Semantic Segmentation in Aerial Imagery Using Multi-level Contrastive Learning with Local Consistency

Maofeng Tang[1], Konstantinos Georgiou[1], Hairong Qi[1], Cody Champion[2], Marc Bosch[2]
[1]Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville
[2]Accenture Federal Services
{mtang4,kgeorgio}@vols.utk.edu, hqi@utk.edu, {cody.champion,marc.bosch.ruiz}@afs.com

Abstract

Semantic segmentation in large-scale aerial images is an extremely challenging task. On one hand, the limited ground truth, as compared to the vast area the images cover, greatly hinders the development of supervised representation learning. On the other hand, the large footprint from remote sensing raises new challenges for semantic segmentation. In addition, the complex and ever changing image acquisition conditions further complicate the problem where domain shifting commonly occurs. In this paper, we exploit self-supervised contrastive learning (CL) methodologies for semantic segmentation in aerial imagery. In addition to performing CL at the feature level as most practices do, we add another level of contrastive learning, at the semantic level, taking advantage of the segmentation output from the downstream task. Further, we embed local mutual information in the semantic-level CL to enforce local consistency. This has largely enhanced the representation power at each pixel and improved the generalization capacity of the trained model. We refer to the proposed approach as multi-level contrastive learning with local consistency (mCL-LC). The experimental results on different benchmarks indicate that the proposed mCL-LC exhibits superior performance as compared to other state-of-the-art contrastive learning frameworks for the semantic segmentation task. mCL-LC also carries better generalization capacity especially when domain shifting exists.

1. Introduction

Remote collection of high-resolution imagery data along both temporal and spatial dimensions has allowed for large areas of the planet to be monitored regularly, thus enabling a wide variety of tasks such as disaster monitoring, urban planning, agricultural planning [35, 39], etc. However, due to the large footprint of aerial images and limited sensor bandwidth, there is considerable interest and investigation into the classification of object types at the pixel level. The extraction of this information is the basis of semantic segmentation in aerial images [42].

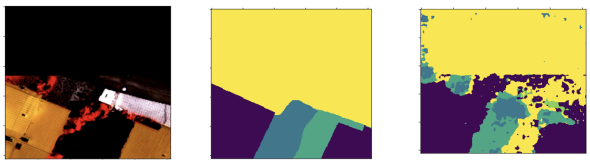


Figure 1. Comparison of semantic segmentation results on an aerial image using different frameworks. (a) Raw image. (b) Ground truth. Segmentation results using (c) DeepLab v3+ [9] (supervised), (d) MoCo v2 [12] (feature-level self-supervised), (e) GLCNet (global/local-level self-supervised) [26], and (f) the proposed mCL-LC method.

With the success of deep learning techniques in high-level and abstract feature learning, e.g., VGG [36], ResNet [20], and MobileNet [22, 33, 21], various semantic segmentation models based on these backbones have been proposed to yield accurate and reproducible results, such as SegNet [2], PSPNet [44], Mask-RCNN [19], DenseASPP [43], DeepLab [7, 8, 9], Fast-SCNN [30], etc. However, these methods need to rely on large amounts of data with high-quality labels [14, 27], which might not be feasible in many scenarios of aerial imagery where the volume of created data is extremely large while the inherent speed of human annotators is extremely limited. When only limited data is available for training, these semantic models tend to over-fit and result in poor performance.

To address the issue of the lack of high-fidelity labeled data, several options are available, including 1) modifying

existing data through augmentations such as cropping, flipping, etc. [16], 2) generating synthetic data through generative adversarial networks [31] or physics-based models, 3) using pre-trained models and fine-tune with the target data [15], and 4) applying transfer learning methodologies to reduce dependency on the labeled data [1]. A major limitation of all these techniques is that they are still very much reliant on a relatively significant amount of labeled data. To this end, self-supervised learning strategies [29, 4, 25, 28, 24] have been gaining more and more attention.

As one of the most effective self-supervised learning techniques developed recently, contrastive learning [10] has achieved significant breakthrough in extracting powerful representations without the need of any annotations. It utilizes augmentations of the image to extract representations of similar and dissimilar images and construct highly discriminant models. Contrastive learning has shown strength in image segmentation [45, 38, 41], but only for natural images.

In contrast to the natural image, where most segmentation tasks are instance based, the large footprint of remote sensing imagery demands the models to be more representative of the local semantic information [26]. Hence, in addition to the commonly-adopted feature-level contrastive learning (CL) based on high-level features extracted from the encoder, we also apply CL at the semantic level taking advantage of the semantic segmentation output from the decoder. We further enforce the local consistency by maximizing the local mutual information, thus boosting the representation power at local pixels.

We refer to the proposed framework as multi-level contrastive learning with local consistency (mCL-LC). Figure 1 illustrates the effectiveness of mCL-LC in semantic segmentation as compared to supervised and state-of-the-art self-supervised approaches. The main contribution is threefold:

(1) We propose a multi-level CL framework where CL is conducted at both the feature level and the semantic level taking advantage of the semantic segmentation results, in order to boost the representative and discriminative power at local details.

(2) We introduce the mutual information as a physical constraint for local consistency to the semantic-level CL such that smoothness can be preserved while revealing local details. By maximizing the mutual information, we can further enhance the model's representation capacity at local pixels.

(3) We identify an effective augmentation scheme, pseudo-cloud noise generation, tailored to the aerial image analysis, showing the importance of augmentation in improving model robustness and generalization capacity.

The remainder of this paper is organized as follows. Sec. 2 reviews recent developments in contrastive-based learning frameworks. Sec. 3 elaborates on the proposed mCL-LC model design. Sec. 4 presents details about experiments and results. Sec. 5 concludes the paper and provides general directions for further improvements.

