DENSE SUPERVISED CONTRASTIVE LEARNING FOR FEW-SHOT REMOTE SENSING SCENE CLASSIFICATION

GE LIU, XIANGZHONG FANG

Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China
E-MAIL: liu.ge@sjtu.edu.cn, xzfang@@sjtu.edu.cn

Abstract:

Few-shot remote sensing scene classification (FSRSSC) tackles the challenge of recognizing novel scene categories with only a limited number of labeled examples, which heavily relies on pre-trained transferable deep representations. Recent advances have widely explored contrastive learning to improve global feature representations for this task. However, we argue that global feature representations often fail to capture the fine-grained local features crucial for distinguishing remote sensing scenes, which typically include numerous small-scale, densely distributed objects. To address this limitation, we propose Dense Supervised Contrastive Learning (DSCL), which applies supervised contrastive learning at the patch level to improve local feature discriminability. By optimizing a dense pairwise similarity loss across local patches, DSCL significantly boosts generalization in few-shot scenarios. Experiments on three benchmark datasets (UCM, WHU-RS19, and NWPU-RESISC45) demonstrate that DSCL achieves competitive performance compared to recent state-of-the-art methods in both 5-way 1-shot and 5-shot settings, highlighting the potential of local-level feature learning for FSRSSC.

Keywords:

Few-shot learning; Representation learning; Remote sensing scene classification

1 Introduction

Scene classification is a fundamental task in remote sensing image analysis [?], which aims to categorize remote sensing images into predefined semantic classes based on their visual content. Remote sensing scene classification (RSSC) has a wide range of applications, including land use mapping [?], environmental monitoring, and disaster assessment.

Traditional RSSC methods rely on hand-crafted features such as Histograms of Oriented Gradients (HOG) [?]. With the development of deep learning, convolutional neural networks (CNNs) [?] have significantly improved RSSC performance [?]. However, training deep CNNs requires a substantial amount of labeled data, and obtaining enough annotated remote sensing data is frequently impractical due to the high costs associated with data acquisition and labeling.

This limitation has led to increasing interest in few-shot remote sensing scene classification (FSRSSC), which aims to recognize novel scene categories from only a few labeled examples. Training CNNs with only limited samples is prone to overfitting. Current FSRSSC methods adopt the transfer learning paradigm [?], typically including two stages: (1) meta-training, where models learn transferable knowledge from base classes, and (2) meta-testing, where the learned knowledge is adapted to recognize novel classes using a small support set. The effectiveness of this paradigm heavily depends on the transferability of the learned representations [?]. Moreover, FSRSSC encounters additional challenges beyond general few-shot learning (FSL) due to the unique characteristics of remote sensing imagery, such as high intra-class variation and inter-class similarity [?], which stem from overhead imaging perspectives and the complex coexistence of multiple ground objects.

Recent state-of-the-art methods have widely adopted contrastive learning to enhance feature transferability for FSRSSC. For instance, SCL-MLNet [?] incorporates self-supervised contrastive learning as an auxiliary task to improve representation quality. TSC [?] employs a task-specific contrastive loss within each episode to strengthen meta-learning. Foreground-background contrastive learning [?] explicitly separates foreground and background re-

gions to enhance feature discrimination. MPCL-Net [?] proposes a multi-pretext prototype-guided contrastive framework that jointly optimizes multiple pretext tasks for representation learning. DCN [?] proposes dual-branch supervised contrastive learning to capture both contextual and fine-grained features. ACL-Net [?] integrates a mutual attention mechanism with a dictionary-based contrastive loss to improve feature representation learning.

