Segmentation of VHR EO Images using Unsupervised Learning

Sudipan Saha[1][0000−0002−9440−0720], Lichao Mou[1,2][0000−0001−8407−6413], Muhammad Shahzad[1][0000−0002−8278−9118], and Xiao Xiang Zhu[1,2][0000−0001−5530−3613]

[1] Data Science in Earth Observation, Technical University of Munich, Germany
{sudipan.saha,muhammad.shahzad}@tum.de
[2] Remote Sensing Technology Institute, German Aerospace Center (DLR), Germany
{lichao.mou,xiaoxiang.zhu}@dlr.de

Abstract. Semantic segmentation is a crucial step in many Earth observation tasks. Large quantity of pixel-level annotation is required to train deep networks for semantic segmentation. Earth observation techniques are applied to varieties of applications and since classes vary widely depending on the applications, therefore, domain knowledge is often required to label Earth observation images, impeding availability of labeled training data in many Earth observation applications. To tackle these challenges, in this paper we propose an unsupervised semantic segmentation method that can be trained using just a single unlabeled scene. Remote sensing scenes are generally large. The proposed method exploits this property to sample smaller patches from the larger scene and uses deep clustering and contrastive learning to refine the weights of a lightweight deep model composed of a series of the convolution layers along with an embedded channel attention. After unsupervised training on the target image/scene, the model automatically segregates the major classes present in the scene and produces the segmentation map. Experimental results on the Vaihingen dataset demonstrate the efficacy of the proposed method.

Keywords: Segmentation · Unsupervised learning · Earth observation.

1 Introduction

Plethora of satellites equipped with High Resolution sensors have been launched in the last decade. Additionally, unmanned aerial vehicles (UAVs) are now widely available, thus generating a large volume of images for detailed Earth observation (EO). Automatic parsing of such images is useful for various applications, including disaster management [19] and urban monitoring [21]. The last decade also witnessed the development of various deep learning methods that outperform the previous methods on EO images. Their superior performance is attributed to their capability to learn complex spatial features from large volume of labeled data.

A crucial step in understanding EO images is semantic segmentation, i.e., assigning a semantically meaningful class/category to each pixel in the image. Semantic segmentation for natural images has progressed fast exploiting availability of vast training data and superior performance of convolutional neural networks (CNNs). While CNN based methods have been adopted to the EO images [17], their applicability in the EO has been limited due to the lack of labeled data [13].

While the lack of training data offers a hurdle for segmentation of VHR EO images, their large spatial size offers an advantage. Being representative of a geographical area, EO scenes are generally large, e.g., each scene in the Potsdam dataset (part of ISPRS semantic labeling dataset [1]) has a size of 6000×6000 pixels. While a typical image in the computer vision datasets [9] rarely captures multiple instances of the same object in the same image, EO images may capture even up to hundreds of instances of the same object (e.g., building) in a single image/scene. Though most state-of-the-art semantic segmentation methods are supervised [14, 18], there are few methods based on the concept of deep clustering [22] that can work in unsupervised manner. The unsupervised paradigm has also been extended for remote sensing images in context of multitemporal analysis by exploiting temporal consistency between images in a time-series [21]. Inspired by their success, we propose an semantic segmentation method that can learn the segmentation clusters in unsupervised manner from a single image using a lightweight model. The proposed method employs deep clustering and contrastive learning and produces segmentation map such that each label corresponds to a semantically meaningful entity. The proposed method is trained on a single scene only. The key contributions of this paper are as follows:

1. Proposing an unsupervised segmentation method that can be trained on a single unlabeled EO scene.
2. Incorporating deep clustering and contrastive learning in same framework for unsupervised segmentation.

Related works are discussed in Section 2. We define the problem statement and detail the proposed single-scene segmentation method in Section 3. Experimental validation is presented Section 4. We conclude the paper and discuss scope of future research in Section 5.

2 Related Work

Considering relevance to our work, in this Section we briefly discuss deep segmentation, unsupervised and self-supervised learning.

