IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium | 979-8-3503-6032-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/IGARSS53475.2024.10641541

LEARNING FROM UNLABELLED DATA WITH TRANSFORMERS: DOMAIN ADAPTATION FOR SEMANTIC SEGMENTATION OF HIGH RESOLUTION AERIAL IMAGES

Nikolaos Dionelis[1], *Francesco Pro*[2], *Luca Maiano*[2], *Irene Amerini*[2], *Bertrand Le Saux*[1]

[1] European Space Agency (ESA), ESRIN, Φ-lab, Italy
[2]Sapienza University of Rome, Italy

ABSTRACT

Data from satellites or aerial vehicles are most of the times unlabelled. Annotating such data accurately is difficult, requires expertise, and is costly in terms of time. Even if Earth Observation (EO) data were correctly labelled, labels might change over time. Learning from unlabelled data within a semi-supervised learning framework for segmentation of aerial images is challenging. In this paper, we develop a new model for semantic segmentation of unlabelled images, the Non-annotated Earth Observation Semantic Segmentation (NEOS) model. NEOS performs domain adaptation as the target domain does not have ground truth masks. The distribution inconsistencies between the target and source domains are due to differences in acquisition scenes, environment conditions, sensors, and times. Our model aligns the learned representations of the different domains to make them coincide. The evaluation results show that it is successful and outperforms other models for semantic segmentation of unlabelled data.

Index Terms— Semantic segmentation, Unlabelled data

1. INTRODUCTION

Importance and overview. Remote Sensing (RS) images from satellites or aerial vehicles can be used to map trees and land cover classes [1]. While both RS technology and AI for data analysis continue to advance [2], the integration and use of airplanes and drones for localized studies is nowadays also increasing. Supervised learning has shown good performance for classification and segmentation. However, it requires high-quality handcrafted *large* labelled datasets. Learning from unlabelled data is challenging as the performance of models depends highly on the *size* and quality of the data. However, for real-world applications [3], labelling large datasets is laborious, expensive, and time-consuming. This holds for Earth Observation (EO), where huge amounts of data are produced *daily*. Also, data from satellites or aircrafts usually require domain expertise. Furthermore, labels for specific geographical regions may change over time (task of change detection) due to nature (seasonality), man-induced changes, and natural hazards (volcano eruptions). Also, for specific regions, some labels might be incorrect (task of learning from *noisy labels*). Because many satellites and aerial images are unlabelled, it is challenging to effectively use these data. Developing semi-supervised learning methods is crucial to improve generalization performance. Semi-supervised learning, which involves training on both a labelled dataset, where both images and their annotations are provided, and on an *unlabelled* set, with only image data, is a more realistic setting than supervised learning, as in RS, unlabelled data are *plentiful*, while labelled data can be hard to find. This holds for semantic segmentation (*pixel-level* labels) [1], which requires assigning a class label to each pixel [4, 5] by understanding its semantics. This task is crucial for several applications, including land cover mapping and urban change detection. In this work, we propose a method to perform semantic segmentation on unlabelled datasets, and we evaluate it on the unlabelled Cross-View USA (CVUSA) dataset [6]. To the best of the authors' knowledge, *accurate* semantic segmentation on the CVUSA aerial dataset, which is used for cross-view aerial-ground matching [7, 8] and has no land cover label annotations, has not yet been performed.

Domain adaptation. In this paper, we develop a model for semantic segmentation of aerial images, the Non-annotated Earth Observation Semantic Segmentation (NEOS) model. NEOS makes the learned representations of the *different domains* to coincide. This is achieved by minimizing the distribution differences of the different domains. Hence, we enforce the model to be able to work well with the different datasets that have distribution inconsistencies due to differences in acquisition scenes, environment conditions, sensors, and times. Our model performs semantic segmentation on the unlabelled dataset CVUSA. During training, a loss function is minimized that makes the network to align the latent features of the *different domains* to minimize distribution differences. Our main contribution is the development of a novel model for semantic segmentation of aerial images that do *not* have ground truth segmentation masks, also performing domain adaptation.

