On Auxiliary Losses for Semi-Supervised Semantic Segmentation

Javiera Castillo-Navarro[1,2,⋆], Bertrand Le Saux[3], Alexandre Boulch[4], and Sébastien Lefèvre[2]

[1] ONERA, Université Paris-Saclay, F-91123 Palaiseau, France.
[2] Université Bretagne Sud, IRISA UMR 6074, F-56000 Vannes, France.
[3] European Space Agency, ESRIN, I-00044 Frascati (Rome), Italy
[4] valeo.ai, F-75008 Paris, France

Abstract. The development of semi-supervised learning methods is essential to Earth Observation applications. Indeed, labeled remote sensing data are scarce and likely insufficient to train fully supervised models with good generalization capacities. Conversely, raw data are abundant and therefore it is crucial to leverage unlabeled inputs to build better deep learning models. This work addresses the problem of semi-supervised semantic segmentation from a multi-task learning perspective. In this context, we explore several auxiliary tasks (reconstruction, unsupervised segmentation or self-supervision), and corresponding unsupervised losses, to perform along with semantic segmentation. Our experiments show the potential of semi-supervised learning approaches in a life-like scenario, outperforming a classical supervised setting.

Keywords: Semi-Supervised Learning · Semantic Segmentation · Multi-task Learning.

1 Introduction

Semantic segmentation, i.e. the problem of pixel-wise image classification, is of special importance in Earth Observation (EO). Indeed, many applications can be addressed as a semantic segmentation task, such as land cover mapping, building recognition or change detection. These tasks help us to deeply analyze and better understand our planet. In the last decade, the development of deep learning techniques has allowed to perform semantic segmentation with impressive success in an automatic way. Unfortunately, most of these methods rely heavily on the availability of large amounts of annotated data to be trained on. For this reason, special attention have been brought recently to the development of semi-supervised techniques, that leverage unlabeled data together with labeled data during the learning process. Semi-supervised learning is of significant interest in remote sensing, since labeled data are hard and costly to obtain, as they usually

⋆ Corresponding author: javiera.castillo_navarro@onera.fr

require expertise knowledge to be annotated, while raw data are continuously generated through satellites or drones.

Therefore, this work explores semi-supervised semantic segmentation from a multi-task learning perspective. In this context, a deep neural network is trained to perform two tasks simultaneously: the supervised semantic segmentation, as the main task, and an unsupervised auxiliary task. The latter enables to use prior information from data only –without labels–, in order to improve the results of the main task. Then, the objective function to optimize can be expressed as a weighted sum of two components: the segmentation loss –usually a standard cross-entropy loss– and the auxiliary loss. Yet, it is still unclear which task to perform along with semantic segmentation or which auxiliary loss to optimize.

In this paper, we review several auxiliary tasks and unsupervised losses for semi-supervised semantic segmentation. In particular, we present the Relaxed K-Means loss [4] and show it is very promising w.r.t. the state-of-the-art. We also highlight the interest of the Chrischurch Aerial Semantic Dataset (CASD) [23] for evaluating semi-supervised learning. More generally, we demonstrate the relevance of semi-supervised learning over supervised approaches in a life-like scenario where labeled data are scarce, while unlabeled data are abundant.

2 Related Work

Semi-Supervised Learning for Semantic Segmentation. The general idea is to learn a representation function –mapping a data point to its target– from labeled data and simultaneously leverage unlabeled data to improve this representation by learning the intrinsic properties of data. Existing semi-supervised methods for semantic segmentation in deep learning rely mostly on weak supervision, using scribbles [8], bounding boxes [11] or image-level annotations [19] to produce pixel-wise predictions. Fewer are the works that use completely unlabeled data during the learning process. For instance, co-training methods usually train an ensemble of segmentation models and use non-annotated images to exchange information between each other [22]. Other studies [25,10] propose adversarial approaches that integrate unlabeled data during training.