Figure 1. Many remote sensing scene images are made up of dense discriminative objects

Despite recent advances, most methods rely on global features, which often overlook the fine-grained local details crucial for distinguishing remote sensing scenes. Unlike natural images, remote sensing imagery typically contains many small, densely packed objects tied to scene semantics. As shown in Figure 1, these localized cues play a key role in accurate classification. We argue that improving local feature discriminability is beneficial to FSRSSC. To this end, we propose Dense Supervised Contrastive Learning (DSCL), a novel framework specifically designed to learn effective representations for remote sensing imagery characterized by densely distributed objects. DSCL builds on the supervised contrastive learning paradigm [?] by optimizing a patch-wise contrastive loss over all local features in the augmented batch. It constructs positive pairs between local features of the same class, while all patches from different classes are regarded as negatives. This formulation encourages representations of semantically similar regions to cluster together while separating dissimilar ones, thereby enhancing local feature discrimination. During meta-testing, we train a linear classifier on the aggregated dense features from the support set, and obtain the final prediction for each query sample by ensembling its patch-wise logits.

Our contributions are summarized as follows: (1) We propose DSCL, a dense supervised contrastive learning framework that improves local feature discrimination by a patch-level contrastive loss, specifically designed for FSRSSC. (2) We conduct extensive experiments on three benchmark datasets, demonstrating that DSCL achieves competitive performance compared to recent state-of-the-art FSRSSC methods.

2 Method

2.1 Task Formulation

Few-shot remote sensing scene classification (FSRSSC) aims to generalize knowledge learned from base classes to recognize novel scene categories in a previously unseen domain. This problem is commonly formulated from the perspective of inductive transfer learning and consists of two primary stages: meta-training and meta-testing. Formally, the meta-training dataset is denoted as $\mathcal{D}_b = \{(\mathbf{x}, y)\} \subset \mathcal{X}_b \times \mathcal{Y}_b$, where $\mathbf{x}$ represents a remote sensing image and y denotes its corresponding scene class label. FSL algorithms aim to extract generalizable and transferable knowledge from the $\mathcal{D}_b$ based on deep neural networks. In the meta-testing stage, the pre-trained model is adapted to a novel few-shot classification task using a small support set $\mathcal{S}$, which is sampled from a target domain $\mathcal{D}_n = \{(\mathbf{x}, y)\} \subset \mathcal{X}_n \times \mathcal{Y}_n$. The support set $\mathcal{S}$ contains N distinct classes, each with K labeled examples, i.e., $\mathcal{S} = \{S_i\}_{i=1}^{N}$ with $|S_i| = K$, and this setting is referred to as an "N-way K-shot" classification task. Notably, there is no class overlap between the meta-training and meta-testing sets, i.e., $\mathcal{Y}_b \cap \mathcal{Y}_n = \emptyset$. Once the model has been adapted to the support set, its performance is evaluated on a query set $\mathcal{Q}$, which contains unseen samples from the same novel classes as in $\mathcal{S}$. The goal is to evaluate the generalization ability of the model to novel categories using only a few labeled examples.

Conventional approaches to FSRSSC mainly focus on learning general-purpose features. In line with this, we aim to train a feature extractor F, parameterized by θ, using an enhanced contrastive learning method. Given an input image $\mathbf{x}$, the model produces a feature representation $F(\mathbf{x})$ that is both semantically meaningful and transferable. A desirable representation in this context enables robust classification of novel remote sensing scene categories, especially under few-shot settings.

2.2 Supervised Contrastive Learning Revisit

Contrastive learning is a widely adopted method for representation learning and has proven effective in enhancing the feature extraction capabilities of deep neural networks. The core idea is to learn an encoder that maps similar (positive) pairs close together in the embedding space while