2. RELATED WORK

Domain adaptation methods in deep learning, as well as in RS, have been developed recently. Furthermore, *domain adaptation* methods for classification and segmentation have also been developed. Models trained on data from one domain

may *not* generalize well on other domains. Even in one domain, accurate semantic segmentation is challenging. There may be a loss in accuracy when deploying a model on *unseen* data due to a shift between the distributions in the source and *target* domains [9]. Domain adaptation tries to overcome this [10, 11]. Domain gaps are common in aerial images [12, 13], e.g. region change. In [8], no off-the-shelf semantic segmentation model transferred well/ accurately on the aerial CVUSA dataset.

Architectures. Several models have been developed recently for semantic segmentation, including *Transformer* (e.g. SegFormer [14]), encoder-decoder like SegNet or U-Net with skip connections, Fully Convolutional Network (FCN), and dilated convolutions with larger receptive field to capture *long-range* information. SegFormer [14] for semantic segmentation combines the Transformer with an efficient Multi-Layer Perceptron (MLP) decoder and outputs *multi-scale* features. The decoder combines these multi-scale feature maps, which use local and global attention. UNetFormer is a UNet-like architecture for segmentation based on a ResNet encoder and a *Transformer* decoder [15]. It uses efficient *attention* in the decoder to model global and local information. The Segment Anything Model (SAM) [16] has also been recently proposed for *instance* segmentation, but not semantic segmentation.

Unsupervised adaptation. In the Unsupervised Domain Adaptation (UDA) setting, the model is trained on both labelled and unlabelled data from the source and target domains, respectively. Accurate *real-world* semantic segmentation without labels is non-trivial. Domain Symmetric Networks (SymNet) [17, 12] design the *source and target domains* classifier symmetrically to learn domain-invariant features for effective domain adaptation. The model Source Hypothesis Transfer (SHOT) [18] uses hypothesis transfer, training only the backbone and making the classifier of the network *non-trainable*. We focus on the scenario in which the number of classes and the classes themselves are the same in the source and target domains, the *closed-set* setting [12, 19]. The aim is to achieve good performance during inference on the *target domain* test dataset, as well as on the source domain test dataset [10, 20].

Semi-supervised learning methods for segmentation in deep learning, as well as in RS, have been developed. Semi-supervision [5, 21], which is halfway between supervised and *unsupervised* learning, deals with settings where labelled sets of data and their targets are provided and unlabelled sets with data *only* are available. Unlabelled data [5] help the learning process to *improve* performance. To improve generalization [21, 22], models should be able to handle labelled and *unlabelled* data, as well as operate within a multi-task optimization framework [1] performing unsupervised minimization tasks.

3. PROPOSED METHODOLOGY

Flowchart. NEOS is presented in Fig. 1. The input is the image from labelled and unlabelled datasets. The output is the estimated *semantic* segmentation mask. NEOS is based on the

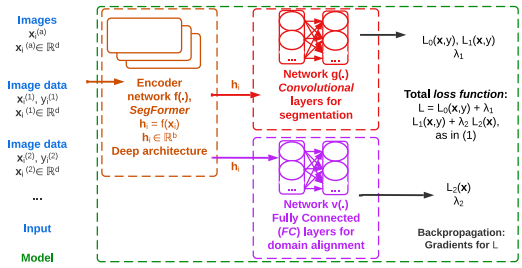


Fig. 1. Flowchart of NEOS for semantic segmentation using domain adaptation on datasets with no *ground truth* labels.

SegFormer B5 [14] architecture and uses a second output for the feature misalignment loss term for domain adaptation.