2. Contrastive-based Learning Frameworks

Contrastive learning is a self-supervised learning framework, aiming to make the network learn significant representations for the downstream task in an unsupervised fashion. The gist of the contrastive learning mechanism is that unlabeled images are used to create data pairs, i.e., pixels that contain "information" from the same image. This pixel information is augmented using various methods to preserve the underlying information but can be viewed as from visually distinct sources. The training process then encodes these pairs of images, typically through a CNN, and generates a compact feature set. The final vectorized representations of similar images are compared and if the vectors agree then the model is re-enforced. Similarly, negative examples (image pairs from different sources) are used to further enforce the model accuracy by providing a repellent force during training. The result of this approach is a model that readily discriminate similar and dissimilar features in the unlabeled training set but do not encode information about the underlying class.

Once the feature extractor is trained to represent highly dense information, a fine-tuning model is trained to reduce these higher dimensional representations into a class label [23] using a smaller dataset with labels. The features learned by this supervised classifier are extendable to the unsupervised dataset due to the deep feature representations learned by the initial model. This paradigm is known as the "unsupervised pre-train, supervised fine-tune, and knowledge distillation" [11].

There have been a couple of contrastive learning-based networks developed. Specifically, the SimCLR method [11] is based on the idea of instance-wise contrastive learning, which learns by forcing positive samples augmented from the same sample to be similar and negative samples augmented from different samples in a mini-batch to be dissimilar. MoCo v2 [18, 12] is also based on the idea of instance-wise contrastive learning but with a focus on obtaining negative samples far beyond the batch size, such that a dynamic queue with the features of negative samples is maintained and the consistency problem alleviated using a momentum update encoder. BYOL [17] and SimSiam [13] are also instance-wise but only focus on the representation learning of positive pairs.

Although effective, all the above mentioned contrastive learning frameworks perform learning based on features obtained from the encoder, which we refer to as feature-level

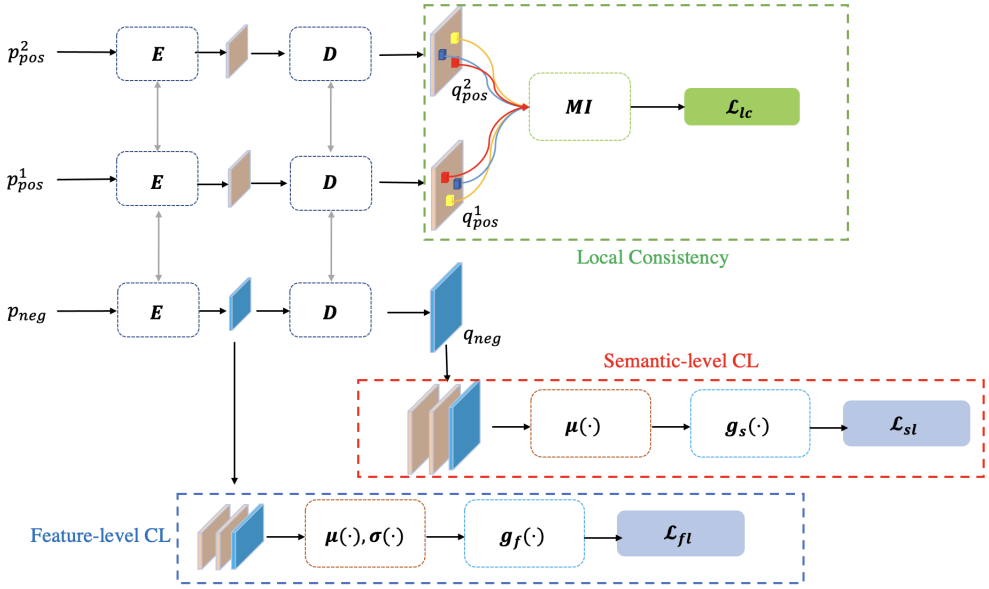


Figure 2. Illustration of the proposed mCL-LC architecture.

CL. Feature-level CL tends to perform well overall for segmentation tasks but not attend to local details, which is specifically important in aerial images. GLCNet [26] alleviates from this issue by adding a local feature contrastive learning module. However, since GLCNet does not provide an effective mechanism for handling local consistency in local features, the segmentation results tend to be noisy within local regions. These can be observed from Fig. 1.

The proposed mCL-LC enforces local consistency on positive pairs within the semantic-level contrastive learning, in addition to the commonly used feature-level learning, striking a good balance between local details and smoothness while preserving global structure.

3. Method

As discussed in Sec. 1, the large footprint of aerial imagery demands a more powerful representation at the local pixel level but the ever-changing acquisition conditions only complicate the problem - making it more difficult to extract invariant features. The proposed mCL-LC is designed to solve these issues. Its architecture is shown in Fig. 2. Generally speaking, a contrastive learning-based framework consists of three components: 1) data augmentation, 2) representation learning, and 3) contrastive loss. In the following, we first elaborate on the multi-level CL (mCL) (Sec. 3.1) at both the feature and semantic levels. We then describe the local consistency module constrained by mutual information that is embedded in the semantic-level CL (Sec. 3.2). In the end, we detail the data augmentation component (Sec. 3.3), especially the pseudo-cloud noise generation, tailored toward segmentation tasks in aerial imagery.

3.1. Multi-level Contrastive Learning (mCL)

Besides the commonly used feature-level contrastive learning, for the semantic segmentation task, we also propose to apply contrastive learning at the semantic level, such that "contrasts" are learned using not only the high-level features, but also the local semantics, in order to boost the representation and discrimination capacity at the local pixel.