Self-Supervised Learning. Self-supervised (or unsupervised) learning aims to learn from the data only, with no labels nor specific task objective [15]. More exactly, it aims to accomplish pretext tasks that are self-induced, such as auto-encoding or blank completion, often for pre-training. In computer vision, usual pretext tasks are: predicting rotations [9], finding relative position of patches [7], solving the jigsaw puzzle[18], inpainting [21] or colorization [31].

Semi- and Self-Supervised Learning in Earth Observation. In the field of remote sensing, different approaches to benefit from unlabeled data have been developed. Non-annotated examples are used to design sharper feature extractors in [30], or to align manifolds for data coming from different modalities in [28]. Lately, deep learning approaches have been developed to leverage weakly annotated

data for different applications: land cover classification [17], change detection [5] or building extraction [2]. Some recent works integrate unlabeled data to the learning process, like [27] that proposes an alternating scheme to perform semi-supervised semantic segmentation, or [32] that considers an adversarial training strategy and uses unlabeled data for domain adaptation purposes.

Application of self-supervised methods in remote sensing is very recent. For instance, [26] proposes a new contrastive approach especially designed for multi-sensor data, while [29] introduces a colorization pretext task adapted to remote sensing data, reconstructing visible colors from high-dimensional spectral bands.

3 Semi-Supervised Semantic Segmentation

In this paper, we approach semi-supervised semantic segmentation as a multi-task problem. On the one hand, we learn to perform semantic segmentation under supervision using labeled data, and on the other hand, we learn to solve an auxiliary, unsupervised, task with raw data.

We choose to use BerundaNet-late [4], a simple and efficient network to address the multi-task semi-supervised semantic segmentation problem. BerundaNet consists of one decoder and a double decoder, where one decoder learns the supervised semantic segmentation task, while the second decoder is trained to perform an auxiliary task. Let $\phi_s(\cdot)$ and $\phi_u(\cdot)$ be the functions learned by the supervised and the unsupervised branch of the network, respectively. Let x be the input image and y the target label, the semi-supervised loss is expressed as a weighted sum of losses for each task:

$$\mathcal{L}(x) = \mathcal{L}_s(\phi_s(x), y) + \lambda \mathcal{L}_u(\phi_u(x), x) \tag{1}$$

where $\mathcal{L}_s$ is a supervised classification loss (usually cross-entropy loss), and thus the only one depending on y, and $\mathcal{L}_u$ is an unsupervised loss chosen according to the auxiliary task to perform. Section 4 presents some unsupervised tasks that can be performed along with semantic segmentation. λ is an hyperparameter of the model.

Since ϕ_s and ϕ_u partially share parameters, the additional unlabeled data will help to capture intrinsic properties through the unsupervised task. BerundaNet-late in particular has an almost-all shared parameters architecture, where only the last layers are split into specific tasks, as shown in Figure 1. Besides, BerundaNet-late is a versatile network and any classic encoder-decoder architecture can be used as backbone.

4 Auxiliary Tasks and Losses

In the semi-supervised setting described above, one could consider different auxiliary or pretext tasks to perform along with supervised semantic segmentation. In particular, we study three groups of auxiliary unsupervised tasks, which are associated to different kinds of unsupervised loss functions.

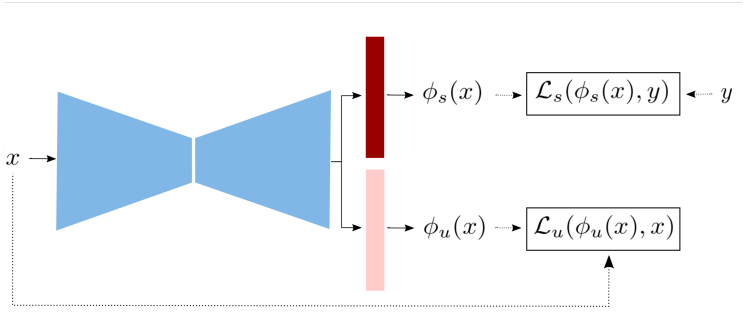

Fig. 1. BerundaNet-late architecture overview. unsupervised/supervised shared layers, is used for supervised layers only and is for unsupervised layers.