Loss function. NEOS minimizes a loss comprising the terms: (a) cross-entropy for pixel-level classification which is computed using the input labelled images and their corresponding ground truth segmentation masks, (b) (1 - *Dice score*) for segmentation, and (c) the *features* misalignment loss. The latter is for domain adaptation to enforce the model to be able to work *well* with the different datasets. Here, an architecture with *two heads* is used. The first two loss terms control the first output head for accurate segmentation and classification on the different *labelled* datasets, while the features misalignment loss, which controls the second output head, enforces the network to reduce the distance between the latent representations of the labelled and unlabelled data in the cross-entropy metric. This forces the features to achieve manifold alignment of the embeddings of the different domains. We denote the data by $\mathbf{x}$ and the ground truth masks by $\mathbf{y}$. Now, the *cost* function is:

$$\text{argmin}_f \, L, \quad L = L_0(\mathbf{x}, \mathbf{y}) + \lambda_1 L_1(\mathbf{x}, \mathbf{y}) + \lambda_2 L_2(\mathbf{x}), \quad (1)$$

where we denote our model by $f(\cdot)$, the first, second, and third *loss* terms by L_0, L_1, and L_2, respectively, and the hyperparameters of the second and third losses by λ_1 and λ_2.

The first term of our model's objective loss function is:

$$L_0 = -\frac{1}{NWH} \sum_{j=1}^{N} \sum_{i=1}^{W} \sum_{l=1}^{H} \log \frac{\exp(f_{y_{j,i,l}}(\mathbf{x}_j))}{\sum_{k=1}^{K} \exp(f_{k,i,l}(\mathbf{x}_j))}, \quad (2)$$

where W and H are the width and height of the images, i and l the indices for the width and height, N the number of samples, j their index, K the number of classes, and k their index. In (2), the N training samples originate from Q labelled datasets. This affects both $\mathbf{x}_j$ and $y_{j,i,l}$. The features before the normalized exponential, i.e. *softmax*, are denoted by $f_{y_{j,i,l}}(\mathbf{x}_j)$. The model computes the estimated probability of the labels for the cross-entropy loss to *reward* correct classification, penalizing deviation from the correct class. The estimated *semantic* segmentation mask, $\hat{y}_{j,i,l}$ is the pixel-wise class label. Here, for the labelled data, to perform accurate classification, NEOS minimizes (2), which is the *pixel-wise* cross-entropy loss.

Next, the second term of the loss function, L_1 in (1), is:

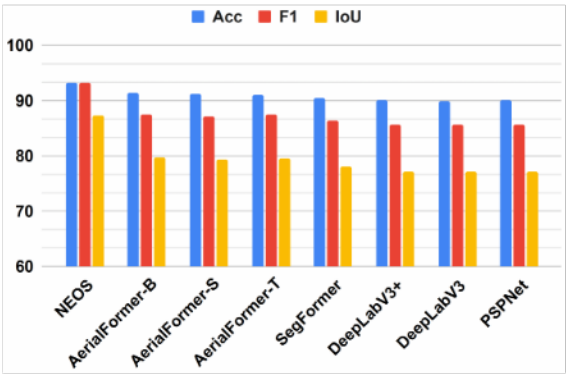


Fig. 2. Evaluation of NEOS in accuracy (Acc), F1-score (F1) and IoU on the dataset Potsdam with the class Clutter [12].

$$1 - \frac{2\sum_{j=1}^{N}\sum_{i=1,l=1}^{W,H} g_{j,i,l}s_{j,i,l}}{\sum_{j=1}^{N}\sum_{i=1,l=1}^{W,H} g_{j,i,l} + \sum_{j=1}^{N}\sum_{i=1,l=1}^{W,H} s_{j,i,l}}, \quad (3)$$

where $g_{j,i,l}$ is the true binary indicator of the class label, and $s_{j,i,l}$ is the estimated *semantic* segmentation probability outputted by the model. The N samples originate from Q labelled datasets. This affects both $g_{j,i,l}$ and $s_{j,i,l}$. To perform accurate segmentation on the labelled data, NEOS minimizes (3), the *Dice* loss. Next, the third term of the loss function, L_2, is:

$$L_2 = \frac{1}{J}\sum_{j=1}^{J} \log \frac{\exp(f_{z_j}(\mathbf{x}_j))}{\sum_{m=1}^{M}\exp(f_m(\mathbf{x}_j))}, \quad (4)$$

where the true domain label is denoted by z_j, and the number of domain labels by M. The samples originate from Q labelled and R *unlabelled* datasets. For domain adaptation, NEOS minimizes (4). To improve generalization (better performance), we perform data augmentation to incorporate *invariances* into the model. We regularize and enforce the model to generalize and be robust to data transformations, and we also perform downsampling, having inputs at different multi-scale levels.