Feature-level CL. The feature-level contrastive learning module, $\boldsymbol{E}$, as shown in Fig. 2, uses the encoder of DeepLab v3+ with ResNet50 as backbone. Upon feeding an input image patch p to $\boldsymbol{E}$, a high-level representation, $\boldsymbol{E}(p)$, can be obtained. These representations are then used to generate the style feature [46], including both the channel-wise mean, $\mu(\cdot)$, and the variance, $\sigma(\cdot)$, as following,

$$f(p) = concat(\mu(\boldsymbol{E}(p)), \sigma(\boldsymbol{E}(p))) \tag{1}$$

Before computing the contrastive loss, a nonlinear projection head $g_f(\cdot)$ is needed, which has been proven to be effective [11]. So, after the encoder and projection head, we obtain the representation, $z = g_f(f(p))$. The feature-level contrastive loss is thus defined as:

$$\mathcal{L}_{fl} = \frac{1}{2N} \sum_{k=1}^{N} \left(\ell_{NTX}(\tilde{p}_k, \hat{p}_k) + \ell_{NTX}(\hat{p}_k, \tilde{p}_k) \right) \tag{2}$$

where $\hat{p}_k$ and $\tilde{p}_k$ are a positive patch pair generated from augmentations of the same patch p_k, and the NT-Xent contrastive loss function ℓ_{NTX} is the same as in SimCLR [10], which is defined as following:

$$\ell_{NTX}\left(\tilde{p}_k, \hat{p}_k\right) = -\log \frac{\exp\left(\operatorname{sim}\left(\tilde{z}_k, \hat{z}_k\right)/\tau\right)}{\sum_{p\in\Lambda^-} \exp\left(\frac{\operatorname{sim}(\tilde{z}_k, g_f(f(p)))}{\tau}\right)} \tag{3}$$

where Λ^- is the rest of the patch sets in the batch and $\tilde{z}_k = g_f\left(f\left(\tilde{p}_k\right)\right)$, $\hat{z}_k = g_f\left(f\left(\hat{p}_k\right)\right)$. Via minimizing the contrastive loss, it learns by forcing the representations from positive view pairs to be similar but those from negative pairs to be dissimilar.

Semantic-level CL. For segmentation tasks, a decoder structure is needed that learns by forcing the semantic feature to be similar in positive pair and dissimilar in negative pair. Here, we use the decoder of DeepLab v3+ as the semantic module, denoted as $\boldsymbol{D}$ (Fig. 2). Given the patch p_k, $\boldsymbol{D}$ receives the output of $\boldsymbol{E}$ and generates a semantic representation of the same size as that of p_k, which is denoted as $q_k = \boldsymbol{D}(\boldsymbol{E}(p_k))$. Integrating the channel information, we thus produce a pseudo semantic map s_k, and $s_k = g_s(\mu(q_k))$, where $\mu(\cdot)$ is the channel-wise average operator and $g_s(\cdot)$ is the projection head. Following the same procedure as in the feature-level contrastive loss, we apply the NT-Xent to calculate the semantic contrastive loss, which is defined as,

$$\mathcal{L}_{sl} = \frac{1}{2N}\sum_{k=1}^{N}\left(\ell_{NTX}\left(\tilde{s}_k, \hat{s}_k\right) + \ell_{NTX}\left(\hat{s}_k, \tilde{s}_k\right)\right) \tag{4}$$

$$\ell_{NTX}\left(\tilde{s}_k, \hat{s}_k\right) = -\log \frac{\exp\left(\operatorname{sim}\left(\tilde{s}_k, \hat{s}_k\right)/\tau\right)}{\sum_{p\in\Lambda_S^-} \exp\left(\frac{\operatorname{sim}(\tilde{s}_k, g_s(\mu(\boldsymbol{D}(\boldsymbol{E}(p)))))}{\tau}\right)} \tag{5}$$

where N denotes the number of patch pairs from a mini-batch of N samples, Λ_S^- is a set of pseudo maps corresponding to all patches except for the positive pair, and $g_s(\cdot)$ is a projection head similar to $g_f(\cdot)$.

3.2. Local Consistency Learning (LC)

So far, we have constructed a multi-level contrastive learning framework. However, the complex acquisition conditions of remote sensing imagery demand more robust representation schemes that would reveal the rich details hidden under the surface of the large footprint. In other words, a module that can understand the local semantic details is needed.

In [5], the contrastive loss in the local region is used to improve the model's performance in learning representation of natural images, which forces similarity of local region of interests (ROIs) in positive pair, but dissimilarity in negative pair. However, this strategy is not adequate for aerial imagery because of its unique characteristics, i.e., spatial auto-correlation, that can cause two ROIs in different images to be similar. This is illustrated in Fig. 3, where two different images form a negative pair, but the three ROI pairs, matched by their geo-location, can be either similar (e.g., the red and yellow ROI pairs) or dissimilar (e.g., the blue ROI pair). However, the contrastive loss would have forced representations of the similar patches to be dissimilar, which is not desirable. Hence we introduce the local consistency loss, to ensure the network only preserves the consistency of the local semantic information in the matching position of positive pairs.

Figure 3. ROIs in negative pair. The red, yellow, blue boxes are the matching ROIs in a negative image pair, but the matching ROIs bounded by the red and yellow boxes are actually similar.

Specifically, to obtain the matching ROIs in the positive pair, we first randomly select an ROI from $\tilde{p}$, then determine the location of the same size matching ROI in $\hat{p}$ according to the position of the ROI in $\tilde{p}$ to ensure the center of the two matching local regions point to the same position in the original image. By repeating these steps, we can select multiple different ROIs. We pass the locations of these ROIs to the last layer of $\boldsymbol{D}$ and select the corresponding ROIs in the pseudo semantic map. Although these matching local regions are from different augmentation, they should share the same content. We thus implement the consistency loss between the matching positive ROI pair, $\tilde{r}_j$ and $\hat{r}_j$, by calculating the mutual information and maximizing it.

