Sparse and Complete Latent Organization for Geospatial Semantic Segmentation

Fengyu Yang* Chenyang Ma*
University of Michigan
{fredyang, dannymcy}@umich.edu

Abstract

Geospatial semantic segmentation on remote sensing images suffers from large intra-class variance in both foreground and background classes. First, foreground objects are tiny in the remote sensing images and are represented by only a few pixels, which leads to large foreground intra-class variance and undermines the discrimination between foreground classes (issue firstly considered in this work). Second, background class contains complex context, which results in false alarms due to large background intra-class variance. To alleviate these two issues, we construct a sparse and complete latent structure via prototypes. In particular, to enhance the sparsity of the latent space, we design a prototypical contrastive learning to have prototypes of the same category clustering together and prototypes of different categories to be far away from each other. Also, we strengthen the completeness of the latent space by modeling all foreground categories and hardest (nearest) background objects. We further design a patch shuffle augmentation for remote sensing images with complicated contexts. Our augmentation encourages the semantic information of an object to be correlated only to the limited context within the patch that is specific to its category, which further reduces large intra-class variance. We conduct extensive evaluations on a large scale remote sensing dataset, showing our approach significantly outperforms state-of-the-art methods by a large margin.

1. Introduction

Remote sensing images are high-resolution images acquired from a long distance from the Earth's surface and contain rich geospatial information. Geospatial semantic segmentation aims to interpret the remote sensing images by assigning each pixel a semantic category. With a wide range of applications such as environmental assessment, infrastructure planning, natural resources management [24, 25], geospatial semantic segmentation draws close attention in the remote sensing community.

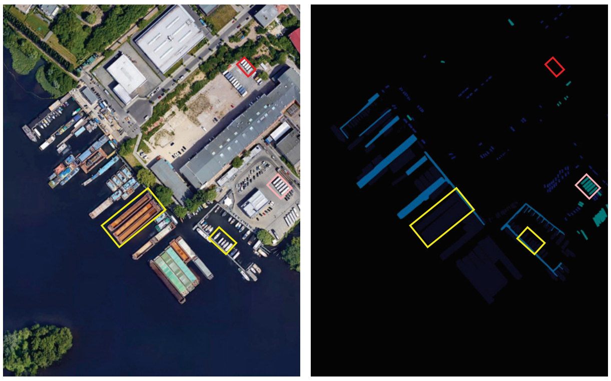


Figure 1. Illustration of large intra-class variance: (1) Foreground intra-class variance that the boats on the right (right yellow bounding box) look more similar to the trucks (pink bounding box) than some other boats (left yellow bounding box). (2) Background intra-class variance (false alarm) that container (red bounding box) in the background looks similar to the trucks (pink bounding box), which is a foreground object.

Compared to general semantic segmentation datasets, remote sensing images have their own challenging problem of large intra-class variance [2, 28, 58]. Since remote sensing images are taken far from the ground, foreground objects are tiny in high resolution images and are represented by only a few pixels. Lack of sufficient information to represent a foreground object leads to a large variation, and the neural network is prone to misclassify different foreground categories. Meanwhile, background class often contains abundant information with high complexity, which causes serious false alarms due to large background intra-class variance [58], as shown in Figure 1.

Current general semantic segmentation methods mainly focus on scale variation [5, 46, 56] while ignoring the above issues. Recent work on geospatial semantic segmentation [58] leverages the symbiotic relation between geospatial

* Indicates equal contribution

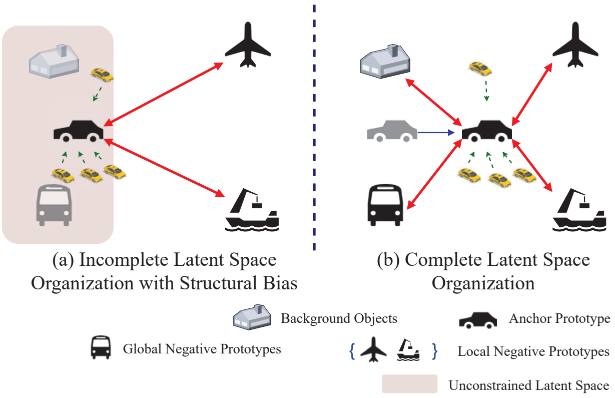

(a) Incomplete Latent Space Organization with Structural Bias
(b) Complete Latent Space Organization
Background Objects
Anchor Prototype
Global Negative Prototypes
Local Negative Prototypes
Unconstrained Latent Space

Figure 2. Left: Incomplete latent space organization with structural bias that only leverages within-image information. The anchor point will be pushed to other unconstrained foreground objects and background objects. Right: Complete latent space organization that leverages global information. The anchor prototype achieves overall optimization in the latent space.

scene and geospatial objects to enhance the discrimination of foreground features and to suppress the false alarm issue due to large background variance. However, they fail to solve the intra-class variance within foreground categories.

In this paper, we propose a **S**parse and **C**omplete latent **O**rganization (**SCO**) to tackle the large intra-class variance issues for both foreground and background class via prototypes, the average of pixels of a category within the image. To enhance the sparsity, we design a prototypical contrastive learning to have prototypes of the same category clustering together, and meanwhile to force prototypes of different categories to be far away from each other. Specifically, given an anchor prototype of a particular category, we treat prototypes of the corresponding category extracted from data augmentation as positive samples while other prototypes from the original image as negative samples. However, this leads to a latent space structural bias, in which only prototypes from partial categories are under constraint since a single image is highly unlikely to contain all categories. Due to the limited size of latent space (the channel of the feature map is restricted), the anchor prototype will be pushed to other categories in the unconstrained latent space, which may undermine the discrimination from those foreground categories and deteriorate the false alarm issue, as shown in Figure 2 (a). To avoid this issue, we strengthen the completeness of the latent space by modeling all foreground categories and hardest (nearest) background objects, as shown in Figure 2 (b), using foreground and background prototype memory banks.