2.1 Deep segmentation

Most supervised methods for semantic segmentation rely on pixelwise classification using a classifier that is trained using available reliable training pixels. Many deep learning based supervised methods have been proposed in the literature [18,

25, 4]. Most deep learning based approaches have an architecture consisting of an encoder and a decoder to achieve discrimination at pixel level. Several supervised segmentation methods have been proposed for EO images [23, 25, 15–17, 7]. The supervised methods require a significant amount of training data. To address the scarcity of training data, [13] proposed a segmentation method that trains the model using incomplete annotations. Unsupervised deep clustering for multi-temporal EO segmentation is proposed in [21]. However, their method focuses on only two classes per target scene.

2.2 Unsupervised and self-supervised learning

Supervised methods are limited in many applications due to the difficulty of labeling data. This has motivated machine learning researchers to develop unsupervised and self-supervised methods. Some works [10, 8] use pre-text tasks like image rotation to learn unsupervised semantic feature. Similar approaches have been adopted in EO, e.g., learning to rearranging randomly shuffled time-series images [20]. In addition to pre-text tasks, some methods rely on deep clustering by jointly learning the cluster assignment and weights of the deep network [3]. Given a collection of unlabeled inputs, deep clustering divides them into groups in terms of inherent latent semantics. Many variants of deep clustering exists, e.g., using convolutional autoencoder [11]. Contrastive methods function by bringing the representation of positive pairs closer while spreading representations of negative pairs apart [5, 6, 24]. [2] demonstrated that the unsupervised methods learn useful semantic features even with a single-image input.

Our work is inspired from the above-mentioned works on unsupervised and self-supervised learning, especially deep clustering [3]. Moreover similar to [2], our work focuses on single scene.

3 Proposed method

Let X be a VHR EO image/scene of spatial dimension $R \times C$ pixels where R and C are much larger than usual image sizes in computer vision (224×224). Originally we do not have label corresponding to any pixel in X. Our goal is to obtain segmentation map corresponding to X, i.e., we want to assign labels to each pixel in X such that those labels are semantically meaningful. We accomplish this by using self-supervised learning that do not require any external label. Smaller patches of size $R' \times C'$ ($R' < R$ and $C' < C$) are extracted from X. Let us assume that a total $\mathcal{B}_{total}$ patches can be extracted from X. One training batch involves only a batch of $\mathcal{B}$ patches sampled from $\mathcal{B}_{total}$, denoted as $\mathcal{X} = \{x^1, ..., x^{\mathcal{B}}\}$. x^b is processed using a deep clustering loss, thus simultaneously refining the weights of the model and the segmentation map. We use a lightweight model that uses a series of convolution layers and a channel attention. At the end of $\mathcal{I}$ epochs, the trained model can be applied to X to obtain its segmentation map. Furthermore we demonstrate that the self supervised network trained on X can be directly applied to another spatially disjoint but semantically similar

scene Z, without requiring any further training or fine tuning. The proposed method is shown in Algorithm 1.

3.1 Basic architecture

Usually the number of images used in training a deep network is in the order of tens of thousands. Compared to that the number of patches that can be used for unsupervised learning from a single EO image/scene is limited. Actual number is a function of R, C, R', C'. Considering this we design a lightweight network consisting of only few (L) convolution layers. Convolving the input patches through convolution layers allow us to capture the pixelwise semantics. Successive convolution layers capture increasingly complex spatial details. The spatial size of the input images are preserved through the successive layers by exclusion of stride or pooling operation from the architecture. Output from each convolution layer is processed through activation function (Rectified Linear Unit - ReLU) to introduce non-linearity and by Batch Normalization layer. The weights of the network, denoted as $\mathbb{W}^1, ..., \mathbb{W}^L$ are initialized using a suitable initialization method and are trainable using a set of loss functions that do no require any external label. While the kernel numbers are fixed as 64 in all layers (any other number could be chosen), last layer projects feature to K-dimensional space, where K is an approximation of desired number of classes. In addition to the convolution layers, a channel attention mechanism is used. Channel attention mechanism has demonstrated potential in improving the performance of CNNs [12]. We apply the channel attention just before the final 1×1 convolution layer. The channel attention is designed following [26], i.e., by joint use of both average pooling and max-pooling.