Mutual information (MI) has been widely used for problems like multi-modality registration [47, 40]. It is a Shannon-entropy based measurement of mutual independence between two random variables, e.g., $\tilde{r}_j$ and $\hat{r}_j$. The mutual information $\mathcal{I}\left(\tilde{r}_j; \hat{r}_j\right)$ measures how much uncertainty of one variable ($\tilde{r}_j$ or $\hat{r}_j$) is reduced given the other variable ($\hat{r}_j$ or $\tilde{r}_j$). Based on the mutual information, the local consistency loss is defined as follows:

$$\mathcal{L}_{lc} = -\frac{1}{N_R}\sum_{j=1}^{N_R}\left(\mathcal{I}\left(\tilde{r}_j; \hat{r}_j\right)\right) \tag{6}$$

with

$$\mathcal{I}\left(\tilde{r}_j; \hat{r}_j\right) = H\left(\tilde{r}_j\right) - H\left(\tilde{r}_j \mid \hat{r}_j\right) \tag{7}$$

$$= \int_{\tilde{r}_j \times \hat{r}_j} \log \frac{\mathbb{P}_{\tilde{r}_j \hat{r}_j}}{\mathbb{P}_{\tilde{r}_j} \otimes \mathbb{P}_{\hat{r}_j}} d\mathbb{P}_{\tilde{r}_j \hat{r}_j}, \tag{8}$$

where H indicates the Shannon entropy and $H(\tilde{r}_j \mid \hat{r}_j)$ is the conditional entropy of $\tilde{r}_j$ given $\hat{r}_j$. N_R is the number of ROIs in one patch, and $\tilde{r}_j$ is the corresponding ROI in the pseudo semantic map. The exact value of the mutual information is difficult to calculate, so we resort to an estimate based on the MINE algorithm [3], which is implemented by back propagation in a two-layer fully connected network.

To summarize, the final loss function of mCL-LC consists of three parts: 1) feature-level contrastive loss, $\mathcal{L}_{fl}$ (Eq. 2); 2) semantic-level contrastive loss, $\mathcal{L}_{sl}$ (Eq. 4); and 3) local consistency loss, $\mathcal{L}_{lc}$ (Eq. 6). The final loss function is defined as follows:

$$\mathcal{L} = \mathcal{L}_{fl} + \mathcal{L}_{sl} + \mathcal{L}_{lc}, \tag{9}$$

The whole processing pipeline is shown in Algorithm 1.

Algorithm 1 mCL-LC Training

Require: hyper parameters τ, batch size N, ROI Number N_R

Input: Training Set I

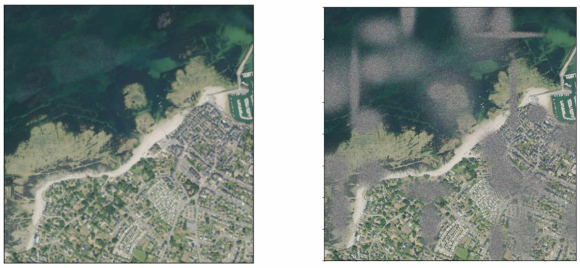
Output: Pretrained $\boldsymbol{E}$ and $\boldsymbol{D}$

for each patch p_k in batch $P = \{p_k\}_{k=1}^{N}$ **do**
  Build ROI position label $O = \{\boldsymbol{o}_k\}_{k=1}^{N}$
  for all samples in the batch **do**
    Get augmented sample and position: $\hat{p_k}, \tilde{p_k}$, $\hat{r_k}, \tilde{r_k}$
  end for
  Extract structure feature: $\tilde{z}_k$ and $\hat{z}_k$
  Extract semantic feature: $\tilde{s}_k$ and $\hat{s}_k$
  Get local feature: $\tilde{r}_k$ and $\hat{r}_k$ by the position $\tilde{o}_k, \hat{o}_k$
  Compute the loss $\mathcal{L}_{fl}$ (Eq. 2), $\mathcal{L}_{sl}$ (Eq. 4), and $\mathcal{L}_{lc}$ (Eq. 6)
  Compute the total loss $\mathcal{L}$ by Eq. 9
  Update network weights
end for =0

3.3. Data Augmentation for Aerial Imagery

Generally speaking, contrastive learning encourages the model to learn spatiotemporal-invariant features, where data augmentation plays an important role. As with common augmentation operations adopted in remote sensing images [26], we perform spatial transformations such as random cropping, resizing, flipping, and rotation for the learning of *spatially-invariant* features and simulate temporal transformations such as color distortion, Gaussian blur, and random noise for the learning of *temporally-invariant* features.

More importantly, considering the intrinsic characteristics of aerial imagery where cloud cover is often the major limiting factor that affects the success of downstream tasks [37], we propose a new augmentation method, referred to as the "pseudo-cloud noise generation", to simulate the potential disturbance caused by cloud, such that the model can also learn *cloud-invariant features*. In this method, some point clouds are randomly created and the associated RGB pixel values are randomly increased by 50 to 100 percent to mimic the variable increase in reflectance caused by cloud formations. The benefit of this approach is its stochastic nature of cloud formation, realized by changing multiple parameters, such as center of cloud, standard deviation of the cloud cluster, and size of cloud coverage, in a random fashion, mimicking the physical appearance and properties of the natural cloud (e.g., formed in clusters and diffuses at the edges). Fig. 4 shows one example of the pseudo-cloud noise



(a) (b)

Figure 4. Illustration of the effect of generated pseudo-cloud noise: (a) original image, (b) original image with the cloud mask added.

generation algorithm. Experiments are conducted (Sec. 4) to show the benefit of including the pseudo-cloud noise addition as a key augmentation method in aerial image processing.