In addition to structuring the latent space, we design a novel data augmentation method, patch shuffle augmentation, to generate the positive samples. Existing data augmentation methods focus on the image-level transformation [10, 41, 52, 54], which provides insufficient variance for pixel-level samples. Since geospatial semantic segmentation is a context-dependent task with highly complicated background information [28, 58], objects of the same category are correlated to different contexts at a large scale (e.g. background of cars in urban scenes varies a lot compared to those in rural scenes). By conducting patch shuffle augmentation, we encourage the semantic information on a pixel to be correlated only to the object itself and its limited surroundings within the patch that is specific to its category (e.g. cars almost always on a road) thus reducing intra-class variance.

Overall, the main contributions of this work are summarized in the following three aspects:

- We propose a sparse and complete latent structure to alleviate the large intra-class variance issue in geospatial segmentation for both foreground and background categories.
- We design a novel patch shuffle augmentation for positive samples generation to restrict context information within the patch, which further reduces foreground intra-class variance and enhances discrimination among objects.
- We evaluate our method through extensive experiments in a large-scale remote sensing dataset, showing our method outperforming state-of-the-art approaches by a large margin.

2. Related Work

Geospatial Semantic Segmentation The success of the semantic segmentation task aligns with the usage of the fully convolutional network (FCN) [32] to conduct pixel-wise classification with end-to-end training, which incorporates more spatial information than convolutional neural network (CNN) [8, 13, 15, 16]. To further improve the performance, various works attempt to retain more spatial information by extending the receptive field and extracting wider and deeper spatial context [1, 4, 36, 38], to extract multi-scale features by designing novel and more robust networks [5, 7, 31, 39, 56], and to introduce new mechanisms such as attention [14, 22, 50, 51] and strip pooling [19] to further exploit the spatial field.

General semantic segmentation methods mainly emphasize on the retainment of spatial information and multi-scale feature extraction, but with few emphasis on the common issues presented in the remote sensing imagery: large intra-class variance and foreground-background imbalance. The task of geospatial semantic segmentation is widely researched in specific application scenarios with some improved techniques [3, 11, 21, 35, 37, 45, 47, 49, 53];

however, few work has been done regarding the common issues in semantic segmentation of remote sensing images. Zheng et al. [58] identified two main unique challenges in geospatial semantic segmentation as stated above, and developed a foreground-aware relation network (FarSeg) to tackle foreground-background related issues via relation-based and optimization-based foreground modeling. Li et al. [28] further pointed out the issue of false alarm and foreground-background imbalance. Our paper focuses on the large intra-class variance issue for both background (identified in [58]) and foreground objects, which is first pointed out in our paper.

Contrastive Learning Contrastive learning aims to learn representations by contrasting similar (positive) data samples against dissimilar (negative) samples often in an unsupervised manner. Various contrastive learning methods develop different strategies to generate instance features. Memory bank was introduced to store the instance class representation vectors [44] and was widely adopted in various tasks [6, 17, 33]. Others explored the approach of in-batch negative sampling [12, 23, 26, 48] as an alternative of the memory bank. These approaches treat each image as an instance where they use augmented images to form positive samples and randomly selected images as negative samples.

Recently, pixel-to-pixel level contrastive learning for semantic segmentation was proposed [20, 42] in a fully supervised manner by introducing a label-based contrastive loss. It enforces pixel embeddings pertained to the same semantic class to be more alike than embeddings from different classes in the latent space. Zhao et al. [57] also proposed three variants of label-based and pixel-level contrastive loss and a two-stage training process of cross-entropy and contrastive losses. These works only focus on discrimination among foreground categories; however, different from the task of general semantic segmentation, geospatial semantic segmentation faces the serious false alarm issue due to large background variance. In our work, we first model the relationship between foreground and background in contrastive learning literature.

3. Method

In this section, we provide a detailed description of our approach as shown in Figure 3. We start by discussing the local prototypical contrastive loss that leverages the information within-image information in Section 3.1; we then discuss the global repulsion force and foreground-background repulsion loss that utilize the global information within the dataset to avoid latent space structural bias (shown in Figure 2) and mitigate false alarm in Section 3.2 and Section 3.3; finally, we discuss the patch shuffle data augmentation, a newly proposed data augmentation for contrastive learning in geospatial semantic segmentation in Section 3.4.

3.1. Within-image Prototypical Contrastive Learning

Ground Truth-Guided Prototype Extraction Given a set of pairs $(\mathbf{x}, \mathbf{y})$ in the training set, we denote $\mathbf{x}$ as the input image and $\mathbf{y}$ as the corresponding ground truth segmentation mask, where $\mathbf{x} \in R^{H\times W\times 3}$ and $\mathbf{y} \in R^{H\times W}$. Semantic segmentation aims to classify each pixel from the input image into a semantic class $c \in C$, where C is the set of all categories appearing in the dataset. Modern semantic segmentation models typically consist of an encoder-decoder network E and a convolutional segmentation head G to output the segmentation score map $S = E \circ G(\mathbf{x}) \in R^{H\times W\times |C|}$. We denote the feature map $\mathbf{F} = E(\mathbf{x})$ with the same spatial dimension H $\times$ W of the ground truth segmentation mask $\mathbf{y}$.

For each image, we extract prototypes from its feature map $\mathbf{F}$ guided by ground truth segmentation masks $\mathbf{y}$. For the class $c \in C_{\mathbf{x}}$, where $C_{\mathbf{x}}$ is the set of foreground categories appearing on the image $\mathbf{x}$, the prototype for class c is computed by taking the average of all the pixels corresponding to class c in ground truth semantic mask $\mathbf{y}$. In formal terms, the prototype of category c, p_c, can be written as:

$$p_c = \frac{1}{|[\![\mathbf{y}_{w,h} = c]\!]|} \sum_{w,h}^{W,H} \mathbf{F}_{w,h} [\![\mathbf{y}_{w,h} = c]\!] \tag{1}$$

where $\mathbf{F}_{w,h}$ denotes the feature vector at the spatial location of (**w**,**h**) in the feature map $\mathbf{F}$ and $[\![\mathbf{y}_{w,h} = c]\!]$ indicates pixels at spatial location (**w**,**h**) in $\mathbf{y}_n$ corresponding to class c and $|\cdot - \cdot|$ is the cardinality.