Notation: In what follows, we still note x the input image and $\mathbf{x}_i$ is the i-th pixel of the image. $\hat{x}$ denotes the reconstructed version of x ($\hat{\mathbf{x}}_i$, respectively). N is the number of pixels in an image.

4.1 Reconstruction Losses

The objective of the reconstruction (or auto-encoding) task is to generate an output as close as possible to the original input. For this reason, reconstruction losses enforce a similarity between the output of the network and the input image. Classical reconstruction losses are based on p-norms. In what follows we consider $\mathcal{L}_1$ and $\mathcal{L}_2$ reconstruction losses as defined in equation (2):

$$\mathcal{L}_1(x) = \frac{1}{N}\sum_{i=1}^{N} ||\mathbf{x}_i - \hat{\mathbf{x}}_i||_1, \quad \mathcal{L}_2(x) = \frac{1}{N}\sum_{i=1}^{N} ||\mathbf{x}_i - \hat{\mathbf{x}}_i||_2^2. \tag{2}$$

4.2 Segmentation Losses

Image segmentation aims to partition an image into multiple segments, where pixels in a segment share some properties, like color, intensity, or texture. This task can be performed in an unsupervised manner –based on the input image only– and might be a better complement to the supervised semantic segmentation task. We consider in this work two different unsupervised losses to perform unsupervised image segmentation.

Relaxed K-means loss [4]. The goal is to find an optimal set of K colors for encoding the image. As in the classic k-means algorithm, the relaxed k-means alternatively optimizes centroids of color clusters $\mathbf{c}_k$ ($k \in \{1, K\}$) and membership matrices $\hat{y}^{(k)}$ of x to cluster k.

Formally, the objective to minimize is given by equation (3):

$$\mathcal{L}_{km}(x) = \mathcal{L}_1(x, x_c) + \alpha \sum_{k=1}^{K} \sum_{i=1}^{N} \hat{y}_i^{(k)} (1 - \hat{y}_i^{(k)}), \tag{3}$$

x_c corresponds to a quantized version of the image, that is obtained by

$$x_c = \sum_{k=1}^{K} c_k \otimes \hat{y}^{(k)}, \quad \text{with} \quad \mathbf{c}_k = \frac{\sum_{i=1}^{N} x_i \hat{y}_i^{(k)}}{\sum_{i=1}^{N} \hat{y}_i^{(k)}}. \tag{4}$$

The relaxed k-means setting considers memberships $\hat{y}_i^{(k)} \in [0, 1]$ that can be obtained through a neural network (usually a soft-max output). $\otimes$ is the outer product.

Mumford-Shah loss. Classical image segmentation methods solve the segmentation problem by minimizing energy functions, such as the Mumford-Shah functional [16]. Recent work [12] shows that this functional can be adapted to be used as an unsupervised loss function in a deep neural network framework.

The unsupervised segmentation loss is then expressed as:

$$\mathcal{L}_{MS}(x) = \sum_{k=1}^{K} \sum_{i=1}^{N} ||\mathbf{x}_i - \mathbf{c}_k||_2^2 \ \hat{y}_i^{(k)} + \alpha \sum_{k=1}^{K} \sum_{i=1}^{N} |\nabla \hat{y}_i^{(k)}| \tag{5}$$

where we kept the same notations as before.

For both unsupervised segmentation losses, there are two hyperparameters to set: α, a regularization weighting parameter and k, the number of unsupervised clusters to identify. In our experiments, we set $\alpha = 1$ and we compare results for $k \in \{5, 10\}$. Section 5 reports only results for $k = 10$ since it exhibits better peformances.

4.3 Self-Supervised Losses

Recently, self-supervised methods have shown impressive results on learning data representations. The main idea behind self-supervision is to build a supervised task from completely unlabeled data by producing labels from the data itself. Many self-supervised tasks have been proposed lately, and we explore here two pretext tasks to perform along with semantic segmentation, that can be easily integrated to our semi-supervised framework.