4. Experimental Evaluation

4.1. Experimental Design

The proposed mCL-LC is evaluated from two aspects: multi-category semantic segmentation accuracy and the generalization capacity.

We evaluate the proposed mCL-LC and other self-supervised methods on different benchmarks including ISPRS Potsdam [32], ISPRS Vaihungen [32], Nice in MiniFrance, and Nantes Saint in MiniFrance [6]. For each dataset, we randomly divide the data to 90%, 2%, 8%, for training, fine-tuning, and testing, respectively. Specifically, the training is in a contrastive learning mode without the label, but fine-tuning is in a supervised manner with associated labels. In the testing phase, the label is used to evaluate the performance quantitatively. More details of the benchmarks used are shown in Table 1. It is worth mentioning that the MiniFrance dataset covers 16 land-use categories, significantly more than other benchmarks, which results in land-use categories being unbalanced and sparse in different regions. In addition, rather than categorizing at the object level (e.g., cars, buildings, trees, etc.), the Nice and

Nantes Saint datasets require the model to be able to understand spatial correlation. For example, if seeing groups of houses and buildings, the site should be identified as an "urban area", which is typical in aerial image analysis [34]. All these characteristics present additional challenge to the segmentation model.

Table 1. DESCRIPTION OF THE FOUR DATASETS. NOTE THAT ALL IMAGES ARE WITH 3 CHANNELS (RGB)

Datasets	Potsdam	Vaihingen	Nice	Nantes Saint
Resolution	0.05 m	0.09 m	0.5 m	0.5 m
Categories	6	6	16	16
Training	13916	12525	14686	19589
Fine-tuning	310	278	330	435
Testing	1237	1113	1405	1741

We compare the semantic segmentation performance of pre-trained mCL-LC with five state-of-the-art CL networks including SimCLR [11], MoCo v2 [18, 12], GLCNet [26], BYOL [17], and SimSiam [13]. During the training process, the backbone for all the methods is set to DeepLab v3+. The patch size is fixed at 256×256. For all the models, we use the Adam optimizer and train for 200 epochs, with a batch size of 64. The initial learning rate is set to 0.001 with a cosine decay schedule. For the proposed mCL-LC, we choose 12 ROIs with a size of 8×8 from a patch. During the fine-tuning, the number of epochs is set to 20 and the initial learning rate is set as 0.0001.

For the evaluation metrics, we select the overall accuracy (OA) and Kappa coefficient (Kappa), the metric indicating the degree of correctness and reliability of a classifier, to measure the overall pixel-level classification accuracy. In addition, the F1-score is used to measure the class-wise classification accuracy. Note that three categories are missing in the Nice and Nantes Saint datasets. They are "cultivation patterns" (class 8), "orchards at the fringe of urban classes" (class 9), and "clouds and shadows" (class 15). In addition, since the first category is "no-information", we ignore it when calculating the F1-score.

4.2. Comparison with State-of-the-Art

In this set of experiments, we evaluate the performance of the proposed mCL-LC from OA, Kappa, and F1-score perspectives, in comparison with the five state-of-the-art contrastive learning frameworks. We also show the effectiveness of the pseudo-cloud noise generation augmentation method. The results are shown in Table 2, where "mCL-LC" means no pseudo cloud in the data augmentation, and "mCL-LC+" is with the pseudo cloud augmentation. From these results, we observe that the proposed mCL-LC outperforms all other contrastive learning frameworks in terms of both OA and Kappa. The addition of pseudo cloud augmentation further improves the performance by roughly 2%.

Besides the OA and Kappa metrics, we also calculate the F1-score for each category. These results are shown in Fig. 5. From the F1-score, we again observe that the proposed mCL-LC achieves the best performance for majority of the categories.

We further study the effect of the number and size of the ROIs when calculating the local consistency loss. The results are shown in Fig. 6. The base setting is three 2×2 ROIs in a patch. From the top row of Fig. 6, we can see that 8×8 and 16×16 ROIs achieve better performance. When the size of ROIs goes beyond 16×16, the performance starts decreasing. Similarly, from the bottom row of Fig. 6, we observe that 12 and 15 ROIs present better performance.

4.3. Ablation Study

The ablation study is two-fold. First, we investigate the important role played by each of the three loss functions. Second, we study the effect of using pseudo cloud augmentation in state-of-the-art contrast learning frameworks. Although we have shown the performance improvement by adding the pseudo cloud augmentation on the proposed mCL-LC, here, we extend the investigation to see if the conclusion can be generalized to other contrast learning frameworks.

The weight update of the encoder and decoder networks is mainly controlled by the loss function in Eq. 9 that consists of three modules, the feature-level contrastive loss, $\mathcal{L}_{fl}$, the semantic-level contrastive loss, $\mathcal{L}_{sl}$, and the local consistency loss, $\mathcal{L}_{lc}$. Table 3 thoroughly compares the segmentation accuracy using different combinations of these three modules, from which we make some interesting observations. First of all, the first three rows of the results, where only one loss module is applied, show that the self-supervised contrastive learning mechanism (using either $\mathcal{L}_{fl}$ or $\mathcal{L}_{sl}$) is very effective in representation learning as compared to using only the local consistency loss ($\mathcal{L}_{lc}$), although the effectiveness of the feature-level or semantic-level contrastive learning is roughly the same according to the OA and Kappa metrics. Second, the multi-level contrastive learning (mCL) using both $\mathcal{L}_{fl}$ and $\mathcal{L}_{sl}$ largely increases the performance (about 4%) as compared to either single-level learning approaches. Third, the addition of the local consistency loss ($\mathcal{L}_{lc}$) to either level of the contrastive learning (i.e., $\mathcal{L}_{fl}$+$\mathcal{L}_{lc}$ and $\mathcal{L}_{sl}$+$\mathcal{L}_{lc}$) also effectively improves the performance by around 4% on the Nice dataset and a much large margin on the Nantes Saint dataset. And finally, using all three losses drastically improves the overall performance, showing the important roles played by each of the three loss modules.