Local Prototypical Contrastive Loss The core idea of prototypical contrastive learning is to have prototypes of the same category clustering together, and meanwhile to force prototypes of the different categories to be far away from the others. For an anchor prototype $\mathbf{P}_c$, we treat prototypes of class c from data augmentation as our positive samples and other prototypes from the original images as our negative samples. Given a set of pairs $(\mathbf{x}, \mathbf{y})$ in the training set, we perform data augmentation on image $\mathbf{x}$ to generate N augmented pairs $\{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2)..,(\mathbf{x}_n, \mathbf{y}_n)\}$ so that the n-th augmented prototype of class $c \in C_{\mathbf{x}}$ are calculated by:

$$p^n_{c+} = \frac{1}{\left|[\![\mathbf{y}^n_{w,h} = c]\!]\right|} \sum_{w,h}^{W,H} \mathbf{F}^n_{h,w} [\![\mathbf{y}^n_{w,h} = c]\!] \tag{2}$$

where $\mathbf{F}^n = E(\mathbf{x}^n)$ is the feature map of $\mathbf{x}_n$.

To encourage the model to be invariant to multiple transformations, we propose an attraction loss to make augmented prototypes to be as close to corresponding

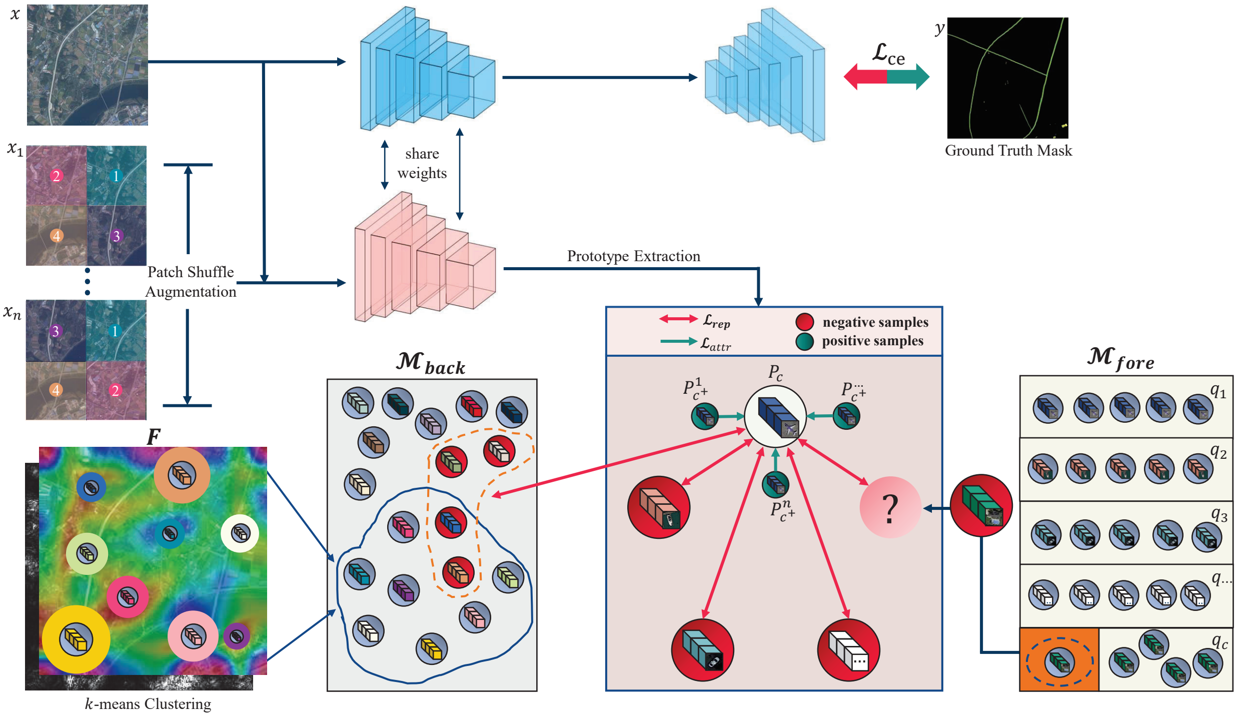


Figure 3. Overview of our approach. The latent space organization is shown in the pink box, where we want our anchor prototype (in the white circle) to be far ways from all negative samples (in red circles) and to be close to positive samples (in green circles). $\mathcal{M}_{fore}$ and $\mathcal{M}_{back}$ denotes foreground and background memory banks storing global prototypes and background prototypes.

the anchor prototype as possible thus reducing the large foreground intra-class variance in remote sensing images, which is defined by:

$$\mathcal{L}_{attr} = \frac{1}{|C_x|} \sum_{c \in C_x} \sum_{i=1}^{n} \left\| (\mathbf{P}_c - p^i_{c+}) \right\|^2 \tag{3}$$

where $\mathbf{p}_c$ is the anchor prototype of class c and $\|\cdot - \cdot\|$ denotes Euclidean distance.

In the meantime, our anchor prototypes are separated from negative prototypes (prototypes of different foreground categories in the same image) to enhance the discrimination. To this end, we propose a local repulsion force within the image, which is defined as:

$$\mathcal{L}^{local}_{rep} = \frac{1}{|C_x|} \sum_{c_i \in C_x} \sum_{\substack{c_j \in C_x \\ c_i \neq c_j}} \frac{1}{\left\| \mathbf{P}_{c_i} - p_{c_j} \right\|^2} \tag{4}$$

So the within-image local prototypical contrastive loss $\mathcal{L}^{local}_{pcl}$ can be expressed as:

$$\mathcal{L}^{local}_{pcl} = \lambda_{attr} \cdot \mathcal{L}_{attr} + \lambda^{local}_{rep} \cdot \mathcal{L}^{local}_{rep} \tag{5}$$

where λ are hyperparamters that balance the loss weights between attraction force and local repulsion force.

