GeoMultiTaskNet: remote sensing unsupervised domain adaptation using geographical coordinates

Valerio Marsocci
Sapienza University of Rome
valerio.marsocci@uniroma1.it

Nicolas Gonthier
Univ Gustave Eiffel, IGN, ENSG, LASTIG
nicolas.gonthier@ign.fr

Anatol Garioud
IGN
anatol.garioud@ign.fr

Simone Scardapane
Sapienza University of Rome
simone.scardapane@uniroma1.it

Clément Mallet
Univ Gustave Eiffel, IGN, ENSG, LASTIG
clement.mallet@ign.fr

Abstract

Land cover maps are a pivotal element in a wide range of Earth Observation (EO) applications. However, annotating large datasets to develop supervised systems for remote sensing (RS) semantic segmentation is costly and time-consuming. Unsupervised Domain Adaption (UDA) could tackle these issues by adapting a model trained on a source domain, where labels are available, to a target domain, without annotations. UDA, while gaining importance in computer vision, is still under-investigated in RS. Thus, we propose a new lightweight model, GeoMultiTaskNet, based on two contributions: a GeoMultiTask module (GeoMT), which utilizes geographical coordinates to align the source and target domains, and a Dynamic Class Sampling (DCS) strategy, to adapt the semantic segmentation loss to the frequency of classes. This approach is the first to use geographical metadata for UDA in semantic segmentation. It reaches state-of-the-art performances (47,22% mIoU), reducing at the same time the number of parameters (33M), on a subset of the FLAIR dataset, a recently proposed dataset properly shaped for RS UDA, used for the first time ever for research scopes here.

1. Introduction

Accurate land cover information is crucial for a wide range of applications, including environmental monitoring and management [9,21,56], urban planning, and monitoring [6,41]. In particular, semantic segmentation is a key task in the analysis of very high-resolution (VHR) remote sensing (RS) images, as it enables the automatic categorization of land cover [32]. However, annotating large datasets for supervised learning is costly and time-consuming, especially when not all data are acquired contemporaneously [31].

In this context, unsupervised domain adaptation (UDA) offers a promising solution for adapting a model trained on a source domain to a target domain, without the need for annotations [13, 20, 26], reducing domain shift. Although this task is gaining importance in computer vision (CV) [17,18,52], in RS it is still under-investigated. On one hand, often new RS UDA methods are applied on datasets not properly developed for this purpose [38] and, consequently, far from the real-world UDA scenario. On the other hand, general CV models are often applied to RS images, with little regard to the EO peculiarities. A clear example is the use of metadata, such as geographical coordinates, which are often discarded [39,59].

For this reason, we experiment a new lightweight Convolutional Neural Network (CNN), named GeoMultiTaskNet (GeoMTNet), on a new dataset (FLAIR i.e., French Land cover from Aerospace ImageRy [15]), properly shaped for UDA (see for example the radiometric shifts in Fig. 1). This contribution is the first in which the FLAIR dataset is used for scientific purposes.

GeoMTNet is a novel algorithm for UDA in semantic segmentation of RS images leveraging geographical coordinates, to align the source and target domains, with two key novelties. First, we propose a simple GeoMultiTask module (GeoMT) that learns to predict the geographic position of the input image. Second, inspired by [24], we propose a Dynamic Class Sampling (DCS) module that adapts the semantic segmentation loss to the frequency of the classes.

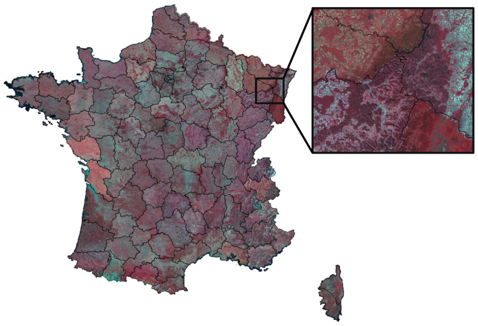

Figure 1. Radiometric discrepancies of the aerial images between domains. The bands displayed are composite of near-infrared, red and green spectral information. Figure adapted from [15].

To our knowledge, this is the first work to address UDA in semantic segmentation using geographical metadata. The proposed approach offers a promising solution for reducing the annotation cost in semantic segmentation of VHR RS images, with a simple and portable module. Our proposed method establishes on a subset of the FLAIR dataset new state-of-the-art performance (47.22% mIoU) with a limited number of parameters (33M), w.r.t. the transformer counterparts (85M).