Inpainting Similarly to the context autoencoder [21], we aim to solve the problem of filling in a missing piece in the image. The loss function is then expressed in terms of L_2 distance as

$$\mathcal{L}_{ca}(x) = \mathcal{L}_2 \left(M \odot x, M \odot \phi_u((1 - M) \odot x) \right) \tag{6}$$

where M is a binary mask (value of 1 for dropped pixels and 0 for input pixels) and $\odot$ the element-wise product.

There is an intrinsic hyperparameter to the inpainting problem: c, the crop size to mask from the image. In our experiments we try $c \in \{80, 160\}$ and in Section 5 we report results for $c = 80$ since it led to the best results. In our settings, masks are randomly chosen over the image.

Jigsaw puzzle Solving jigsaw puzzles using neural networks was first proposed by [18] to learn visual representations. In brief, the task consists in cutting out the image into 9 patches, shuffle them and train the network to retrieve the original image.

In practice, we follow here a similar approach to [3], where a network is trained to solve two tasks simultaneously (in our case, the jigsaw puzzle and the semantic segmentation) and the input is an image with permuted patches. The problem is then formulated as a classification task, using standard cross-entropy loss. We use the maximal Hamming distance algorithm from [18] to define a set of P allowed patch permutations. In our experiments we compare results for $P \in \{30, 100\}$. Since $P = 100$ led to the best results, we report them in Section 5.

5 Experiments

The Christchurch Aerial Semantic Dataset [14]

The CASD comprises aerial imagery ($\approx 5000 \times 4000$ px per image) at 10 cm/px resolution over Christchurch, New Zealand. Images were captured after the earthquake that struck the area on February 2011 and made available by Land Information New Zealand[1]. Dense semantic annotations were produced by ONERA/DTIS on 4 images [1,23], considering 4 classes: buildings, cars, vegetation and background. The dataset also includes 20 aerial images without any kind of annotation, which makes it suitable for semi-supervised learning algorithms. In practice, we use a training partition containing labeled and unlabeled data –2 annotated tiles and 20 non-annotated tiles–, and keep 2 annotated tiles for validation.

Implementation Details

BerundaNet-late is used here with a U-Net [24] backbone. For all experiments, it is trained using Adam optimizer [13] with a fixed learning rate of 10^{-4}, during 50 pseudo-epochs. Each pseudo-epoch consists in 5000 labeled samples and 5000 unlabeled samples (for the fully supervised experiments only labeled data is used). One sample is a 320×320 tile randomly chosen from training data. During testing time, tiles are processed with a sliding window of 320×320 pixels and overlap of 75%. PyTorch [20] is used for all implementations.

[1] https://www.linz.govt.nz/land/maps/linz-topographic-maps/imagery-orthophotos/christchurch-earthquake-imagery

Due to the stochastic nature of the optimization process, all experiments are averaged on 4 runs to obtain statistically significant results.

5.1 How to Harness Unlabeled Data?

Table 1 summarizes the results of our experiments with semi-supervision. Interestingly, for every loss studied in Section 4, there exists one (or more) λ value that allows to outperform the supervised setting.

Nevertheless, the best scores are obtained with unsupervised segmentation losses, where the approach with the *relaxed K-means loss* shows significant improvements, with respect to the supervised setting: mIoU score improves by +3.39%, while overall accuracy increases by +1.97%.

Table 1. Results comparison for supervised and semi-supervised methods over the Christchurch Aerial Semantic Dataset.