After processing a patch x^b in $\mathcal{X}$ through the network, for each input pixel x^b_n ($n = 1, ..., N$), we obtain deep features y^b_n of dimension K. The basic architecture (showing only convolution and attention layers) is shown in Table 1.

Table 1. Basic architecture of the network. Activation function and batch-normalization is excluded for sake of brevity.

Layer	**Kernel**	**Kernel size**	**Stride**
convolution	64	(3,3)	1
convolution	64	(3,3)	1
convolution	64	(3,3)	1
convolution	64	(3,3)	1
convolution	64	(3,3)	1
Attention	NA	NA	NA
convolution	K	(1,1)	1

Algorithm 1 Self-supervised training for single-scene segmentation

Input: A VHR EO image/scene X
Output: A lightweight model that can segment X or any other similar scene
Initialize $\mathbb{W}^1, ..., \mathbb{W}^L$
Extract $\mathcal{B}_{total}$ patches from X
for $i \leftarrow 1$ to $\mathcal{I}$ **do**
    while all $\mathcal{B}_{total}$ patches are not sampled **do**
        Sample $\mathcal{B}$ patches from $\mathcal{B}_{total}$ patches, denoted as $\mathcal{X} = \{x^1, ..., x^{\mathcal{B}}\}$
        for $j \leftarrow 1$ to $\mathcal{J}$ **do**
            for $b \in \mathcal{B}$ **do**
                for n-th pixel in x^b **do**
                    Compute feature y_n^b
                    Compute pseudo-label c_n^b
                    Compute loss ℓ_n^b
                Compute $\mathcal{L}^b$ by considering all n in x^b
            Compute $\mathcal{L}$ by considering $b = 1, ...B$
            Shuffle $\mathcal{X}$ to $\mathcal{X}'$
            Compute contrastive loss $\mathcal{L}'$
            Update $\mathbb{W}^1, ..., \mathbb{W}^L$ with $\mathcal{L}$ and $\mathcal{L}'$

3.2 Pseudo label assignment

Semantically similar inputs (in our case, pixels) generate strong activations in similar feature. Following this principle, we can assign each pixel to a label/cluster by using argmax classification. More specifically, label c_n^b for an input pixel x_n^b is estimated by selecting the feature in which y_n^b has maximum value. Representing the k-th feature of y_n^b as $y_n^b(k)$, c_n^b is obtained as following:

$$c_n^b = \arg\max_{k \in K} y_n^b(k) \tag{1}$$

Considering that the last layer has K different neurons, c_n^b can take at most K values. Thus this is equivalent to clustering with K number of classes. Please note that we assign label to each pixel for each patch in the training batch, i.e., our deep clustering process works at pixel level.

3.3 Deep clustering

Training the self-supervised network is composed of two processes, assignment of labels to each pixel, estimation of loss based on assigned labels. This process continues in iteration by reassigning the weights and re-estimating loss. Label assignment of each pixels needs to be meaningfully refined so that semantic information of the image is captured and label assignment converges with iterations, performed for $\mathcal{J}$ iterations for each batch. Towards this, we compute the cross-entropy loss between the continuous-valued deep feature representation y_n^b and the discrete valued estimated labels c_n^b.

$$\ell_n^b = \text{crossentropy}(y_n^b, c_n^b) \tag{2}$$

In practice the loss term $\mathcal{L}$ is computed by considering all pixels in a patch and all patches in a training batch.