2. Related Work

2.1. Unsupervised Domain Adaptation

UDA approaches could be divided into three main branches: feature alignment, labeling adjustment, and discriminative methods. Feature alignment methods have the aim of aligning some characteristics (e.g., color histograms or features) of source and target domains. Some examples are DeepCORAL [43], KeepItSimple [2], CoVi [33], GtA [50]. Labeling adjustment makes use of pseudo-labeling to force the predictions of the target domain to be consistent. Several works followed these strategies, such as NoisyStudent [53], CBST [60]. Discriminative methods are based on loss terms that force the net to distinguish among source and target features, e.g., DANN [14], AdaptSegNet [46], ADVENT [48], DADA [49]. Moreover, some hybrid approaches are also developed. For example, we can recall methods based on a combination of the presented strategies, such as SePiCo [52], DISE [7] and DAFormer [17, 18]. Finally, hybrid UDA approaches such as self-supervised learning (SSL) [8, 34, 50] or continual learning [40] have been explored.

2.2. UDA for Remote Sensing

Different methods, not all aiming properly for UDA, have been proposed. StandardGAN [45] works with multi-source domains, forcing the domains to have similar distributions. Seasonal Contrast (SeCo) [30] is based on two steps: gathering uncurated RS images, then, using SSL. Bidirectional sample-class alignment (BSCA) [19] addresses semi-supervised domain adaption for cross-domain scene classification. ConSecutive Pre-Training (CSPT) [57], similarly to [30], aims to leverage knowledge from unlabeled data through a self-supervision approach. MemoryAdaptNet [58] constructs an output space adversarial learning framework to tackle domain shift. UDAT [55] addresses UDA for nighttime aerial tracking, through a transformer. MATerial and TExture Representation Learning (MATTER) [3] aligns domains of different datasets, through a self-supervision task, on several tasks. UDA_for_RS [24], complementing [17], proposes a Gradual Class Weights (GCW) and a Local Dynamic Quality (LDQ) module.

2.3. Using Geographical Metadata

The first attempts at using geoinformation, outside the UDA framework, were presented in [25, 44]. In [28], the authors provide a comprehensive review of location encoding. [27] proposes an efficient spatiotemporal prior, that estimates the probability that a given object category occurs at a specific location. GeoKR [23] uses metadata for an efficient pre-training strategy on a wide dataset. In [5], geographical coordinates are used for map translation. Geography-Aware SSL [4] proposes an SSL algorithm based on the geoinformation of the patches. In [29], the authors present Space2Vec to encode the absolute positions and location spatial relationships. PE-GNN [22] follows a similar approach, using graphs.

3. Methodology

As stated, the aim of GeoMTNet is to reduce domain shift using geographic coordinates, by designing a lightweight and easy-to-use architecture. In particular, given a set of source images $\mathbf{X}_S \in \mathbb{R}^{H \times W \times B}$, where H is the height of the images, W is the width and B are the bands of the images in input, and a set of target images $\mathbf{X}_T \in \mathbb{R}^{H \times W \times B}$, we want to predict the annotation maps of the target, $\hat{\mathbf{Y}}_T \in \mathbb{R}^{H \times W}$, making use only of $\mathbf{X}_S$, $\mathbf{X}_T$ and the labels of the source images, $\mathbf{Y}_S \in \mathbb{R}^{H \times W}$. The labels of the target images, $\mathbf{Y}_T \in \mathbb{R}^{H \times W}$, could only be used for evaluation purposes. To achieve these targets, we decided to adopt a classic U-Net [36] as the backbone model. It is easy to train and commonly used. It has a reduced number of parameters w.r.t. transformer counterparts. It exploits a semantic segmentation training, through the use of a pixel level-classification loss. On the other hand, to tackle domain shift, visible in Fig. 1, the GeoMultiTask (GeoMT) module is added. The net is further trained with the Dynamic Class Sampling (DCS) strategy. Both of them have been shaped to be easily portable to different architectures. First, the GeoMT makes use of the geographical coordinates