4. EVALUATION AND RESULTS

Labelled and unlabelled datasets. We train NEOS on the aerial image datasets: labelled Potsdam and Vaihingen, and unlabelled CVUSA by performing domain adaptation. We test NEOS on Potsdam and Vaihingen, as well as on CVUSA. We also *compare* our model to other baseline models, including SAM [16]. For the labelled datasets, the classes are: Buildings (blue colour), Trees (green), Cars (yellow), Low vegetation (cyan), Roads (white), and Clutter (red) [23, 24]. Also, for the different datasets and the domain adaptation loss term, as well as the *tags* for the different datasets, Tag A is used for the dataset Potsdam, Tag B for Vaihingen, and C for CVUSA.

Evaluation of NEOS on Potsdam. We evaluate NEOS on the dataset Potsdam in Fig. 2. In this experiment, the class

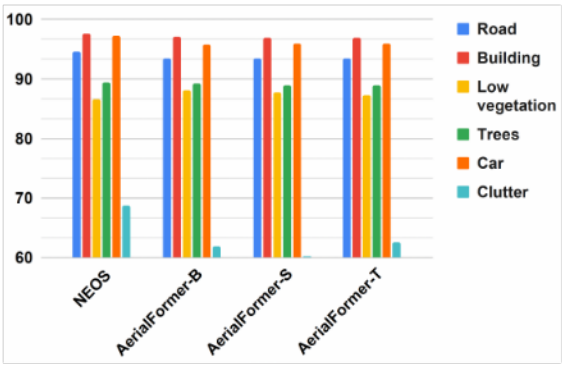
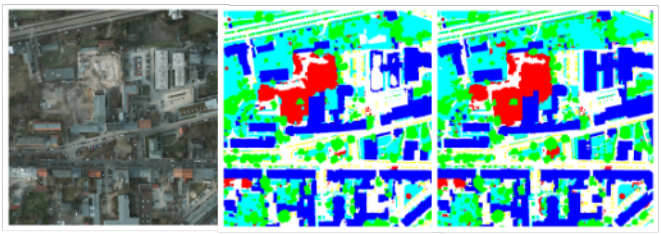


Fig. 3. Per-class *F1-score* evaluation (in %) of NEOS on the Potsdam dataset including the class Clutter in the evaluation.



a) Input b) NEOS (Ours) c) Ground truth

Fig. 4. Semantic segmentation masks by NEOS on Potsdam.

Clutter is considered in the evaluation. Here, the evaluation is based on the accuracy, F1-score, and Intersection over Union (IoU) metrics. It can be observed in Fig. 2 that the proposed model outperforms all the other baseline models [23]. In Fig. 3, we evaluate the *per-class* F1-score performance of NEOS on Potsdam. In Fig. 4, we present NEOS *qualitative* results.

Evaluation of NEOS on Vaihingen. We evaluate NEOS in Fig. 5 in accuracy, F1-score and IoU. Here, the results of NEOS on Vaihingen are comparable to other models. In Fig. 6, we also evaluate the *F1-score* of NEOS for each class [15]. For Roads, Buildings and Cars, NEOS outperforms other models.

Evaluation on the unlabelled dataset CVUSA. We evaluate NEOS on the *unlabelled* dataset aerial CVUSA. The testing is done on a dataset that does *not* have ground truth masks. In Fig. 7, we present the *qualitative* results of NEOS. We observe in (b) and (e) that NEOS performs semantic segmentation and is able to *recognize* effectively classes such as Roads and Low vegetation. This holds for NEOS for the vertical roads that have *shadows* (occlusion) in (a)-(b). In the next paragraphs, we present a numerical evaluation of NEOS on the unlabelled CVUSA dataset and a further comparison to other models.