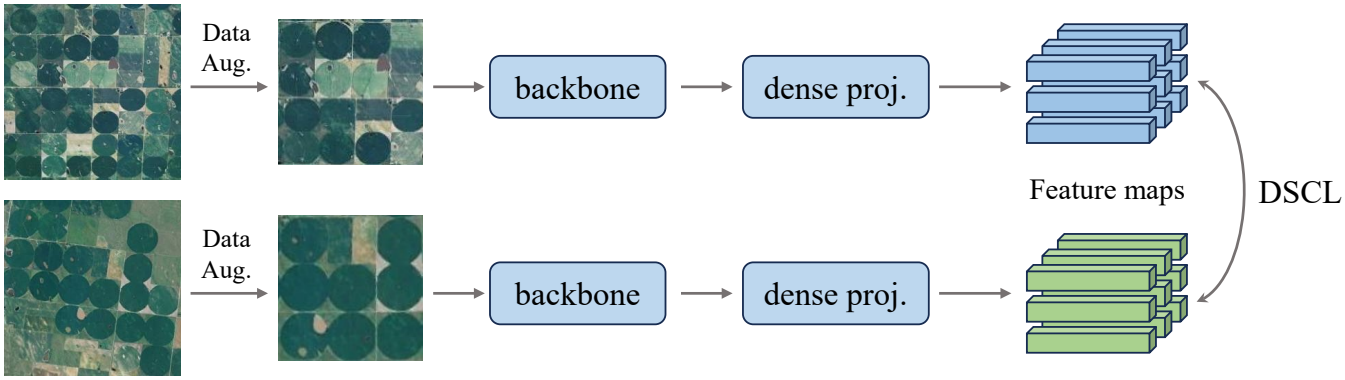

Figure 2. Framwork of Dense Supervised Contrastive Learning

pushing dissimilar (negative) pairs farther apart. While most prior works focus on self-supervised contrastive learning, which learns from unlabeled data through instance discrimination [?,?], supervised contrastive learning (SCL) [?] leverages label information to encourage representations of samples from the same class to be closer together, while pushing apart those from different classes. We briefly revisit the supervised contrastive learning framework, which we later extend for dense scene representation learning.

Given a batch of labeled training data $\{\mathbf{x}_i, y_i\}_{i=1}^{N}$, we apply data augmentation to generate two views per sample, resulting in a multiviewed batch $\{\tilde{\mathbf{x}}_j, \tilde{y}_j\}_{j=1}^{2N}$, where $\{\tilde{\mathbf{x}}_{2i-1}, \tilde{\mathbf{x}}_{2i}\}$ are two random augmentations of $\mathbf{x}_i$, and $\tilde{y}_{2i-1} = \tilde{y}_{2i} = y_i$. Each augmented image $\tilde{\mathbf{x}}$ is passed through an encoder network $F(\cdot)$ to obtain a feature representation $\boldsymbol{r} = F(\tilde{\mathbf{x}}) \in \mathbb{R}^D$. This representation is further mapped by a projection head $Proj(\cdot)$, implemented as a two-layer MLP with a ReLU activation in between, producing the embedding $\boldsymbol{z} = Proj(\boldsymbol{r}) \in \mathbb{R}^D$ for contrastive learning. To compute the contrastive loss, we define $I \equiv \{1, \cdots, 2N\}$ as the index set of the augmented batch. For each anchor sample $i \in I$, the positive set $\mathcal{P}(i)$ includes all other samples with the same label ($\mathcal{P}(i) \equiv \{p \in I : y_p = y_i\} \setminus \{i\}$), and the negative set $\mathcal{N}(i)$ includes those with different labels. The supervised contrastive loss is defined as:

$$\mathcal{L}_{sup} = \frac{1}{|I|} \sum_{i \in I} \frac{1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} -\log \hat{P}_i \,, \tag{1}$$

$$\hat{P}_i = \frac{\exp\left(\cos(z_i, z_p)/\tau\right)}{\exp\left(\cos(z_i, z_p)/\tau\right) + \sum\limits_{a \in \mathcal{N}(i)} \exp\left(\cos(z_i, z_a)/\tau\right)}, \tag{2}$$

where $\cos(\cdot, \cdot)$ denotes the cosine similarity between two vectors, and τ is a temperature scaling parameter.

After training, the encoder $F(\cdot)$ is retained for downstream tasks, while the projection head $Proj(\cdot)$ is discarded. In this paper, we aim to expand the supervised contrastive learning framework to develop dense representations that are more appropriate for remote sensing scene classification.