3.2. Global Repulsion Loss

Latent space constructed with only within-image information will lead to the latent space structural bias, as shown in Figure 2. In this case, the anchor prototype will be pushed to other global foreground categories (categories existing in the dataset while not appearing in the image), which undermines the discrimination from other global foreground classes. To alleviate this issue and increase the richness of our latent representation, we maintain a foreground prototype memory bank $\mathcal{M}_{fore}$. We adopt the nearest neighbor algorithm to fetch prototypes of global foreground categories, and fulfill them into the latent space as complementary negative samples besides within-image negative samples. Then, similar to the local repulsion force, we impose a global repulsion force to separate the anchor prototype from the additional negative sample. In formal terms, the global repulsion loss $\mathcal{L}^{global}_{rep}$ can be defined as:

$$\mathcal{L}^{global}_{rep} = \frac{1}{|C_x|} \sum_{c_i \in C_x} \sum_{\substack{c_j \in C \\ c_j \notin C_x}} \frac{1}{\left\| \mathbf{P}_{c_i} - NN(\mathbf{P}_{c_i}, c_j, \mathcal{M}_{fore}) \right\|^2} \tag{6}$$

where $\mathbf{P}_{c_i}$ is the anchor prototype of class c_i and $NN(p_{c_i}, c_j, M)$ is the nearest neighbor operator to find the closest prototype of class c_j in memory bank $\mathcal{M}$ to the prototype p_{c_i}, as defined below:

$$NN(p_{c_i}, c_j, \mathcal{M}) = \underset{p_{c_j} \in \mathcal{M}}{\arg\min} \left\| p_{c_i} - p_{c_j} \right\|^2 \qquad (7)$$

Foreground Prototype Memory Bank The foreground memory bank consists of independent queues q to store prototypes of each foreground category. The size of the foreground memory bank is $m_f \cdot d \cdot |C|$, where m_f is a hyperparameter of the length of each queue, d is the dimension of each prototype (i.e., the channel of the feature map where the prototypes are extracted from), and $|C|$ is the number of all the foreground categories in the dataset. The memory bank is first randomly initialized, and we adopt the FIFO (first-in-first-out) method to update these queues at the end of each training step and remove the oldest prototypes.

3.3. Background Prototype

Background class is a unique category in semantic segmentation and other pixel-level tasks. It contains pixels that do not belong to any of the foreground categories. Recent approaches in contrastive learning of semantic segmentation only model the relationship among pixels from foreground classes while ignoring the background pixels. However, in geospatial semantic segmentation, with much more complex context than general semantic segmentation, background class suffers from large intra-class variation. Failing to model the relationship between anchor prototype and background class will aggravate the false alarm issue due to the limited size of latent space (anchor prototypes may be pushed closer to background objects). To this end, we propose to take k-nearest background prototypes as our negative samples while structuring the latent space.

Background Prototype Extraction To extract background prototypes, we adopt average pooling on background areas. However, background class contains more diverse information than foreground categories, so simply taking the average pooling as the prototype is an ineffective and inaccurate representation of the background information. Thus we utilize k-means clustering to obtain a better representation of the background. The objective function can be expressed as follow:

$$\min_{\mu_k} \sum_{w,h}^{W,H} \sum_{n=1}^{k} \| F_{w,h} [\![M_{w,h} = 0]\!] - \mu_n \|^2 \qquad (8)$$

where μ represents the center of the clusters. $M \in R^{H \times W}$ is a binary foreground mask where $M_{w,h} = 0$ represents pixels from the background class. The k-th background prototype $\mathbf{P}_b^k$ are then equal to μ_k, which is the average pooling of all the embedding of pixels belonging to the cluster k.

Foreground-Background Repulsion Loss To reduce false alarm issue in geospatial semantic segmentation and mitigate the latent space structural bias shown in Figure 2, we propose a foreground-background repulsion loss $\mathcal{L}_{rep}^{fb}$ to set the anchor prototype apart from the k nearest background prototypes in the background memory bank $\mathcal{M}_{back}$, which is defined as:

$$\mathcal{L}_{rep}^{fb} = \frac{1}{|C_x|} \sum_{c \in C_x} \sum_{n=1}^{k} \frac{1}{\| \mathbf{P}_c - kNN(\mathbf{P}_c, n, \mathcal{M}_{back}) \|^2} \qquad (9)$$

where $\mathbf{P}_c$ is the anchor prototype of class $c \in C_x$ and $kNN(p_c, n, \mathcal{M})$ is the k-nearest neighbor operator that finds the k^{th} nearest prototypes to the prototype p_c in the memory bank $\mathcal{M}$.

Background Prototype Memory Bank The background memory bank consists of background prototypes $\mathbf{P}_b$ extracted using the k-means clustering algorithm, as mentioned above. The size of the background memory bank is $m_b \cdot d$, where m_b is a hyperparameter representing the number of background prototypes stored in this memory bank and d is the dimension of each prototype. It is randomly initialized at the beginning of the training. We first adopt FIFO for update until all randomly initialized background prototypes are removed. After that, we update the background memory bank by replacing the earliest used (least active) background prototypes after each epoch.