The second part of the ablation analysis studies the effect of the proposed pseudo cloud augmentation. In Table 2, we have reported how this new augmentation technique improves the proposed mCL-LC by roughly 2% in

Table 2. COMPARISON WITH STATE-OF-THE-ART CONTRASTIVE LEARNING FRAMEWORKS IN TERMS OF PIXEL-BASED SEMANTIC SEGMENTATION ACCURACY. NOTE THAT SUPERVISED BASELINE REFERS TO DEEPLAB V3+

	Nice		Nantes Saint		Potsdam		Vaihungen	
	OA	Kappa	OA	Kappa	OA	Kappa	OA	Kappa
Supervised Baseline	0.6712	0.5410	0.6628	0.5312	0.7518	0.6812	0.7603	0.7037
SimCLR	0.6127	0.5013	0.6025	0.4413	0.7327	0.6472	0.7286	0.6304
MoCo v2	0.6277	0.5082	0.6201	0.4513	0.7371	0.6735	0.7309	0.6213
BYOL	0.6366	0.5230	0.6311	0.4850	0.7509	0.6715	0.7546	0.6407
SimSiam	0.6305	0.5022	0.6036	0.4735	0.7439	0.6903	0.7492	0.6333
GLCNet	0.6494	0.5302	0.6407	0.5291	0.7811	0.7179	0.7855	0.6807
mCL-LC	**0.6944**	**0.5379**	**0.6793**	**0.5343**	**0.8053**	**0.7301**	**0.8281**	**0.7377**
mCL-LC +	**0.7120**	**0.5520**	**0.7008**	**0.5689**	**0.8217**	**0.7440**	**0.8453**	**0.7506**

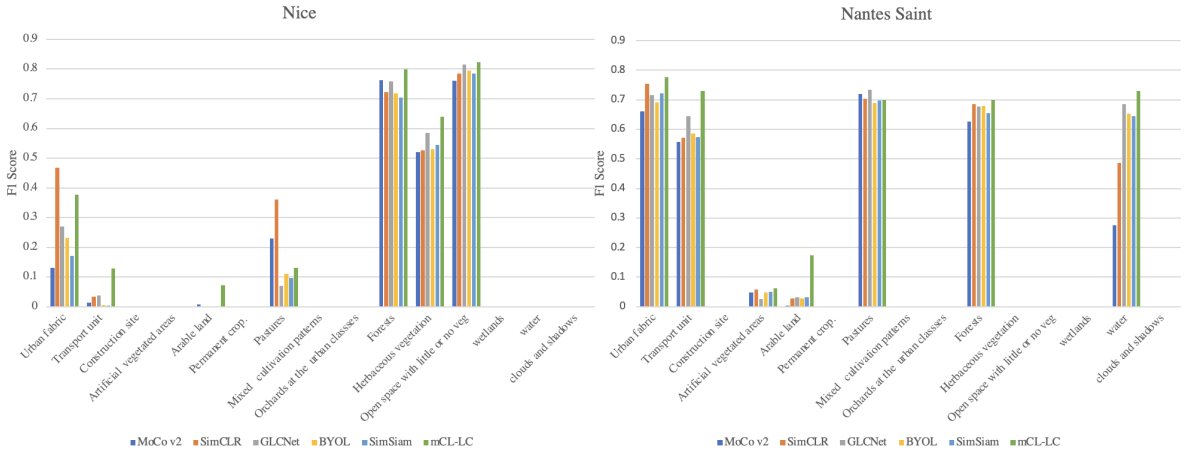


Figure 5. Comparison of class-wise F1-score on Nice and Nantes Saint.

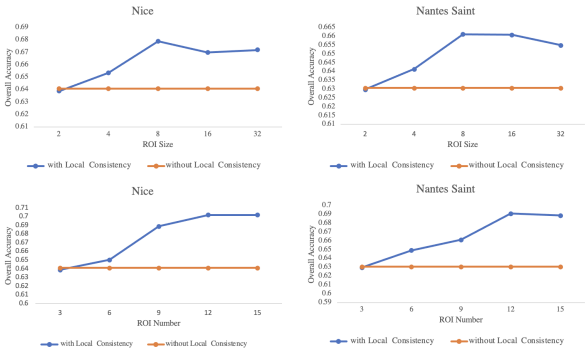


Figure 6. Effect of the size and number of ROIs in a patch when calculating the local consistency.

terms of OA on all four datasets. Here, we extend the investigation and exploit the potential of deploying pseudo cloud generation as a standard augmentation approach for other contrastive learning frameworks in aerial image processing. The results shown in Fig. 7 demonstrate a consistent 1% - 3% OA increment across all the six frameworks, providing convincing evidence of pseudo cloud generation as a standard augmentation approach benefiting aerial image analysis.