2.3 Dense Supervised Contrastive Learning

While conventional supervised contrastive learning focuses on global image-level features, it falls short in remote sensing scenarios where critical discriminative cues often reside in small-scale, densely packed local regions. To better capture such local semantics, we propose Dense Supervised Contrastive Learning (DSCL), which extends supervised contrastive learning to the patch (or spatial) level, enabling fine-grained representation learning for FSRSSC, as illustrated in Figure 2.

Given an input image $\mathbf{x}$, similar to global SCL, random data augmentation is first applied to obtain two views $\tilde{\mathbf{x}}_1$ and $\tilde{\mathbf{x}}_2$. Each augmented view is passed through a shared feature extractor $F(\cdot)$, yielding dense feature maps $\mathbf{R}_1 = F(\tilde{\mathbf{x}}_1) \in \mathbb{R}^{H \times W \times D}$ and $\mathbf{R}_2 = F(\tilde{\mathbf{x}}_2) \in \mathbb{R}^{H \times W \times D}$, where H and W denote the spatial dimensions and D is the channel dimension. We treat each spatial location (patch) in the two feature maps as an individual unit of the same class for dense contrastive learning. To this end, we apply a spatially shared projection head $Proj(\cdot)$ to each embedding vector $\mathbf{r}_{ij} \in \mathbb{R}^D$ at location (i, j), producing projected vectors $\mathbf{z}_{ij} = Proj(\mathbf{r}_{ij}) \in \mathbb{R}^D$.

Given a batch of N images and their augmented views, we collect all spatial patch-level vectors across all views

and construct positive pairs between corresponding spatial locations of the same class, while treating all patches from different classes as negatives. To compute the contrastive loss, we define $U \equiv \{1, \cdots, 2N*H*W\}$ as the index set of all local embedding. Formally, for a local patch embedding $\mathbf{z}_u$ with label y_u, we define its positive set $\mathcal{P}(u) = \{v \in U : y_v = y_u \wedge v \neq u\}$ and its negative set $\mathcal{N}(u) = \{v \in U : y_v \neq y_u\}$. The dense supervised contrastive loss is computed as:

$$\mathcal{L}_{dense} = \frac{1}{|U|} \sum_{u \in U} \frac{1}{|\mathcal{P}(u)|} \sum_{v \in \mathcal{P}(u)} -\log \hat{P}_u \,, \tag{3}$$

$$\hat{P}_u = \frac{\exp\left(\cos(z_u, z_v)/\tau\right)}{\exp\left(\cos(z_u, z_v)/\tau\right) + \sum\limits_{w \in \mathcal{N}(u)} \exp\left(\cos(z_u, z_w)/\tau\right)}, \tag{4}$$

where U indexes all spatial positions across the augmented batch, and τ is the temperature hyperparameter as in Eq. (2).

This dense contrastive formulation encourages the network to learn discriminative and invariant local representations that generalize better in few-shot settings, particularly for remote sensing imagery with dense discriminative objects.

During meta-testing, we discard the projection head and use the trained backbone $F(\cdot)$ to extract dense features for both support and query images. The dense features aggregated from the support set are used to train a linear classifier. The final prediction for each query image is obtained by ensembling its patch-wise logits.

3 Experimental Results

3.1 Benchmark Datasets

Following recent studies, we evaluate our method on three widely used remote sensing scene classification benchmark datasets: UC Merced LandUse (UCM) [?], WHU-RS19 (WHURS [?], and NWPU-RESISC45 (NWPU) [?]. The class splits for few-shot learning are as follows:

UCM [?] contains 2,100 images across 21 scene categories. Following prior work [?, ?], we use 10 classes for training, 6 classes for validation, and 5 classes for testing.

WHURS [?] includes 1,005 images from 19 categories. Following prior work [?], we use 9 classes for training, 5 classes for validation, and 5 classes for testing.

NWPU [?] consists of 31,500 images covering 45 categories, with 700 images per class. Following [?, ?], we use 25 classes for training, 10 classes for validation, and 10 classes for testing.