3.4. Patch Shuffle Augmentation

Data augmentation is a critical technique in contrastive learning. By conducting data augmentation, we expect our model can learn better features that are robust and invariant to multiple data transformations of a single sample. Common data augmentation focuses on the image transformations to generate image-level samples for contrastive learning such as rotation, cropping, mixing, and color transformation [10, 41]. However, semantic segmentation is considered as a pixel-level dense prediction task, where we treat pixels as our samples. In this case, conventional data augmentation methods can not provide sufficient variance, and positive samples extracted from those data augmentation are considered relatively easy. To tackle this issue, we design a patch shuffle augmentation for contrastive learning in semantic segmentation, which has been only marginally investigated in the current literature. Specifically, we first split an image into fixed-sized patches and then randomly rearrange these patches to form augmented images. It is worth noting that different augmented images contain dif-

Table 1. Comparison with the state-of-the-art results on iSAID dataset in terms of MeanIoU, best in **bold**. The categories are defined as: ship (Ship), storage tank (ST), baseball court (BC), ground field track (GTF), bridge (Bridge), large vehicle (LV), small vehicle (SV), helicopter (HC), swimming pool (SP), roundabout (RA), soccerball field (SBF), plane (Plane), harbor (Harbor). All the results except for ours are from [28].

Method	**Backbone**	**mIoU (%)**	**IoU per Category (%)**														
			Ship	ST	BD	TC	BC	GTF	Bridge	LV	SV	HC	SP	RA	SBF	Plane	Harbor
DenseASPP [46]	ResNet-50	57.3	55.7	63.5	67.2	81.7	54.8	52.6	34.7	55.6	36.3	33.4	37.5	53.4	73.3	74.7	46.7
RefineNet [31]	ResNet-50	60.2	63.8	58.6	72.3	85.3	61.1	52.8	32.6	58.2	42.4	23.0	43.4	65.6	74.4	79.9	51.1
PSPNet [56]	ResNet-50	60.3	65.2	52.1	75.7	85.6	61.1	60.2	32.5	58.0	3.0	10.9	46.8	68.6	71.9	79.5	54.3
OCNet-(ASP-OC) [51]	ResNet-50	40.2	47.3	40.2	44.4	65.0	24.1	29.9	2.71	46.3	13.6	10.3	34.6	37.9	41.4	68.1	38.0
EMANet [30]	ResNet-50	55.4	63.1	68.4	66.2	82.7	56.0	18.8	42.1	58.2	41.0	33.4	38.9	46.9	46.4	78.5	47.5
CCNet [22]	ResNet-50	58.3	61.4	65.7	68.9	82.9	57.1	56.8	34.0	57.6	38.3	31.6	36.5	57.2	75.0	75.8	45.9
EncodingNet [55]	ResNet-50	58.9	59.7	64.9	70.0	84.2	55.2	46.3	36.8	57.2	38.7	34.8	42.4	59.8	69.8	76.1	48.0
SemanticFPN [27]	ResNet-50	62.1	68.9	62.0	72.1	85.4	54.1	48.9	44.9	61.0	48.6	37.4	42.8	70.2	58.6	84.7	54.9
UPerNet [27]	ResNet-50	63.8	68.7	71.0	73.1	85.5	55.3	57.3	43.0	61.3	45.6	30.3	45.7	68.7	75.1	84.3	56.2
SFNet [29]	ResNet-50	64.3	68.8	71.3	72.1	85.6	58.8	60.9	43.1	62.9	47.7	30.4	47.8	69.8	75.1	83.1	57.3
GSCNN [40]	ResNet-50	63.4	65.9	71.2	72.6	85.5	56.1	58.4	40.7	63.8	**51.1**	33.8	48.8	58.5	72.5	83.6	54.4
RANet [34]	ResNet-50	62.1	67.1	61.3	72.5	85.1	53.2	47.1	45.3	60.1	49.3	38.1	41.8	70.5	58.8	83.1	55.6
FarSeg [58]	ResNet-50	63.7	65.4	61.8	77.7	86.4	62.1	56.7	36.7	60.6	46.3	35.8	51.2	71.4	72.5	82.0	53.9
PFSegNet [28]	ResNet-50	66.9	70.3	74.7	77.8	87.7	62.2	59.5	45.2	**64.6**	50.2	37.9	50.1	71.7	75.4	85.0	59.3
SCO (Ours)	ResNet-50	**69.1**	**74.7**	**75.0**	**78.5**	**89.0**	**66.3**	**63.6**	**46.3**	63.0	46.9	**41.1**	**56.5**	**73.3**	**84.0**	**85.3**	**64.3**

ferent patch arrangements without repetition so that augmented images are different from each other and from the original image. Meanwhile, we split the corresponding ground truth labels into patches and rearrange them into the same pattern to match the corresponding augmented images. Compared to image-level data augmentation, patch shuffle augmentation leverages the inherent properties of geospatial semantic segmentation: 1) semantic segmentation is a context-dependent task, where semantic information on a pixel is strongly correlated with its surrounding context. 2) the background is much more complex in the remote sensing images where the same foreground objects have very different contexts at a large scale, intensifying foreground intra-class variance. By doing patch shuffle augmentation, we restrict the correlation of foreground objects to the limited information within the patch specific to its category to further reduce large intra-class variance within foreground objects and enhance the discrimination.

4. Experiments

In this section, we conduct extensive experiments to validate the effectiveness of our approach on iSAID [43], a large scale remote sensing dataset. We first start with the description of our experimental setup and implementation details in Section 4.1. We then demonstrate our experimental results compared to existing state-of-the-art methods in Section 4.2. Finally, we show our results from comprehensive ablation studies for our method in Section 4.3.

4.1. Experiments Setup and Implementation Details

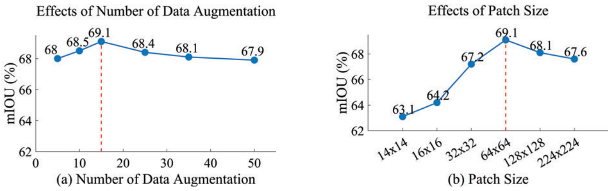
Dataset We evaluate our method on a commonly used large scale remote sensing dataset iSAID. iSAID consists of 2,806 remote sensing images acquired by multiple satellites and sensors with original image sizes ranging from 800×800 pixels to 4000×13000 pixels. As one of the largest datasets for geospatial semantic segmentation on remote sensing imagery, iSAID contains 655,451 object instances for 15 categories across 2,806 high-resolution images that are densely annotated. For predefined training set, validation set, and test set, iSAID dataset has 1411, 458, and 937 images respectively.

