not need to visually scan vast featureless areas, and instead could focus their efforts on categorizing, annotating, and enriching the attribute data on autonomously-identified features. This approach would also eliminate inter-observer differences in feature detection sensitivity and specificity, while also enabling greater transparency and reproducibility through the reporting of model parameters.

Acknowledgements

This work is supported by Scaling Success Grant from Vanderbilt University. The imagery analyzed in this paper was provided by the Digital Globe Foundation through a generous satellite imagery grant (S. Wernke, P.I.). Computational resources were provided by the Vanderbilt University Spatial Analysis Research Laboratory (https://wernkelab.org/). This work has not been submitted for publication or presentation elsewhere.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The work was supported by the DigitalGlobe Foundation .

ORCID

Jiachen Xu http://orcid.org/0000-0001-8869-2045

Yuankai Huo http://orcid.org/0000-0002-2096-8065

Data Availability Statement

The data that support the findings of this study are available from the author, Steven A. Wernke, upon reasonable request. Due to restrictions imposed by the data provider, we are unable to provide a URL for the data used in this study.

References

Abrams, M. J., and D. C. Comer. 2013. "Multispectral and Hyperspectral Technology and Archaeological Applications." In *Mapping Archaeological Landscapes from Space*. Springer Science & Business Media.

Alcock, S., and J. Cherry. 2016. *Side-By-Side Survey: Comparative Regional Studies in the Mediterranean World*. Oxbow Books.

Ali, I., F. Greifeneder, J. Stamenkovic, M. Neumann, and C. Notarnicola. 2015. "Review of Machine Learning Approaches for Biomass and Soil Moisture Retrievals from Remote Sensing Data." *Remote Sensing* 7 (12): 16398–16421. doi:10.3390/rs71215841.

Azamathulla, H. M., and W. Fu-Chun. 2011. "Support Vector Machine Approach for Longitudinal Dispersion Coefficients in Natural Streams." *Applied Soft Computing* 11 (2): 2902–2905. doi:10.1016/j.asoc.2010.11.026.

Balkansky, A. K., S. A. Kowalewski, V. Pérez Rodrguez, T. J. Pluckhahn, C. A. Smith, L. R. Stiver, D. Beliaev, J. F. Chamblee, V. Y. Heredia Espinoza, and R. Santos Pérez. 2000. "Archaeological Survey in the Mixteca Alta of Oaxaca, Mexico." *Journal of Field Archaeology* 27 (4): 365–389. doi:10.1179/jfa.2000.27.4.365.

Banning, E. B. 2002. *Archaeological Survey*. Springer.

Bewley, R., A. Wilson, D. Kennedy, D. Mattingly, R. Banks, M. Bishop, J. Bradbury, et al. 2016. "Endangered Archaeology in the Middle East and North Africa: Introducing the EAMENA Project." In *CAA2015. Keep the revolution going: Proceedings of the 43rd annual conference on computer applications and quantitative methods in archaeology*, Newcastle University.

Camps-Valls, G. 2009. "Machine Learning in Remote Sensing Data Processing." In *2009 IEEE international workshop on machine learning for signal processing*, 1–6. IEEE.

Caron, M., I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin. (2020). "Unsupervised Learning of Visual Features by Contrasting Cluster Assignments." *arXiv preprint arXiv:2006.09882*.

Casana, J. 2014. "Regional-Scale Archaeological Remote Sensing in the Age of Big Data: Automated Site Discovery Vs. Brute Force Methods." *Advances in Archaeological Practice* 2 (3): 222–233. doi:10.7183/2326-3768.2.3.222.

Casana, J., and J. Cothren. 2013. "The CORONA Atlas Project: Orthorectification of CORONA Satellite Imagery and Regional-Scale Archaeological Exploration in the Near East." In *Mapping Archaeological Landscapes from Space*, 33–43. Springer.

Chen, X., and H. Kaiming 2021. "Exploring Simple Siamese Representation Learning." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15750–15758.

Chen, T., S. Kornblith, M. Norouzi, and G. Hinton. 2020. "A Simple Framework for Contrastive Learning of Visual Representations." In *International conference on machine learning*, 1597–1607. PMLR.

Chuang, C.Y., J. Robinson, L. Yen-Chen, A. Torralba, and S. Jegelka. 2020. "Debiased Contrastive Learning." *arXiv preprint arXiv:2007.00224*.