In Fig. 7, we examine the qualitative results of NEOS, and we also do this at a *large scale*, automating the process. We assess the performance of NEOS on many images, and to capture the big picture, we evaluate NEOS *numerically* by computing the Segments of Predictions and Inputs Error (SPIE). We first perform detection of segments on the estimated mask and input images, and for this, we use a variant of SAM [26]. Then, the

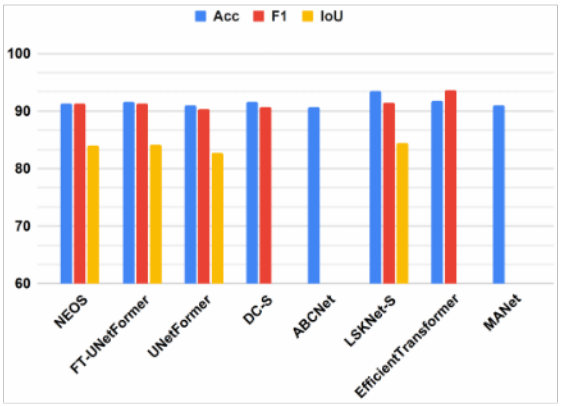


Fig. 5. Evaluation of NEOS on Vaihingen in Acc, F1 and IoU.

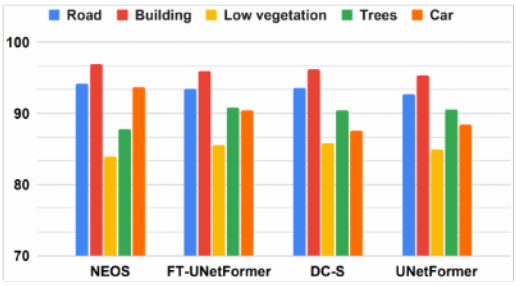


Fig. 6. Per-class F1-score evaluation of NEOS on Vaihingen.

Table 1. Evaluation of NEOS on the CVUSA dataset, on both the aerial (Aer) and street (Str), and the improvement (I) over the base model.

SPIE for aerial & street	**Aer**	**I Aer**	**Str**	**I Str**
NEOS (Ours)	0.047	32%	0.041	21%
Base model, SegFormer	0.069	N/A	0.052	N/A
CNN-based using Eq. (1)	0.064	7.2%	0.049	5.8%

error is calculated and normalized. For perfect segmentation without considering semantic information (no labels), SPIE is *zero*. For completely inaccurate segmentation, SPIE is equal to one. SPIE is the mean residual image where the residual is between the *estimated* mask and the input after being modified by a detection of segments algorithm, and its definition is:

$$\text{SPIE} \;=\; \frac{1}{R}\sum_{j=1}^{R} g(f(\mathbf{x}_j)) - g(\mathbf{x}_j) \qquad (5)$$

where $f(\cdot)$ is NEOS from (1) (or another model), R the number of evaluation samples, and $g(\cdot)$ a detection of segments algorithm. We use SPIE as an indicator of good performance and as an empirical metric that works in practice. Using SPIE in (5), the numerical results *match* the qualitative results we obtain. In addition, we have also included the code in [27].

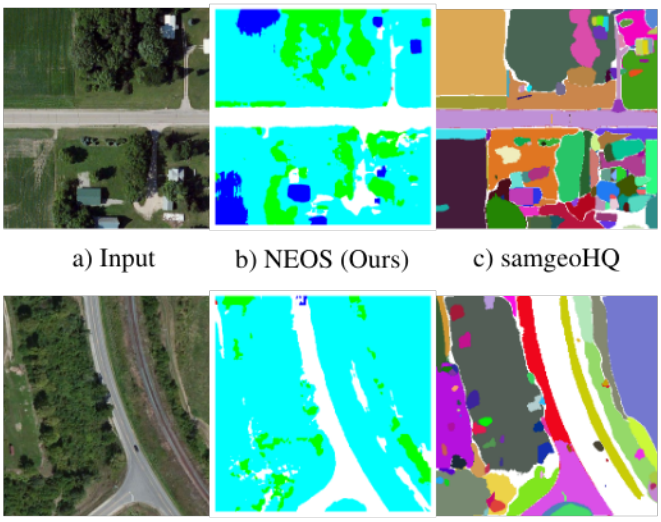


Fig. 7. Qualitative evaluation of NEOS on the unlabelled CVUSA aerial dataset, and comparison to samgeoHQ [25].