Table 2. Ablation study on the effectiveness of modules on iSAID *val* set. Starting from baseline, the proposed modules are gradually added for the module analysis.

$\mathcal{L}_{ce}$	$\mathcal{L}_{attr}$	$\mathcal{L}_{rep}^{local}$	$\mathcal{L}_{rep}^{global}$	$\mathcal{L}_{rep}^{fb}$	**mIoU**
✓					63.7
✓	✓				65.1
✓		✓			64.8
✓	✓	✓			65.4
✓	✓	✓	✓		67.8
✓	✓	✓		✓	67.4
✓	✓	✓	✓	✓	**69.1**

Table 3. Ablation study of different augmentation methods on our approach.

Data Augmentation Method	**mIoU**
Cutout	67.8
Mixup	68.1
Manifold Mixup	68.3
CutMix	68.5
Patch Shuffle Augmentation	**69.1**



(a) Number of Data Augmentation

(b) Patch Size

Figure 4. Ablation studies on the number of data augmentation and the size of patches in patch shuffle data augmentation.

Evaluation metric We report the performance using mean Intersection-over-Union (mIoU) in percentage, which calculates the average of Intersection-over-Union (IoU) of all classes including the background class. We adopt mIoU as our main metric unless specifically specified, which is a common practice.

Compared Methods We benchmark our model against the latest state-of-the-art methods in remote sensing semantic segmentation: FarSeg [58] and PFSegNet [28]. Additionally, we compare our method to state-of-the-art approaches for general semantic segmentation methods on remote sensing datasets.

Implementation Details Our method is implemented with PyTorch. We follow the implementation of the state-of-the-art method, FarSeg [58] and adopt it as our baseline. Following the same setting of FarSeg, we adopt ResNet-50 [18] pretrained on ImageNet [9] as our backbone within all experiments. We use the "poly" learning rate policy where current learning rate equals to the base one multiplying $(1 - \frac{step}{max\_step}^{power})$. Models are all trained with 60,000 iterations using the above "poly" learning rate policy and we set base learning rate to 0.007 and power to 0.9. We train our network using SGD with weight decay of 0.0001 and momentum of 0.9. The length of foreground memory bank, m_f, is set to 10 and the length of background memory bank, m_b, is set to 256. To balance the number of foreground and background prototypes in latent space, we take 30 background prototypes. The number of the patch shuffle augmentation (positive samples) is set to 15. Also, we set the patch size equal to 64 × 64 in the patch shuffle augmentation. We train our model on 2 NVIDIA 2080 Ti GPU for all datasets and models. We set the batch size to 4 and in total 8 images are allocated on two GPUs. We adopt the synchronized batch normalization for multi GPU training. We also use apex to speed up the training and the opt_level O1 of mixed precision. We crop the image into the size of 896×896 using a sliding window striding 512 pixels.

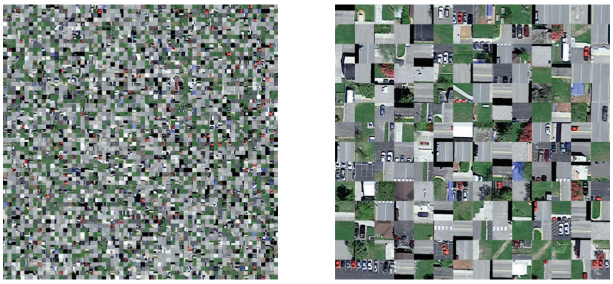


Figure 5. Patch Size = 16 × 16 (left), Patch Size = 64 × 64 (right)

4.2. Comparison to state-of-the-art

Table 1 demonstrates the quantitative results of the overall mIoU and IoU per category. The results suggest that our method outperforms its closest contender, PFSegNet [28] by a significant margin of 2.2 % increasing from 66.9% to 69.1%. Particularly, our method shows significant improvements in categories of ship, baseball court, ground field track, swimming pool, soccerball field, and harbor. Under these categories, baseball court - ground field track, soccerball field - ground field track, ship-harbors are pairs that are visually similar, and often appear together and easy to be misclassified. We can show numerically that our method successfully enhances the discrimination between these categories.

4.3. Ablation Studies

In this section, we perform ablation studies on several different aspects to analyze our proposed modules and some important hyperparameters setting in our method.

Effects of local (within-image) contrastive loss We evaluate the effectiveness of local (within-image) contrastive loss in Table 2, which is composed of $\mathcal{L}_{attr}$ and $\mathcal{L}_{rep}^{local}$. From the table, we can see that adding $\mathcal{L}_{attr}$ and $\mathcal{L}_{rep}^{local}$ separately resulting in the increase in performance by 1.4% and 1.1% compared to the baseline with only pixel-wise cross entropy loss ($\mathcal{L}_{ce}$). After combining the above loss together, the performance reaches 65.4%, which is 1.7% higher than the baseline and 0.3% and 0.6% higher than adding two losses separately.

Effects of global repulsion loss To demonstrate the effectiveness of our global repulsion loss ($\mathcal{L}_{rep}^{global}$), we evaluate performance of local (within-image) contrastive loss with global repulsion loss ($\mathcal{L}_{rep}^{global}$). From Table 2, we can see that the performance with $\mathcal{L}_{rep}^{global}$ surpass that without $\mathcal{L}_{rep}^{global}$. In particular, by adding $\mathcal{L}_{rep}^{global}$ to the local (within-image) contrastive loss, we obtain a significant increase of 2.4% with respect to the mIoU (from 65.4% to 67.8%).