Mode	*Aux. Task*	*Aux. Loss*	λ	*OA (%)*	*mIoU (%)*
Sup	-	-	-	81.06 ± 0.46	67.43 ± 0.49
Semi-sup	Rec	$\mathcal{L}_1$	0.5	82.28 ± 0.55	68.78 ± 1.27
		$\mathcal{L}_2$	5	82.36 ± 0.42	68.99 ± 0.85
	Seg	$\mathcal{L}_{km}$	1	$\mathbf{83.03 \pm 0.42}$	$\mathbf{70.82 \pm 0.35}$
		$\mathcal{L}_{MS}$	1	82.94 ± 0.26	70.24 ± 0.84
	Self	$\mathcal{L}_{ca}$	5	82.57 ± 0.59	69.47 ± 0.7
		$\mathcal{L}_{js}$	0.5	82.88 ± 0.95	70.17 ± 1.12

Figure 2 shows two visual examples of the different methods. In the first example, the supervised approach is the only one that mistakes the shadow of trees over the river as a building; the supplementary information provided by unlabeled images to the semi-supervised methods allows to prevent this error. In the second image, the $\mathcal{L}_{km}$ loss is the only one that correctly segments the central building, likely thanks to the color clustering capacity.

5.2 Influence of the λ Hyperparameter

We also study the impact of the weighting parameter λ on the segmentation performance. Figure 3 illustrates the average behavior of each loss with respect to the value of λ.

Three behavioral groups appear. Segmentation losses are robust to the choice of λ and show, in general, better performances. $\mathcal{L}_1$ and $\mathcal{L}_{js}$ work better for small λ and require cautious hyperparameter tuning, as they are close to the fully-supervised case. $\mathcal{L}_2$ and $\mathcal{L}_{ca}$ losses show the same optimum for $\lambda = 5$, which comes likely from the fact that inpainting uses $\mathcal{L}_2$ to estimate discrepancies.

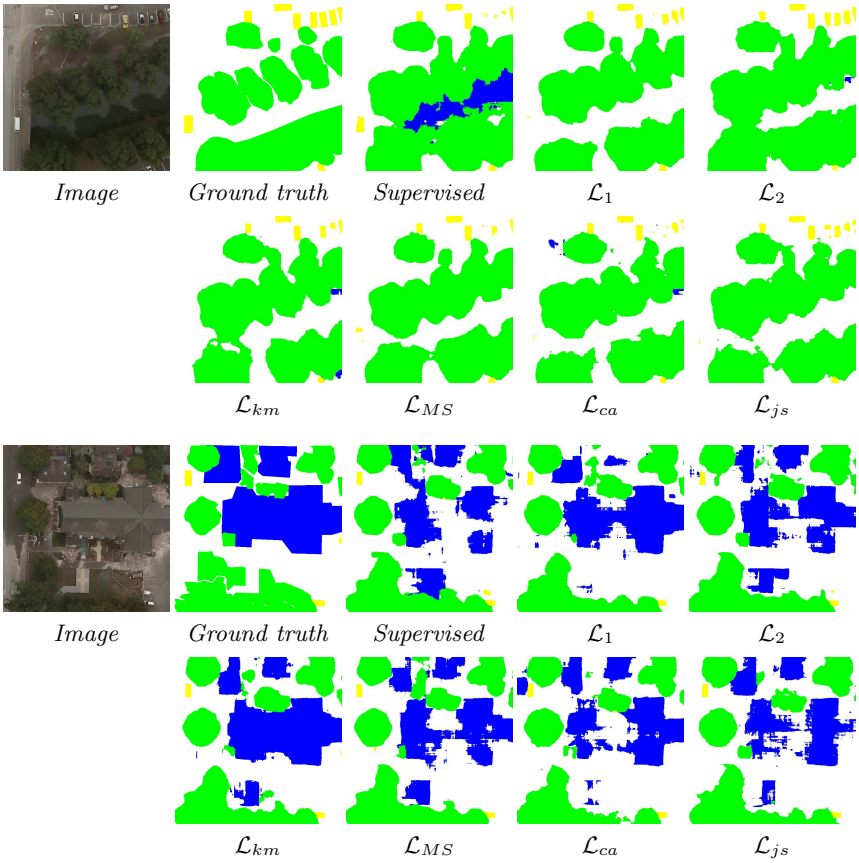

Fig. 2. Two examples of inference over the CASD dataset. buildings, cars, vegetation and background.