3.4 Contrastive learning

Contrastive learning is employed to encourage the network to produce dissimilar feature for different input. Though we do not have any negative samples under the unsupervised setting, we shuffle the batch of patches $\mathcal{X}$ to $\mathcal{X}'$. This implies that $b-$th patch in $\mathcal{X}$ (x^b) and $\mathcal{X}'$ ($x^{b'}$) are unpaired. Thus we compute negative absolute error loss for each input pixel x_n^b and $x_n^{b'}$:

$$\ell_n^{b'} = -||(y_n^b - y_n^{b'})||_1 \tag{3}$$

Loss term $\mathcal{L}'$ is computed as mean of exponentials of $\ell_n^{b'}$ over all considered pixels for all patches in the batch.

The sum of loss term $\mathcal{L}$ and $\mathcal{L}'$ is used to update the model weights $\mathbb{W}^1, ..., \mathbb{W}^L$. Note that the computation of $\mathcal{L}$ does not require any external label and hence the mechanism is unsupervised.

4 Experiments

4.1 Dataset

We use the Vaihingen dataset that is a benchmark dataset for semantic segmentation provided by the International Society for Photogrammetry and Remote Sensing (ISPRS) [1]. The images are collected over the city of Vaihingen with a spatial resolution of 9 cm/pixel. Each image in the dataset convers an average area of 1.38 square km. Three bands are available - near infrared (NIR), red (R), and green (G). Additionally digital surface models (DSMs) are available that are not used in this work. In total, six land-cover classes are considered: impervious surface, building, low vegetation, tree, car, and clutter/background. As used in [13], we use image IDs 11, 15, 28, 30, and 34 as test set. Since we need only a single scene for training, image ID 1 is used for training the unsupervised model. Our result is shown as an average of three runs with different seeds.

4.2 Comparison method

To the best of our knowledge, our work is first attempt to obtain multi-class segmentation maps from VHR images in unsupervised manner. Hence, comparison to supervised paradigms is unfair and instead comparison needs to be performed with methods that can work in label-constrained manner. Considering this, we compared the proposed method to FEature and Spatial relaTional regulArization (FESTA) [13] that trains semantic segmentation model based on incomplete annotations. For comparison, we trained the FESTA model in [13] using image ID 1 and using different number of training points. Please note that inspite of working on the incomplete annotations, the method in [13] has access to some labeled point during training, while the proposed method does not use any annotated data during training.

4.3 Result

The training process is accomplished with $\mathcal{I} = 2$, $\mathcal{J} = 50$, and number of kernels in the final layer $K = 8$, a value considerably close to number of classes in the images (6). Choice of $K = 8$ is the only component of the proposed method, where prior knowledge about the target scene is used. Since our approach is unsupervised, it is not possible to automatically know the name of each class unlike supervised segmentation. Here we have assigned each class a name as per their overlap with the classes in the reference map.

Multi-class segmentation: We show the result for image ID 11 and 15 in Figure 1. Input images are shown in first column. Second and third columns show the reference segmentation masks and the result obtained by proposed method, respectively. We observe that in both cases the two major classes - buildings (blue) and impervious surfaces (white) are satisfactorily detected. A significant overlap is observed between low vegetation (cyan) and trees (green), especially in image ID 15. Considering the unsupervised nature of the proposed method, it is difficult for it to know the real class divisions as desired in the reference map. Thus it identifies similar (as per spectral characteristics) low vegetation and trees as same class.

The quantitative result averaged over 5 test tiles are shown in Table 2 in terms of F1 score and Intersection-Over-Union (IoU). The proposed method clearly outperforms FESTA [13] for 5 point and 20 point annotations. This result shows that proposed method, despite not using any annotated data during training, can outperform existing state-of-the-art method when using few annotated points. When FESTA uses all annotated points in tile 1, the proposed method still outperforms FESTA, however the margin reduces.

Binary segmentation: In many urban applications, it is more important to know only the man-made urban structures than low vegetation and trees. Considering that, we also show the performance of the proposed method as binary segmentation map, considering two classes: buildings as one class (white) and rest as one (black). For image 11, Figure 2(a) shows the reference binary segmentation map and Figure 2(b) shows the binary segmentation obtained by the proposed method. It is visually evident that there is high match between the binary reference map and the binary segmentation map.