Comer, D. C., and M. J. Harrower. 2013. "Mapping Archaeological Landscapes from Space." In *SpringerBriefs in Archaeological Heritage Management*, Google-Books-ID: yWJDAAAAQBAJ New York: Springer Science & Business Media. https://www.springer.com/gp/book/9781461460732

Cooner, A. J., Y. Shao, and J. B. Campbell. 2016. "Detection of Urban Damage Using Remote Sensing and Machine Learning Algorithms: Revisiting the 2010 Haiti Earthquake." *Remote Sensing* 8 (10): 868. doi:10.3390/rs8100868.

Cui, J., Z. Zhong, S. Liu, Y. Bei, and J. Jia. 2021. "Parametric Contrastive Learning." *arXiv preprint arXiv:2107.12028*.

Cunningham, P., M. Cord, and S. Jane Delany. 2008. "Supervised Learning." In *Machine Learning Techniques for Multimedia*, 21–49. Springer.

Feng, Y., M. Zhou, and X. Tong. 2021. "Imbalanced Classification: A Paradigm-Based Review." *Statistical Analysis and Data Mining: The ASA Data Science Journal* 14 (5): 383–406. doi:10.1002/sam.11538.

Fowler, M. J. 2002. "Satellite Remote Sensing and Archaeology: A Comparative Study of Satellite Imagery of the Environs of Figsbury Ring, Wiltshire." *Archaeological Prospection* 9 (2): 55–69. doi:10.1002/arp.181.

Friedl, M. A., and C. E. Brodley. 1997. "Decision Tree Classification of Land Cover from Remotely Sensed Data." *Remote Sensing of Environment* 61 (3): 399–409. doi:10.1016/S0034-4257(97)00049-7.

Grill, J.B., F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, et al. 2020. "Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning." *arXiv preprint arXiv:2006.07733*.

Grizonnet, M., J. Michel, V. Poughon, J. Inglada, M. Savinaud, and R. Cresson. 2017. "Orfeo ToolBox: Open Source Processing of Remote Sensing Images." *Open Geospatial Data, Software and Standards* 2 (1): 15. doi:10.1186/s40965-017-0031-6.

Hanson, W. S., and I. A. Oltean. 2012. *Archaeology from Historical Aerial and Satellite Archives*. Springer Science & Business Media.

Harris, R. 1987. "Satellite Remote Sensing. An Introduction."

Hjelm, R. D., A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio. 2018. "Learning Deep Representations by Mutual Information Estimation and Maximization." *arXiv preprint arXiv:1808.06670*.

Hou, S., H. Shi, X. Cao, X. Zhang, and L. Jiao. 2021. "Hyperspectral Imagery Classification Based on Contrastive Learning." *IEEE Transactions on Geoscience and Remote Sensing* 60: 1–13. doi:10.1109/TGRS.2021.3139099.

Kaiming, H., H. Fan, W. Yuxin, S. Xie, and R. Girshick. 2020. "Momentum Contrast for Unsupervised Visual Representation Learning." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.

Kendall, A., Y. Gal, and R. Cipolla. 2018. "Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7482–7491.

Khosla, P., P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. 2020. "Supervised Contrastive Learning." *arXiv preprint arXiv:2004.11362*.

Lary, D. J., A. H. Alavi, A. H. Gandomi, and A. L. Walker. 2016. "Machine Learning in Geosciences and Remote Sensing." *Geoscience Frontiers* 7 (1): 3–10. doi:10.1016/j.gsf.2015.07.003.

Lasaponara, R., and N. Masini. 2012. *Satellite Remote Sensing: A New Tool for Archaeology*. Vol. 16. Springer Science & Business Media.

Le, Q. V. 2013. "Building High-Level Features Using Large Scale Unsupervised Learning." In *2013 IEEE international conference on acoustics, speech and signal processing*, 8595–8598. IEEE.

Leisz, S. J. 2013. "An Overview of the Application of Remote Sensing to Archaeology During the Twentieth Century." In *Mapping Archaeological Landscapes from Space*, edited by D. C. Comer and M. J. Harrower, 11–19. New York, NY: Springer New York.

Lin, A.Y.M., A. Huynh, G. Lanckriet, L. Barrington, and M. D. Petraglia. 2014. "Crowdsourcing the Unknown: The Satellite Search for Genghis Khan." *Plos One* 9 (12): e114046. doi:10.1371/journal.pone.0114046.

Liu, B., Y. Anzhu, Y. Xuchu, R. Wang, K. Gao, and W. Guo. 2020. "Deep Multiview Learning for Hyperspectral Image Classification." *IEEE Transactions on Geoscience and Remote Sensing* 59 (9): 7758–7772. doi:10.1109/TGRS.2020.3034133.