Effects of foreground-background repulsion loss We evaluate the performance of local (within-image) contrastive loss with foreground-background repulsion force

https://github.com/Z-Zheng/FarSeg

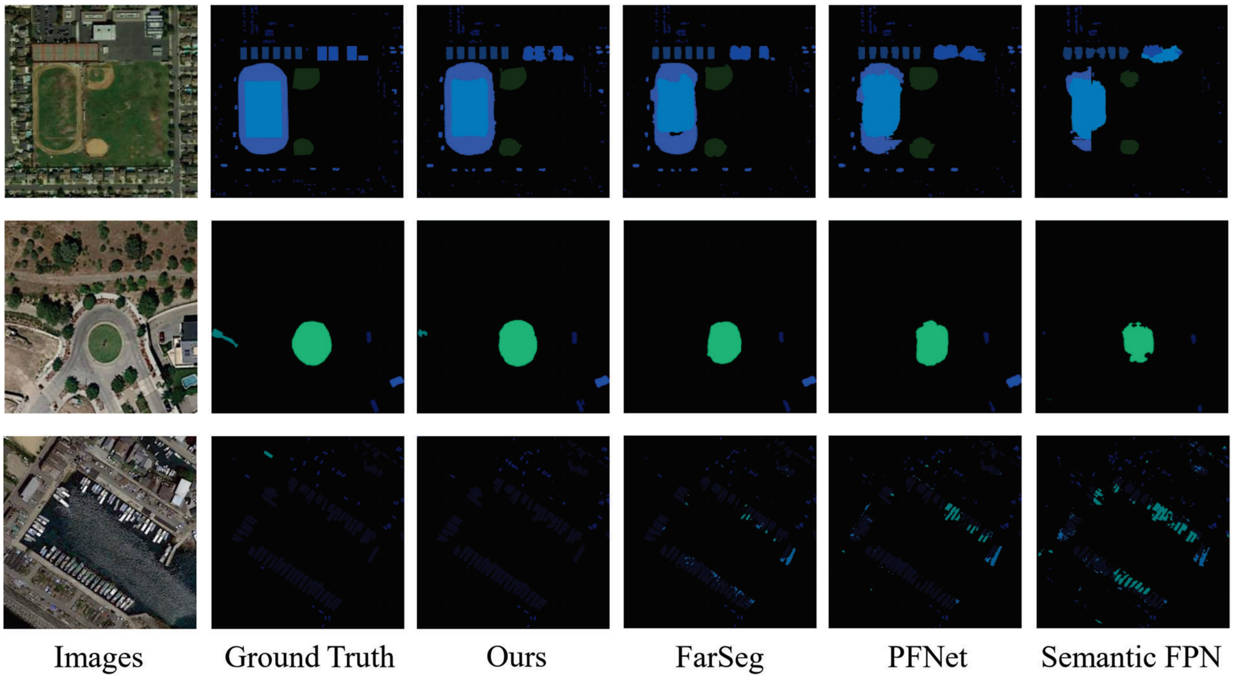


Figure 6. Visualization results on iSAID validation dataset.

($\mathcal{L}_{rep}^{fb}$) and Table 2 summarizes the experimental results on iSAID validation set. The mIoU of that adds the $\mathcal{L}_{rep}^{fb}$ is 2.0% higher than that without it (from 65.4% to 67.4%), which demonstrates that taking background information into account can significantly boost the performance.

Effectiveness of patch shuffle data augmentation In Table 3, we compare the patch shuffle augmentation method with state-of-the-art data augmentation methods including Cutout [10], Manifold Mixup [41], Mixup. [54], and CutMix [52] based on our approach. Compared to previous work, our patch shuffle augmentation prominently exceeds its closest contender CutMix with an increase from 68.5% to 69.1 %, showing the effectiveness of our approach under remote sensing images.

Design choice of patch shuffle augmentation In Figure 4 (a), we evaluate the best number of data augmentation. We observe that the optimal number is 15, which is approximately the sum of negative foreground prototypes. After generating more augmented images, the performance encounters a drop. The reason is there will be a stronger attraction force to guide the movement of prototypes in the latent space that undermines the overall structure. We further evaluate the best size of patches in patch shuffle augmentation in Figure 4. (b). Our idea is to have the patch to cover object itself with local context (Figure 5 Right) but not too small to cut objects into unrecognizable fragments (Figure 5 Left). We find that the best patch size is 64 × 64. When patch size is too small (e.g., 14x14, 16x16), the foreground objects are split into fractions. Thus the model learns little about foreground objects but noise (consider the extreme case when patch size is 1x1), which harms the performance. On the other hand, if patches are generated on a larger scale, we do not add enough variance for contrastive learning.

Visualization Figure 6 demonstrates the visualization of our model compared to several existing state-of-the-art methods including PFNet [28], FarSeg [58], and a baseline semantic segmentation Semantic FPN [27] on iSAID validation dataset. Overall, our method has a better segmentation results handling easily misclassified objects and complicated context.

5. Conclusion

In this paper, we propose a sparse and complete latent organization for geospatial semantic segmentation in remote sensing images to tackle the large intra-class variance issue in both foreground and background categories jointly. We further design a novel data augmentation method for geospatial semantic segmentation that further reduces intra-class variance and enhances the discrimination among objects. Lastly, we perform extensive evaluations on a large scale remote sensing dataset to demonstrate the effectiveness of our model.

Acknowledgements We would like to thank Prof. David Fouhey for his valuable advice and insightful discussions.