Table 2. Quantitative comparison of the proposed method to FESTA [13].

Method	F1 score	IOU
Proposed Unsupervised	0.43	0.30
FESTA 5 points	0.26	0.16
FESTA 10 points	0.32	0.23
FESTA All points	0.41	0.28

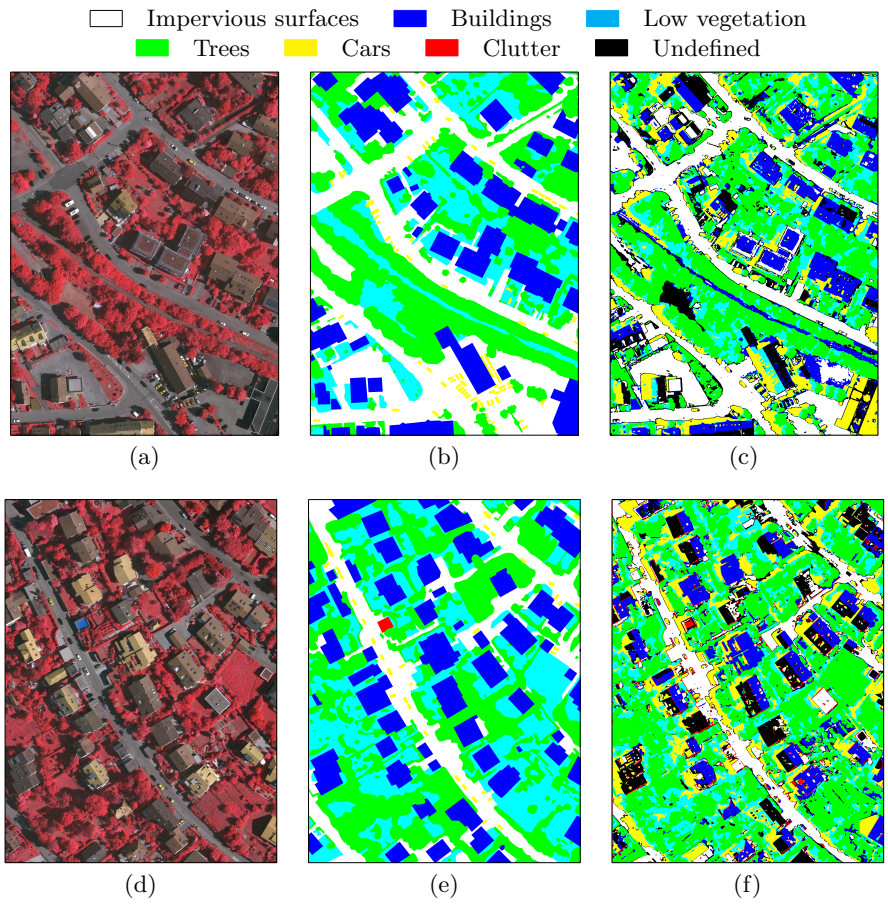
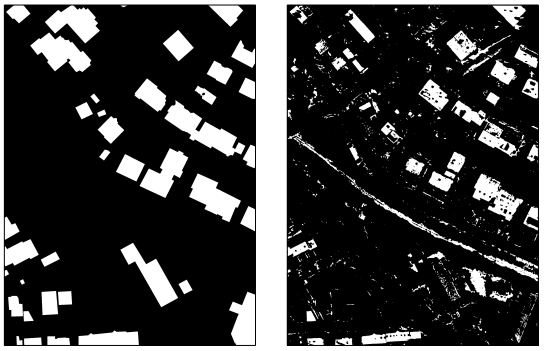


Fig. 1. Visualization of segmentation on Vaihingen dataset: (a), (d) input images 11 and 15 (false color composition), (b), (e) corresponding reference segmentation, and (c), (f) are segmentation produced by the proposed unsupervised method.