Mnih, V., and G. E. Hinton. 2012. "Learning to Label Aerial Images from Noisy Data." In *Proceedings of the 29th International conference on machine learning (ICML-12)*, 567–574.
Noroozi, M., and P. Favaro. 2016. "Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles." In *European conference on computer vision*, 69–84. Springer.
Pal, M. 2005. "Random Forest Classifier for Remote Sensing Classification." *International Journal of Remote Sensing* 26 (1): 217–222. doi:10.1080/01431160412331269698.
Parcak, S. H. 2019. *Archaeology from Space: How the Future Shapes Our Past*. Illustrated edition ed. Henry Holt and Co.
Phillips, P., and G. R. Willey. 1953. "Method and Theory in American Archeology: An Operational Basis for Culture-Historical Integration." *American Anthropologist* 55 (5): 615–633. doi:10.1525/aa.1953.55.5.02a00030.
Samui, P. 2008. "Support Vector Machine Applied to Settlement of Shallow Foundations on Cohesionless Soils." *Computers and Geotechnics* 35 (3): 419–427. doi:10.1016/j.compgeo.2007.06.014.
Sanders, W. T. 1961. "A Developmental Concept of Pre-Spanish Urbanization in the Valley of Mexico." *American Antiquity* 27 (2): 259–260. William J. Mayer-Oakes. Middle American Research Records, Vol. II, No. 8 (preprinted from Pub. No. 18, pp. 165-76), Middle American Research Institute, Tulane University, New Orleans, 1960. 9 pp., 2 charts. doi:10.2307/277859.
Tian, Y., C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola. 2020. "What Makes for Good Views for Contrastive Learning?" *arXiv preprint arXiv:2005.10243*.
Ur, J. A. 2013. "CORONA Satellite Imagery and Ancient Near Eastern Landscapes." In *Mapping Archaeological Landscapes from Space*, 21–31. Springer.
Wang, Y., C. M. Albrecht, N. Ait Ali Braham, L. Mou, and X. Xiang Zhu. 2022. "Self-Supervised Learning in Remote Sensing: A Review." *arXiv preprint arXiv:2206.13188*.
Wegier, W., and P. Ksieniewicz. 2020. "Application of Imbalanced Data Classification Quality Metrics as Weighting Methods of the Ensemble Data Stream Classification Algorithms." *Entropy* 22 (8): 849. doi:10.3390/e22080849.
Wernke, S., P. VanValkenburgh, and A. Saito. 2020. "Interregional Archaeology in the Age of Big Data: Building Online Collaborative Platforms for Virtual Survey in the Andes." *Journal of Field Archaeology* 45 (sup1): S61–74. doi:10.1080/00934690.2020.1713286.
Xiang, H., L. Teng, T. Zhou, Y. Liu, and Y. Peng. 2021. "Contrastive Learning Based on Transformer for Hyperspectral Image Classification." *Applied Sciences* 11 (18): 8670. doi:10.3390/app11188670.
Yang, F., W. Kai, S. Zhang, G. Jiang, Y. Liu, F. Zheng, W. Zhang, C. Wang, and L. Zeng. 2022. "Class-Aware Contrastive Semi-Supervised Learning." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14421–14430.
Yang, Y., and X. Zhi. 2020. "Rethinking the Value of Labels for Improving Class-Imbalanced Learning." *Advances in Neural Information Processing Systems* 33: 19290–19301.
Yue, J., L. Fang, H. Rahmani, and P. Ghamisi. 2021. "Self-Supervised Learning with Adaptive Distillation for Hyperspectral Image Classification." *IEEE Transactions on Geoscience and Remote Sensing* 60: 1–13. doi:10.1109/TGRS.2021.3057768.
Zhang, Z., and M. Sabuncu. 2018. "Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels." *Advances in Neural Information Processing Systems* 31.
Zhang, Y., X. Zhang, L. Jie, R. Qiu, X. Haohang, and Q. Tian. 2022. "Semi-Supervised Contrastive Learning with Similarity Co-Calibration." *IEEE Transactions on Multimedia*.
Zhirong, W., Y. Xiong, S. X. Yu, and D. Lin. 2018. "Unsupervised Feature Learning via Non-Parametric Instance Discrimination." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3733–3742.
Zhuang, C., A. Lin Zhai, and D. Yamins. 2019. "Local Aggregation for Unsupervised Learning of Visual Embeddings." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6002–6012.