3.2 Implementation Details

Model Structure. Following prior methods [?, ?, ?, ?], we adopt ResNet12 [?] as the feature extraction backbone for all experiments. ResNet12 is a variant of residual networks tailored for FSL tasks. It consists of four residual blocks, each comprising three convolutional layers and one 2×2 max-pooling layer. All convolutional layers use 3×3 kernels, followed by batch normalization and a Leaky ReLU activation with a negative slope of 0.1. The output channels of the four blocks are 64, 160, 320, and 640, respectively. Given an input image of size $3 \times 84 \times 84$, the network produces a feature map of size $640 \times 5 \times 5$. On top of the backbone, we attach a 2-layer multi-layer perceptron (MLP) with a ReLU activation in between as the projection head, which maps the local CNN features to a 640-dimensional contrastive space.

Training Details. During meta-training, we optimize the model using SGD with Nesterov momentum of 0.9 and apply a weight decay of 1×10^{-4} to all parameters. The ResNet12 backbone is first pre-trained using a standard cross-entropy loss on the training set for initialization. Subsequently, the model is trained with DSCL for 300 epochs on UCM and WHURS, and 30 epochs on NWPU, using an initial learning rate of 0.01, which decays by a factor of 0.5 every one-third of the total epochs. We adopt the data augmentation strategies from [?], including random resized cropping, horizontal flipping, color jittering, and random rotation. The temperature parameter τ in the contrastive loss is set to 0.07.

Evaluation Protocol. We evaluate our model on 5-way 1-shot and 5-way 5-shot classification tasks, following the standard FSL protocol [?]. Each testing task contains 15 query samples per class. We sample 2,000 tasks from the novel classes for testing and report the mean classification accuracy along with the 95% confidence interval across all tasks as the result.

3.3 Results

The comparison of different methods on three widely used benchmark datasets (UCM, WHU-RS19, and NWPU-RESISC45) under the 5-way 1-shot and 5-way 5-shot settings is shown in Table 1. Our proposed method,

Table 1. Few-shot remote sensing scene classification results on UCM, WHURS, NWPU datasets.

Methods	UCM		WHURS		NWPU	
	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
MAML [?]	48.94 ± 0.31	60.61 ± 0.29	50.87 ± 0.23	64.26 ± 0.32	48.04 ± 0.21	62.98 ± 0.47
Meta-SGD [?]	51.13 ± 0.95	63.68 ± 0.59	51.78 ± 1.05	65.47 ± 0.65	40.96 ± 0.08	47.46 ± 0.37
MatchingNet [?]	34.68 ± 0.91	53.34 ± 0.17	51.25 ± 0.61	54.36 ± 0.38	40.31 ± 0.13	47.27 ± 0.28
ProtoNet [?]	52.34 ± 0.19	69.28 ± 0.67	58.17 ± 0.56	80.54 ± 0.42	41.38 ± 0.26	62.77 ± 0.14
RelationNet [?]	48.48 ± 0.75	62.17 ± 0.33	61.74 ± 0.51	79.15 ± 0.35	66.21 ± 0.28	78.37 ± 0.28
TPN [?]	53.36 ± 0.77	68.23 ± 0.52	66.51 ± 0.87	78.50 ± 0.56	66.51 ± 0.87	78.50 ± 0.56
DLA-Match [?]	53.76 ± 0.62	63.01 ± 0.51	68.27 ± 1.83	79.89 ± 0.33	68.80 ± 0.70	81.63 ± 0.46
SCL-MLNet [?]	51.37 ± 0.79	68.09 ± 0.92	63.36 ± 0.88	77.62 ± 0.81	62.21 ± 1.12	80.86 ± 0.76
SPNet [?]	57.64 ± 0.73	73.52 ± 0.51	81.06 ± 0.60	88.04 ± 0.28	67.84 ± 0.87	83.94 ± 0.50
TSC [?]	55.11 ± 0.68	69.20 ± 0.64	70.99 ± 0.74	82.18 ± 0.32	73.26 ± 0.15	84.62 ± 0.35
S2M2 [?]	56.42 ± 0.40	71.97 ± 0.27	69.00 ± 0.41	82.14 ± 0.21	63.24 ± 0.47	83.23 ± 0.28
MPCL-Net [?]	56.46 ± 0.21	76.57 ± 0.07	61.84 ± 0.12	80.34 ± 0.54	55.94 ± 0.04	76.24 ± 0.12
DCN [?]	58.64 ± 0.71	76.61 ± 0.49	81.74 ± 0.55	91.67 ± 0.25	74.40 ± 0.78	89.22 ± 0.41
FBCL [?]	58.45 ± 0.74	74.59 ± 0.59	-	-	75.35 ± 0.82	88.90 ± 0.41
ACL-Net [?]	59.74 ± 0.46	74.89 ± 0.29	78.30 ± 0.32	90.43 ± 0.15	76.13 ± 0.24	86.54 ± 0.23
DSCL (ours)	63.63 ± 0.38	81.35 ± 0.24	86.19 ± 0.30	95.06 ± 0.11	73.81 ± 0.44	89.57 ± 0.22