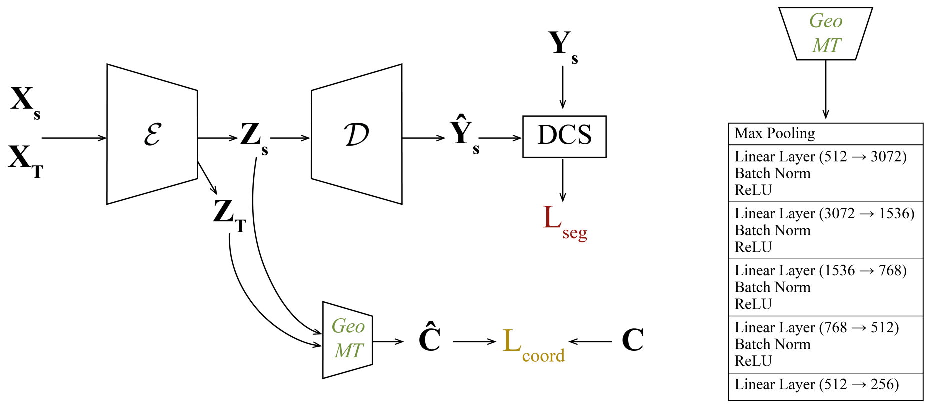

Figure 2. On the left the overview of the proposed architecture, made of: an encoder ($\mathcal{E}$), a decoder ($\mathcal{D}$), the GeoMultiTask module, and the Dynamic Class Sampling module. On the right, the structure of the GeoMultiTask module with input and output sizes.

as a proxy to supervise the domain shift. Inspired by different self-supervised approaches [4, 30], we consider that an effective method to improve the performance on the target domains is to constrain, through a loss term, the features of the encoder to understand where the target images are located. Considering that also the source domain is made of different sub-domains (i.e. departments), the GeoMT is employed to constrain the encoder to learn generalized representations of all the data. Second, as the distribution of the labels is skewed, we propose DCS, to limit the errors on the under-represented classes, inspired by [24]. The whole architecture is shown in Fig. 2. In the next sections, the GeoMultiTask module (Section 3.1) and the Dynamic Class Sampling (Section 3.2) are presented in a formal and detailed way.

3.1. GeoMultiTask Module

In other EO tasks, some approaches used geographical coordinates, such as using them as residual [5] or skip connections, or even being stacked to the input [54]. In our case, inspired by SSL [4, 30], we decided to use coordinates to drive and constrain the encoder features. Specifically, both $\mathbf{X}_S$ and $\mathbf{X}_T$ images pass through the encoder $\mathcal{E}$. This results in $\mathbf{Z}_S \in \mathbb{R}^{H' \times W' \times C}$ and $\mathbf{Z}_T \in \mathbb{R}^{H' \times W' \times C}$ feature maps, where H', W' and C are the height, width and number of channels of the feature maps. $\mathbf{Z}_S$ passes through the decoder $\mathcal{D}$ to obtain $\hat{\mathbf{Y}}_S$. In parallel, both representations $\mathbf{Z}_S$ and $\mathbf{Z}_T$ enter the GeoMT to predict a vector $\widehat{\mathbf{C}} \in \mathbb{R}^D$ containing localization information. Specifically, each patch is assigned a pair of coordinates (C_{lon}, C_{lat}), referring to the centroid of the patch itself. These coordinates undergo the following transformations to be used as supervision for $\widehat{\mathbf{C}}$:

1. centering them in the reference system EPSG:2154, w.r.t. whom the coordinates are expressed. Particularly, we subtract $x = 489353.59\ m$ to C_{lon} and $y = 6587552.20\, m$ to C_{lat}, to make the median values equal $(0, 0)$;
2. noise injection of 30 km to let the net capture large-scale patterns not too specifically referred to the patches in the batch, but rather to wider areas of France, that may eventually even cross the boundaries of individual departments;
3. positional encoding of the coordinates, for similar reasons of the noise injection. The strategy uses the following formula:

$$\mathbf{C} = \begin{bmatrix} \sin\left(C_{lon}\omega_1\right) \\ \cos\left(C_{lon}\omega_1\right) \\ \vdots \\ \sin\left(C_{lat}\omega_{d/4}\right) \\ \cos\left(C_{lat}\omega_{d/4}\right) \end{bmatrix}_d \quad \text{with } \omega_i = \frac{1}{f^{2i/d}}, \tag{1}$$

where $D = 256$ and $f = 20,000$. Particularly, for the same reasons of the noise injection, f is set to 20,000 and not 10,000 like in most applications [5, 47].