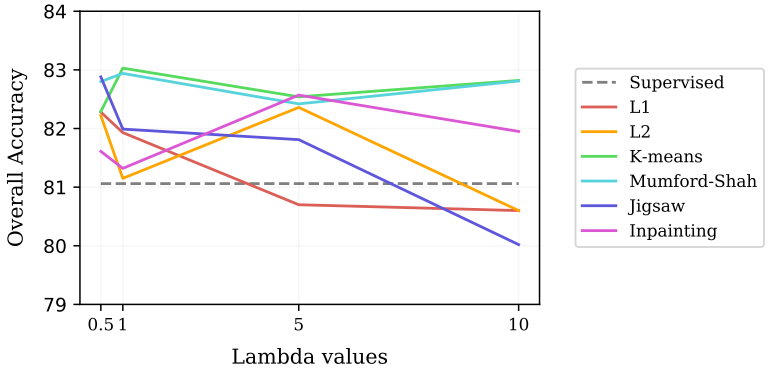

Fig. 3. Impact of the λ parameter on the semantic segmentation performance.

In a multi-task setting, different tasks might have very different behaviors and orders of magnitude. Tuning a weighting hyperparameter is not straightforward and further work is needed to find a neat normalization. Some works have even focused on adapting the multi-task loss balancing during training [6].

6 Conclusions

In this work, we have explored semi-supervised semantic segmentation in Earth Observation from a multi-task learning perspective. We presented a review of several auxiliary tasks and unsupervised losses to be used in such a setting. We performed experiments over a life-like dataset, the CASD dataset, and showed that it is well-suited for semi-supervised learning. This study also brought out that unsupervised segmentation losses –and in particular the relaxed K-means loss– are suitable as auxiliary losses for semantic segmentation. Indeed, they outperformed other approaches and showed to be robust to the balancing hyperparameter (λ) of the model.

Finally, our experiments have shown the potential of semi-supervised learning approaches over simple supervised learning in a realistic scenario, where labeled examples are limited, while unlabeled images are abundant.

References

1. Audebert, N., Le Saux, B., Lefèvre, S.: Segment-before-Detect: Vehicle Detection and Classification through Semantic Segmentation of Aerial Images. Remote Sensing **9**(4), 368 (2017)
2. Bonafilia, D., Gill, J., Basu, S., et al.: Building High Resolution Maps for Humanitarian Aid and Development with Weakly-and Semi-Supervised Learning. In: CVPR-W. pp. 1–9 (2019)
3. Carlucci, F.M., D'Innocente, A., Bucci, S., et al.: Domain Generalization by Solving Jigsaw Puzzles. In: CVPR. pp. 2229–2238 (2019)
4. Castillo-Navarro, J., Audebert, N., Boulch, A., et al.: Semi-Supervised Semantic Segmentation in Earth Observation: The MiniFrance suite, dataset analysis and multi-task network study. Under review. (2020)
5. Caye Daudt, R., Le Saux, B., Boulch, A., et al.: Guided Anisotropic Diffusion and Iterative Learning for Weakly Supervised Change Detection. In: CVPR-W (2019)
6. Chen, Z., Badrinarayanan, V., Lee, C.Y., et al.: GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks. In: ICML. pp. 794–803 (2018)
7. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised Visual Representation Learning by Context Prediction. In: ICCV. pp. 1422–1430 (2015)
8. Durand, T., Mordan, T., Thome, N., et al.: WILDCAT: Weakly Supervised Learning of Deep ConvNets for Image Classification, Pointwise Localization and Segmentation. In: CVPR. vol. 2 (2017)
9. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised Representation Learning by Predicting Image Rotations. arXiv preprint arXiv:1803.07728 (2018)
10. Hung, W.C., Tsai, Y.H., Liou, Y.T., et al.: Adversarial Learning for Semi-Supervised Semantic Segmentation. BMVC (2018)