Table 3. ABLATION STUDY OF THE EFFECT OF EACH OF THE THREE LOSSES IN EQ. 9

Modules	Nice		Nantes Saint	
	OA	Kappa	OA	Kappa
$\mathcal{L}_{fl}$	0.6133	0.5107	0.5933	0.4395
$\mathcal{L}_{sl}$	0.6027	0.5283	0.5756	0.4510
$\mathcal{L}_{lc}$	0.4033	0.3212	0.4308	0.2977
$\mathcal{L}_{fl} + \mathcal{L}_{sl}$	0.6409	0.5539	0.6307	0.5371
$\mathcal{L}_{fl} + \mathcal{L}_{lc}$	0.6517	0.5324	0.6463	0.5133
$\mathcal{L}_{sl} + \mathcal{L}_{lc}$	0.6320	0.5681	0.6215	0.5482
$\mathcal{L}_{fl} + \mathcal{L}_{sl} + \mathcal{L}_{lc}$	**0.6944**	**0.5623**	**0.6793**	**0.5543**

4.4. Generalization Analysis

In this set of experiments, we evaluate the generalization capacity of mCL-LC using the zero-shot domain testing. Specifically, the training and fine-tuning are performed in one city and testing is conducted in the other city. The OA and Kappa for the different contrastive learning frameworks are shown in Table 4, where the city to the left of the arrow indicates the training and fine-tuning city, and that to the

Table 4. ZERO-SHOT GENERALIZATION COMPARISON IN SITES TRANSFER (OA)

OA	SimCLR	MoCo v2	GLCNet	BYOL	SimSiam	mCL-LC
Nice → Nantes Saint	0.4402	0.4684	0.4744	0.4325	0.3926	**0.6108**
Nantes Saint → Nice	0.0816	0.1762	0.3018	0.2004	0.2131	**0.4719**

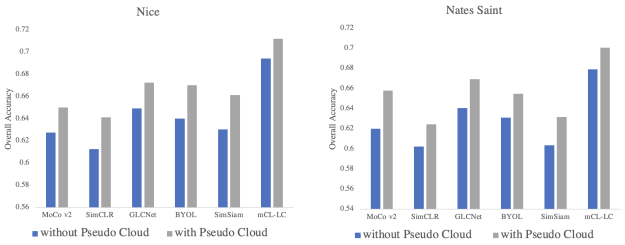


Figure 7. Results of an ablation experiment exploring the effectiveness of the pseudo cloud augmentation.

right is the testing city. From this table, we can observe that mCL-LC outperforms all other contrastive learning frameworks by more than 15%.

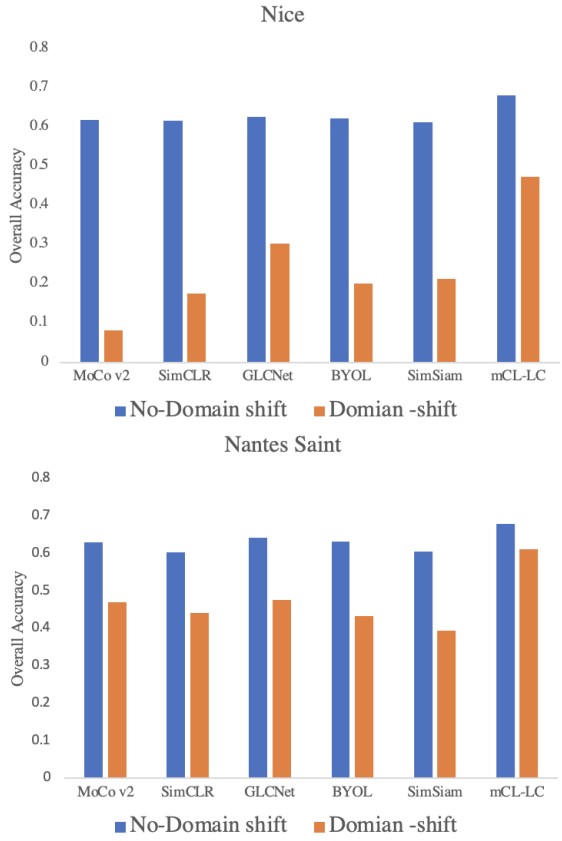


Figure 8. Comparison of generalization capacity of different frameworks with and without domain-shift.

Fig. 8 shows a more thorough comparison on the model's robustness to domain shift. In the figure, "no-domain shift" means the training, fine-tuning, and testing are conducted using the same city dataset, and "domain-shift" refers to the training and fine-tuning conducted on one city (e.g., Nice) and testing on the other city (e.g., Nantes Saint). We observe performance drop in all frameworks when domain-shift is present. However, the proposed mCL-LC drops the least as compared to other frameworks. For example, MoCo v2 decreases close to 90% in Nice data and 86.8% in Nantes Saint data with domain shift, but the proposed mCL-LC only drops less than 30% in Nice and 24.8% in Nantes Saint. This shows the superior performance of mCL-LC in generalization.

5. Conclusion

In this paper, we proposed a multi-level contrastive learning (CL) framework taking advantage not only the popular feature-level CL from the encoder output, but also the semantic-level CL from the decoder output, in order to boost the representation power at the local pixel level. This is essential especially for aerial image analysis where pixels tend to cover a large footprint. To further balance the tradeoff between local detail and local smoothness, we introduced mutual information as a physical constraint to enforce local consistency while preserving details. We further showed the great potential of pseudo-cloud generation as a standard augmentation technique for aerial imageries. The proposed mCL-LC framework has shown superior performance as compare to other single-level or multi-level CL frameworks, demonstrating strong generalization capacity especially when domain shift is present.

In the future, we plan to extend this work mainly in two aspects. The first is to investigate pixel-level representation framework and its contribution to the segmentation problem in aerial and natural imagery. The second is to explore the potential of CL in multi-modality representation in remote sensing.