We evaluate NEOS numerically on CVUSA in Table 1 using SPIE, and we compare with other models for *semantic* segmentation. The improvement of NEOS for aerial images over the base model, SegFormer [14], is 32%. NEOS also outperforms [14] for street images when we use the labelled dataset CityScapes [28] for the domain adaptation loss in Sec. 3.

Further comparison of NEOS to other models. In Fig. 7, we examine the results of NEOS and its comparison to SAM [16]. NEOS performs *joint* classification and segmentation, while SAM does *not* consider semantics [29]. There are several *variants* of SAM: Geospatial SAM [25] *fine-tunes* [16] on aerial data. SAM-HQ [30] achieves improved accuracy (FastSAM [31], speed). Semantic SAM (SSAM) [32, 33] modifies [16] to perform semantic (rather than instance) segmentation, but *not* for aerial data. We examined the performance of SAM and its variants on the unlabelled CVUSA dataset. In Fig. 7, samgeoHQ [25], which does *not* perform semantic segmentation, is sensitive to even small changes in the scene. For scenes with details, we need to adjust the several tunable parameters of SAM to control how *dense* the estimated masks are. For NEOS, we do not need to adjust its parameters, and this can potentially lead to improved user convenience and ease of use.

5. CONCLUSION

We have developed a semantic segmentation method that is effective for unlabelled datasets. The results show that NEOS outperforms other models. We have used the unlabelled aerial CVUSA dataset, where accurate semantic segmentation has not yet been performed, to the best of the authors' knowledge, and we plan to also use the results for cross-view geo-location matching to accurately match aerial and street images [34, 35].

6. REFERENCES

[1] Gaston Lenczner, Adrien Chan-Hon-Tong, Bertrand Le Saux, et al., "DIAL: Deep interactive and active learning for semantic segmentation in remote sensing," *IEEE J Sel Top Appl Ear Obs Rem Se, v. 15, p. 3376-3389*, 2022.

[2] Devis Tuia, Konrad Schindler, Begüm Demir, Gustau Camps-Valls, Xiao Xiang Zhu, Mrinalini Kochupillai, Sašo Džeroski, Jan N van Rijn, Holger H Hoos, Fabio Del Frate, et al., "Artificial intelligence to advance Earth observation: a perspective," *arXiv preprint arXiv:2305.08413*, 2023.

[3] Devis Tuia, Moser Gabriele, Bertrand Le Saux, and Benjamin Bechtel, "Data Fusion Contest 2017 (DFC2017)," *IEEE GRSS*, 2022. DOI: 10.21227/e56j-eh82.

[4] Ziyang Zhang, Plamen Angelov, Eduardo Soares, Nicolas Longepe, and Pierre P. Mathieu, "An interpretable deep semantic segmentation method for Earth Observation," *In IEEE Int Conf Intelligent Systems (IS)*, 2022.

[5] Javiera Castillo-Navarro, Bertrand Le Saux, et al., "Semi-supervised semantic segmentation in earth observation: The MiniFrance suite, dataset analysis and multi-task network," *Machine Learning (111), p. 3125-3160*, 2022.

[6] Scott Workman et al., "Wide-area image geolocalization with aerial reference imagery," *In ICCV*, pp. 1–9, 2015.

[7] Yujiao Shi, Xin Yu, et al., "Where am I looking at? Joint location and orientation estimation by cross-view matching," *In CVPR*, pp. 4064–4072, 2020.

[8] Royston Rodrigues and Masahiro Tani, "Are these from the same place? Seeing the unseen in cross-view image geo-localization," In WACV, p. 3753-3761, 2021.

[9] Jiangtao Peng et al., "Domain adaptation in remote sensing image classification: A survey," *IEEE J Sel Top Appl Earth Obs Remote Sen, v. 15, p. 9842-9859*, 2022.

[10] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, et al., "Domain-adversarial training of neural networks," *JMLR*, 2016.

[11] Xiaofeng Liu et al., "Deep unsupervised domain adaptation: A review of recent advances and perspectives," *APSIPA Trans. Signal Inf. Process. 11 (1)*, 2022.