References

[1] 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. 2

[2] Haohao Bai, Tingzhu Bai, Wei Li, and Xun Liu. A building segmentation network based on improved spatial pyramid in remote sensing images. *Applied Sciences*, 11:5069, 2021. 1

[3] Alexey Bokhovkin and Evgeny V. Burnaev. Boundary loss for remote sensing imagery semantic segmentation. *ArXiv*, abs/1905.07852, 2019. 2

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs, 2016. 2

[5] 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, 2017. 1, 2

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. 3

[7] Wuyang Chen, Ziyu Jiang, Zhangyang Wang, Kexin Cui, and Xiaoning Qian. Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[8] Dan Cirean, Alessandro Giusti, Luca Maria Gambardella, and Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. *Proceedings of Neural Information Processing Systems*, 25, 01 2012. 2

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 7

[10] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 2, 5, 8

[11] Matt Dickenson and Lionel Gueguen. Rotated rectangles for symbolized building footprint extraction. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 215–2153, 2018. 2

[12] Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning, 2017. 3

[13] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1915–1929, 2013. 2

[14] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation, 2019. 2

[15] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from rgb-d images for object detection and segmentation, 2014. 2

[16] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation, 2014. 2

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2020. 3

[18] 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 (CVPR)*, 2016. 7

[19] Qibin Hou, Li Zhang, Ming-Ming Cheng, and Jiashi Feng. Strip pooling: Rethinking spatial pooling for scene parsing, 2020. 2

[20] Hanzhe Hu, Jinshi Cui, and Liwei Wang. Region-aware contrastive learning for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16291–16301, October 2021. 3

[21] Bo Huang, Bei Zhao, and Yimeng Song. Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery. *Remote Sensing of Environment*, 214:73–86, 2018. 2

[22] Zilong Huang, Xinggang Wang, Yunchao Wei, Lichao Huang, Humphrey Shi, Wenyu Liu, and Thomas S. Huang. Ccnet: Criss-cross attention for semantic segmentation, 2020. 2, 6

[23] Xu Ji, João F. Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation, 2019. 3

[24] Michael Kampffmeyer, Arnt-Børre Salberg, and Robert Jenssen. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 680–688, 2016. 1

[25] Ronald Kemker, C. Salvaggio, and Christopher Kanan. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *Isprs Journal of Photogrammetry and Remote Sensing*, 145:60–77, 2018. 1

[26] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning, 2021. 3

[27] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollar. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 6, 8

[28] Xiangtai Li, Hao He, Xia Li, Duo Li, Guangliang Cheng, Jianping Shi, Lubin Weng, Yunhai Tong, and Zhouchen Lin. Pointflow: Flowing semantics through points for aerial image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4217–4226, 2021. 1, 2, 3, 6, 7, 8

[29] Xiangtai Li, Ansheng You, Zhen Zhu, Houlong Zhao, Maoke Yang, Kuiyuan Yang, Shaohua Tan, and Yunhai Tong. Semantic flow for fast and accurate scene parsing. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 775–793, Cham, 2020. Springer International Publishing. 6

[30] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 6

[31] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-

resolution semantic segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5168–5177, 2017. 2, 6

[32] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015. 2

[33] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations, 2019. 3

[34] Lichao Mou, Yuansheng Hua, and Xiao Xiang Zhu. A relation-augmented fully convolutional network for semantic segmentation in aerial scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 6

[35] Lichao Mou and Xiao Xiang Zhu. Vehicle instance segmentation from aerial image and video using a multitask learning residual fully convolutional network. *IEEE Transactions on Geoscience and Remote Sensing*, 56(11):6699–6711, 2018. 2

[36] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation, 2015. 2

[37] Doruk Oner, Mateusz Koziński, Leonardo Citraro, Nathan C. Dadap, Alexandra G. Konings, and Pascal Fua. Promoting connectivity of network-like structures by enforcing region separation, 2020. 2

[38] Franz Rottensteiner, Gunho Sohn, Jaewook Jung, Markus Gerke, Caroline Baillard, Sébastien Bénitez, and U Breitkopf. The isprs benchmark on urban object classification and 3d building reconstruction. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, I-3, 07 2012. 2

[39] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions, 2019. 2

[40] Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. Gated-scnn: Gated shape cnns for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 6

[41] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6438–6447, Long Beach, California, USA, 09–15 Jun 2019. PMLR. 2, 5, 8

[42] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. *arXiv preprint arXiv:2101.11939*, 2021. 3

[43] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–37, 2019. 6

[44] Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance-level discrimination, 2018. 3

[45] Yongyang Xu, Liang Wu, Zhong Xie, and Zhanlong Chen. Building extraction in very high resolution remote sensing imagery using deep learning and guided filters. *Remote Sensing*, 10(1), 2018. 2

[46] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3684–3692, 2018. 1, 6

[47] Naisen Yang and Hong Tang. Semantic segmentation of satellite images: A deep learning approach integrated with geospatial hash codes. *Remote Sensing*, 13(14), 2021. 2

[48] Mang Ye, Xu Zhang, Pong C. Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature, 2019. 3

[49] Jiangye Yuan. Learning building extraction in aerial scenes with convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(11):2793–2798, 2018. 2

[50] Yuhui Yuan, Xiaokang Chen, Xilin Chen, and Jingdong Wang. Segmentation transformer: Object-contextual representations for semantic segmentation, 2021. 2

[51] Yuhui Yuan, Lang Huang, Jianyuan Guo, Chao Zhang, Xilin Chen, and Jingdong Wang. Ocnet: Object context network for scene parsing, 2021. 2, 6

[52] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *International Conference on Computer Vision (ICCV)*, 2019. 2, 8

[53] Ce Zhang, Isabel Sargent, Xin Pan, Huapeng Li, Andrew Gardiner, Jonathon Hare, and Peter Atkinson. Joint deep learning for land cover and land use classification. *Remote Sensing of Environment*, 221:173–187, 02 2019. 2

[54] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 2, 8

[55] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 6

[56] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 1, 2, 6

[57] Xiangyun Zhao, Raviteja Vemulapalli, Philip Mansfield, Boqing Gong, Bradley Green, Lior Shapira, and Ying Wu. Contrastive learning for label-efficient semantic segmentation, 2021. 3

[58] Zhuo Zheng, Yanfei Zhong, Junjue Wang, and Ailong Ma. Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4096–4105, 2020. 1, 2, 3, 6, 7, 8