DSCL, can achieve competitive performance compared to state-of-the-art approaches in both settings, demonstrating strong generalization capabilities for FSRSSC.

Comparison with conventional few-shot learning methods. Traditional FSL methods are primarily meta-learning based, including MAML [?], Meta-SGD [?], MatchingNet [?], ProtoNet [?], and RelationNet [?]. These approaches generally perform suboptimally across all datasets. For example, ProtoNet achieves only 52.34% (1-shot) and 69.28% (5-shot) on UCM, while DSCL achieves 63.63% and 81.35%, yielding improvements of 11.3% and 12.1%, respectively. These results indicate the limitations of early FSL methods in capturing the complex semantics of remote sensing scenes. SPNet [?], an enhanced version of ProtoNet with self- and inter-calibration mechanisms, also underperforms compared to DSCL.

Comparison with contrastive learning-based methods. Recent works have widely adopted contrastive learning to enhance feature representations, leading to notable performance gains for FSRSSC. For example, SCL-MLNet [?] incorporates self-supervised contrastive learning as an auxiliary task. TSC [?] introduces task-specific contrastive losses within episodes. MPCL-Net [?] proposes a multi-pretext prototype-guided contrastive framework. DCN [?] employs a dual-branch design to capture both contextual and detailed features using supervised contrastive learning. Foreground-background contrastive learning [?] separates discriminative regions to improve representation learning. ACL-Net [?] incorporates a hybrid attention module and a dictionary-based contrastive loss.

Despite these advancements, DSCL achieves superior results, particularly on the UCM and WHURS datasets. On UCM, DSCL outperforms the next-best method by 3.9% in the 1-shot setting and 4.7% in the 5-shot setting. On WHURS, it surpasses DCN by 4.5% (1-shot) and 3.4% (5-shot). On NWPU, DSCL achieves the best performance in the 5-shot setting and remains competitive in the 1-shot setting. In addition to its strong performance, our model, built on the original ResNet12 backbone for FSL, is also more efficient than existing methods. Most competing approaches introduce additional parametric components into the feature extractor, such as attention modules [?,?,?,?,?], to enhance feature learning. Finally, DSCL significantly outperforms earlier methods such as [?, ?, ?], further validating the effectiveness of our approach.

4 Conclusion

This paper proposes Dense Supervised Contrastive Learning (DSCL) for FSRSSC. Unlike existing methods

that primarily rely on global features, DSCL applies supervised contrastive learning at the patch level to enhance the discriminability of local features. This approach is motivated by the unique characteristics of remote sensing imagery, which often includes numerous small, densely distributed, and discriminative objects. Experiments on three standard benchmark datasets (UCM, WHU-RS19, and NWPU-RESISC45) demonstrate that DSCL achieves competitive performance compared to recent state-of-the-art methods, validating the effectiveness of local-level contrastive learning for FSRSSC. DSCL offers a promising alternative and highlights the advantages of strengthening local representation learning in this domain. Future work may explore the integration of multi-scale features to further improve generalization in more challenging few-shot scenarios.