[12] Navya Nagananda, Abu Md Niamul Taufique, Raaga Madappa, et al., "Benchmarking domain adaptation methods on aerial datasets," *Sensors, MDPI*, 2021.

[13] Ying Chen et al., "Semantic segmentation in aerial images using class-aware unsupervised domain adaptation," *ACM SIGSPATIAL Int Workshop GEOAI, p. 9-16*, 2021.

[14] Enze Xie, Wenhai Wang, Zhiding Yu, et al., "SegFormer: Simple and efficient design for semantic segmentation with Transformers," *In NeurIPS, 34:12077-12090*, 2021.

[15] Libo Wang et al., "UNetFormer: A UNet-like Transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS Journal Photogrammetry Remote Sensing, v. 190, p. 196-214*, 2022.

[16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, et al., "Segment anything," *arXiv:2304.02643*, 2023.

[17] Yabin Zhang et al., "Domain-symmetric networks for adversarial domain adaptation," *In CVPR*, 2019.

[18] Jian Liang et al., "Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation," *In ICML, PMLR 119*, 2020.

[19] Kaichao You, Mingsheng Long, et al., "Universal domain adaptation," *In CVPR, p. 2720-2729*, 2019.

[20] Valerio Marsocci et al., "GeoMultiTaskNet: Remote sensing unsupervised domain adaptation using geographical coordinates," *Workshop CVPR*, 2023.

[21] Kihyuk Sohn, David Berthelot, et al., "FixMatch: Simplifying semi-supervised learning with consistency and confidence," *In NeurIPS, v. 33, p. 596-608*, 2020.

[22] Pablo Gómez and Gabriele Meoni, "MSMatch: Semi-supervised multispectral scene classification with few labels," *IEEE Journal Selected Topics Applied Earth Observations and Remote Sensing, 14:11643-11654*, 2021.

[23] Kashu Yamazaki, Taisei Hanyu, Minh Tran, et al., "AerialFormer: Multi-resolution Transformer for aerial image segmentation," *arXiv:2306.06842*, 2023.

[24] X. He et al., "Swin Transformer embedding U-Net for RS image semantic segmentation," T Geo RS (60), 2022.

[25] Qiusheng Wu and Lucas P. Osco, "samgeo: A Python package for segmenting geospatial data with SAM," *Journal of Open Source Software, 8(89), 5663*, 2023.

[26] A. Hancharenka, "SAMEO: Segment anything EO tools," *GitHub*, 2023.

[27] Nikolaos Dionelis, Francesco Pro, Luca Maiano, Irene Amerini, and Bertrand Le Saux, "Learning from unlabelled data: Domain adaptation for semantic segmentation," *GitHub repository*, 2023. `http://github.com/ESA-PhiLab/Learning_from_Unlabeled_Data_for_Domain_Adaptation_for_Semantic_Segmentation`.

[28] Marius Cordts et al., "The Cityscapes dataset for semantic urban scene understanding," *In Proc. CVPR*, 2016.

[29] Yu Qiao, Chaoning Zhang, Taegoo Kang, Donghun Kim, et al., "Robustness of SAM: Segment anything under corruptions and beyond," *arXiv:2306.07713*, 2023.

[30] Lei Ke, Mingqiao Ye, et al., "Segment anything in high quality," *In NeurIPS, Poster Session 4*, 2023.

[31] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, et al., "Fast segment anything," *arXiv:2306.12156*, 2023.

[32] Jiaqi Chen, Zeyu Yang, and Li Zhang, "Semantic segment anything," 2023. http://github.com/fudan-zvg/Semantic-Segment-Anything.

[33] Feng Li et al., "Semantic-SAM: Segment and recognize anything at any granularity," *arXiv:2307.04767*, 2023.

[34] Francesco Pro, Nikolaos Dionelis, Luca Maiano, Bertrand Le Saux, and Irene Amerini, "A semantic segmentation-guided approach for ground-to-aerial image matching," *IGARSS*, 2024. Also: arXiv:2404.11302.

[35] Francesco Pro. (2023), "Ground-to-Aerial Image Matching for Geospatial Applications," [Unpublished Master's thesis]. Sapienza University of Rome.