Acknowledgement

This research is based upon work supported in part by the Office of the Director of National Intelligence (Intelligence Advanced Research Projects Activity) via 2021-20111000006. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U S Government. The U S Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

[1] Nouman Ahmed, Sudipan Saha, Muhammad Shahzad, Muhammad Moazam Fraz, and Xiao Xiang Zhu. Progressive unsupervised deep transfer learning for forest mapping in satellite image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 752–761, 2021.

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[3] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 531–540. PMLR, 10–15 Jul 2018.

[4] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018.

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

[6] Javiera Castillo-Navarro, Bertrand Le Saux, Alexandre Boulch, Nicolas Audebert, and Sébastien Lefèvre. Semi-supervised semantic segmentation in earth observation: the MiniFrance suite, dataset analysis and multi-task network study. *Machine Learning*, Apr. 2021.

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[8] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[9] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. Pmlr, 13–18 Jul 2020.

[11] 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.

[12] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. In *https://arxiv.org/abs/2003.04297*, 2020.

[13] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15750–15758, June 2021.

[14] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding, 2016.

[15] Maayan Frid-Adar, Avi Ben-Cohen, Rula Amer, and Hayit Greenspan. Improving the segmentation of anatomical structures in chest radiographs using u-net with an imagenet pre-trained encoder. In *Image Analysis for Moving Organ, Breast, and Thoracic Images*, pages 159–168. Springer, 2018.

[16] MAA Ghaffar, A McKinstry, T Maul, and TT Vu. Data augmentation approaches for satellite image super-resolution. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4:47–54, 2019.

[17] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc., 2020.

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, Ross Girshick, Huang Xun, and Serge Belongie. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, 2020.

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[21] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.

[22] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[23] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Process-*

ing Systems, volume 33, pages 18661–18673. Curran Associates, Inc., 2020.

[24] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. *Advances in Neural Information Processing Systems*, 31, 2018.

[25] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6874–6883, 2017.

[26] Haifeng Li, Yi Li, Guo Zhang, Ruoyun Liu, Haozhe Huang, Qing Zhu, and Chao Tao. Global and local contrastive self-supervised learning for semantic segmentation of hr remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, page 1–1, 2022.

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.

[28] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European conference on computer vision*, pages 527–544. Springer, 2016.

[29] Deepak Pathak, Ross Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. Learning features by watching objects move. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2701–2710, 2017.

[30] Rudra PK Poudel, Stephan Liwicki, and Roberto Cipolla. Fast-scnn: Fast semantic segmentation network. *arXiv preprint arXiv:1902.04502*, 2019.

[31] Caijun Ren, Xiangyu Wang, Jian Gao, Xiren Zhou, and Huanhuan Chen. Unsupervised change detection in satellite images with generative adversarial network. *IEEE Transactions on Geoscience and Remote Sensing*, 59(12):10047–10061, 2020.

[32] Franz Rottensteiner. Isprs test project on urban classification and 3d building reconstruction: Evaluation of building reconstruction results. Technical report, Technical report, 2013.

[33] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[34] Ronny Hänsch; Claudio Persello; Gemine Vivone; Javiera Castillo Navarro; Alexandre Boulch; Sebastien Lefevre; Bertrand Le Saux. Data fusion contest 2022 (dfc2022), 2022.

[35] Guy J-P. Schumann, G. Robert Brakenridge, Albert J. Kettner, Rashid Kashif, and Emily Niebuhr. Assisting flood disaster response with earth observation data and products: A critical assessment. *Remote Sensing*, 10(8), 2018.

[36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[37] Sergii Skakun, Jan Wevers, Carsten Brockmann, Georgia Doxani, Matej Aleksandrov, Matej Batič, David Frantz, Ferran Gascon, Luis Gómez-Chova, Olivier Hagolle, Dan López-Puigdollers, Jérôme Louis, Matic Lubej, Gonzalo Mateo-García, Julien Osman, Devis Peressutti, Bringfried Pflug, Jernej Puc, Rudolf Richter, Jean-Claude Roger, Pat Scaramuzza, Eric Vermote, Nejc Vesel, Anže Zupanc, and Lojze Žust. Cloud mask intercomparison exercise (cmix): An evaluation of cloud masking algorithms for landsat 8 and sentinel-2. *Remote Sensing of Environment*, 274:112990, 2022.

[38] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7303–7313, 2021.

[39] M. Weiss, F. Jacob, and G. Duveiller. Remote sensing for agricultural applications: A meta-review. *Remote Sensing of Environment*, 236:111402, 2020.

[40] Jonghye Woo, Maureen Stone, and Jerry L. Prince. Multimodal registration via mutual information incorporating geometric and spatial context. *IEEE Transactions on Image Processing*, 24(2):757–769, 2015.

[41] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16684–16693, 2021.

[42] Jin Xing, Renée Sieber, and Margaret Kalacska. The challenges of image segmentation in big remotely sensed imagery data. *Annals of GIS*, 20(4):233–244, 2014.

[43] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3684–3692, 2018.

[44] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[45] Xiangyun Zhao, Raviteja Vemulapalli, Philip Andrew Mansfield, Boqing Gong, Bradley Green, Lior Shapira, and Ying Wu. Contrastive learning for label efficient semantic segmentation. In *CVPR*, pages 10623–10633, 2021.

[46] Xian Zhong, Cheng Gu, Mang Ye, Wenxin Huang, and Chia-Wen Lin. Graph complemented latent representation for few-shot image classification. *IEEE Transactions on Multimedia*, 2022.

[47] Barbara Zitová and Jan Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21(11):977–1000, 2003.