11. Khoreva, A., Benenson, R., Hosang, J.H., et al.: Simple Does It: Weakly Supervised Instance and Semantic Segmentation. In: CVPR (2017)
12. Kim, B., Ye, J.C.: Mumford-Shah Loss Functional for Image Segmentation with Deep Learning. IEEE Transactions on Image Processing (2019)
13. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. In: ICLR (2015)
14. Le Saux, B., Randrianarivo, H., Audebert, N.: Christchurch Aerial Semantic Dataset (2019). https://doi.org/10.5281/zenodo.3566005
15. LeCun, Y.: Self-Supervised Learning (2 2020), Keynote Lecture, AAAI
16. Mumford, D., Shah, J.: Boundary detection by Minimizing Functionals. In: CVPR. vol. 17, pp. 137–154 (1985)
17. Nivaggioli, A., Randrianarivo, H.: Weakly Supervised Semantic Segmentation of Satellite Images. In: JURSE. pp. 1–4. IEEE (2019)
18. Noroozi, M., Favaro, P.: Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. In: ECCV. pp. 69–84. Springer (2016)
19. Papandreou, G., Chen, L.C., Murphy, K.P., et al.: Weakly- and Semi-Supervised Learning of a Deep Convolutional Network for Semantic Image Segmentation. In: ICCV. pp. 1742–1750 (2015)
20. Paszke, A., Gross, S., Massa, F., et al.: PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: NeurIPS (2019)
21. Pathak, D., Krahenbuhl, P., Donahue, J., et al.: Context Encoders: Feature Learning by Inpainting. In: CVPR. pp. 2536–2544 (2016)
22. Peng, J., Estrada, G., Pedersoli, M., et al.: Deep co-training for semi-supervised image segmentation. Pattern Recognition (2020)
23. Randrianarivo, H., Le Saux, B., Ferecatu, M.: Urban Structure Detection with Deformable Part-based Models. In: IGARSS. pp. 200–203. IEEE (2013)
24. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: MICCAI. pp. 234–241. Springer (2015)
25. Souly, N., Spampinato, C., Shah, M.: Semi-supervised Semantic Segmentation using Generative Adversarial Network. In: ICCV. pp. 5688–5696 (2017)
26. Swope, A.M., Rudelis, X.H., Story, K.T.: Representation Learning for Remote Sensing: An Unsupervised Sensor Fusion Approach (2020), https://openreview.net/forum?id=SJlVn6NKPB
27. Tao, Y., Xu, M., Zhang, F., et al.: Unsupervised-Restricted Deconvolutional Neural Network for Very High Resolution Remote-Sensing Image Classification. IEEE Transactions on Geoscience and Remote Sensing **55**(12), 6805–6823 (2017)
28. Tuia, D., Volpi, M., Trolliet, M., et al.: Semi-supervised Manifold Alignment of Multimodal Remote Sensing Images. IEEE Transactions on Geoscience and Remote Sensing **52**(12), 7708–7720 (2014)
29. Vincenzi, S., Porrello, A., Buzzega, P., et al.: The Color out of Space: Learning Self-supervised Representations for Earth Observation Imagery. arXiv preprint arXiv:2006.12119 (2020)
30. Xia, J., Chanussot, J., Du, P., et al.: (Semi-) Supervised Probabilistic Principal Component Analysis for Hyperspectral Remote Sensing Image Classification. IEEE J. of Sel. Top. in Applied Earth Obs. and Remote Sensing **7**(6), 2224–2236 (2013)
31. Zhang, R., Isola, P., Efros, A.: Colorful Image Colorization. In: ECCV. pp. 649–666. Springer (2016)
32. Zhu, R., Yan, L., Mo, N., et al.: Semi-supervised Center-based Discriminative Adversarial Learning for Cross-Domain Scene-Level Land-Cover Classification of Aerial Images. ISPRS J. of Photogram. and Remote Sensing **155**, 72–89 (2019)