5 Conclusions

This paper proposed a deep clustering and contrastive learning based unsupervised semantic segmentation method for single scene EO images. Exploiting the large spatial size of the EO images, the proposed method divides the image into patches that are further used for training the unsupervised network. Pseudo labels are obtained by argmax classification of the final layer. The proposed method optimizes the labels and weights in iterations. The experimental results on Vaihingen dataset show the efficacy of the proposed method to obtain meaningful segmentation labels. Instead of seen as a competitor, the proposed method should be seen as a complementary to the existing supervised segmentation methods in EO. Since the proposed method provides a fast way to predict



(a) (b)

Fig. 2. Visualization of binary (building/other classes) segmentation on Vaihingen dataset image ID 11: (a) reference 11 , (b) segmentation produced by the proposed method.

reasonably accurate segmentation map using single scene, it may be useful in conjunction with supervised methods to generate pseudo-labels. Though we applied the proposed method to urban scenes, the model is application-agnostic. Our future work will aim towards improving the unsupervised segmentation of smaller classes (e.g., cars) and extending the proposed approach for ingesting other sensors, e.g., VHR Synthetic Aperture Radar (SAR) sensor.

References

1. Isprs benchmark. `https://www2.isprs.org/commissions/comm2/wg4/benchmark/`
2. Asano, Y.M., Rupprecht, C., Vedaldi, A.: A critical analysis of self-supervision, or what we can learn from a single image. arXiv preprint arXiv:1904.13132 (2019)
3. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 132–149 (2018)
4. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: 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 (2018)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709 (2020)
6. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
7. Ding, L., Zhang, J., Bruzzone, L.: Semantic segmentation of large-size vhr remote sensing images using a two-stage multiscale training architecture. IEEE Transactions on Geoscience and Remote Sensing **58**(8), 5367–5376 (2020)
8. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: Proceedings of the IEEE international conference on computer vision. pp. 1422–1430 (2015)

9. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision **111**(1), 98–136 (2015)
10. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. arXiv preprint arXiv:1803.07728 (2018)
11. Guo, X., Liu, X., Zhu, E., Yin, J.: Deep clustering with convolutional autoencoders. In: International conference on neural information processing. pp. 373–382. Springer (2017)
12. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
13. Hua, Y., Marcos, D., Mou, L., Zhu, X.X., Tuia, D.: Semantic segmentation of remote sensing images with sparse annotations. arXiv preprint arXiv:2101.03492 (2021)
14. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440 (2015)
15. Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P.: High-resolution aerial image labeling with convolutional neural networks. IEEE Transactions on Geoscience and Remote Sensing **55**(12), 7092–7103 (2017)
16. Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U.: Classification with an edge: Improving semantic image segmentation with boundary detection. ISPRS Journal of Photogrammetry and Remote Sensing **135**, 158–172 (2018)
17. Mou, L., Hua, Y., Zhu, X.X.: Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images. IEEE Transactions on Geoscience and Remote Sensing **58**(11), 7557–7569 (2020)
18. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
19. Saha, S., Bovolo, F., Bruzzone, L.: Building change detection in vhr sar images via unsupervised deep transcoding. IEEE Transactions on Geoscience and Remote Sensing (2020)
20. Saha, S., Bovolo, F., Bruzzone, L.: Change detection in image time-series using unsupervised lstm. IEEE Geoscience and Remote Sensing Letters (2020)
21. Saha, S., Mou, L., Qiu, C., Zhu, X.X., Bovolo, F., Bruzzone, L.: Unsupervised deep joint segmentation of multitemporal high-resolution images. IEEE Transactions on Geoscience and Remote Sensing (2020)
22. Saha, S., Sudhakaran, S., Banerjee, B., Pendurkar, S.: Semantic guided deep unsupervised image segmentation. In: International Conference on Image Analysis and Processing. pp. 499–510. Springer (2019)
23. Sherrah, J.: Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. arXiv preprint arXiv:1606.02585 (2016)
24. Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., Isola, P.: What makes for good views for contrastive learning. arXiv preprint arXiv:2005.10243 (2020)
25. Volpi, M., Tuia, D.: Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. IEEE TGRS **55**, 881–893 (2017)
26. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)