GeoMT consists, firstly, of a max-pooling layer, which is used to reduce dimensionality and select the most meaningful features. After this, 5 linear layers, 4 of which employ batch normalization and ReLUs, are stacked. The detailed sizes are given in right part of Fig. 2. This module produces $\widehat{\mathbf{C}}$ from which we compute the self-supervised loss L_{coord} that has the form of a mean squared error:

$$L_{coord} = \frac{1}{n} \sum_{i=1}^{n} \left(\widehat{\mathbf{C}}_i - \mathbf{C}_i\right)^2, \tag{2}$$

where n is the number of samples.

The final loss of the GeoMTNet is thus:

$$L = L_{seg} + L^{S}_{coord} + L^{T}_{coord}, \tag{3}$$

where L_{seg} is the segmentation loss, computed among $\widehat{\mathbf{Y}}_S$ and $\mathbf{Y}_S$, L^{S}_{coord} is the loss term referred to the source domain images, and L^{T}_{coord} to the target ones.

3.2. Dynamic Class Sampling

Class imbalance is a common problem in deep learning, that leads to poor model generalization, especially in rare classes. To address this issue, researchers have proposed various methods, such as assigning class weights inversely proportional to the frequency of the class in the dataset [60]. The class weight for class c, referred to the n-th label, is calculated as follows:

$$w(n,c) = \frac{N_c \cdot \exp\left[(1-f_c)/t\right]}{\sum_{c'=1}^{C} \exp\left[(1-f_{c'})/t\right]}, \tag{4}$$

where f_c is the frequency of class c in the training dataset, N_c is the total number of classes, and t is a temperature parameter. The frequency f_c is calculated as:

$$f_c = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} \left(y_S^{(h,w)} \right)_c, \tag{5}$$

where $y_S^{(h,w)}$ denotes the one-hot source label at location (h, w) in the image, and $(\cdot)_c$ denotes the c-th scalar of a vector. Inspired by [24], which applies a similar mechanism to the pseudo-labels predicted by the student network, the class weight is updated iteratively for each image using an exponentially weighted average:

$$\mathrm{DCS}(n,c) = \alpha \cdot \mathrm{DCS}(n-1,c) + (1-\alpha) \cdot w(n,c). \tag{6}$$

α is the decay rate of the exponential average. It helps to reduce volatility, especially in the early stages of training. Unlike other approaches [24], this weighting strategy does not impact the pseudo-labels but the predicted labels directly. The distribution of the classes will be different from the whole dataset in advance, due to sampling randomness: the weights will be updated iteratively for each image. It is also worth noting that, instead of directly initializing the class weights to the distributions estimated from the first sample, they are initialized to 1 and then updated iteratively by an exponentially weighted average. A higher t leads to a more uniform distribution. A lower one makes the model pay more attention to the rare classes.

The final segmentation loss is:

$$L_{seg} = -\sum_{h=1}^{H} \sum_{w=1}^{W} \mathrm{DCS}(n,c) \cdot y_S^{(h,w)} \cdot \log\left(h_\theta \left(x_S^{(h,w)} \right) \right), \tag{7}$$

where h_θ is the model with weights θ.

4. Dataset

The French National Institute of Geographical and Forest Information (IGN) [1] is a French public state administrative establishment in charge of measuring large-scale changes on the French territory. It is constructing the French national reference land cover map *Occupation du sol à grande échelle* (OCS-GE), also making use of AI-based data and techniques. To this purpose, IGN developed the FLAIR dataset[1].

4.1. FLAIR dataset

The French Land cover from Aerospace ImageRy (FLAIR) dataset [15] includes 50 spatial domains varying along the different landscapes and climates of metropolitan France. Each domain is a French department (Fig. 3).

The complete dataset is composed of 77,412 patches, covering approximately 810 km^2. Each patch is 512 × 512 pixels, with a ground sample distance (GSD) of $0.2m$. Each domain is composed of $1725 - 1800$ patches. The domains were selected considering the major landscapes (e.g., urban, agricultural, etc..) and per semantic class radiometries (see Fig. 1). To acquire the images, more than three years were needed. This led to a high intra- and inter-domain variance in the acquisitions (see Fig. 3 and 1). The images have 5 bands corresponding to blue, green, red, near-infrared and elevation channels. The first 4 channels are retrieved from VHR aerial images ORTHO HR® [12]. The fifth channel is obtained through the difference between the Digital Surface Model and the Digital Terrain Model (see [15] for more details). The corresponding ground truth labels describe the semantic class for each pixel. Nineteen classes are annotated. The *other* class corresponds to pixels impossible to define with certainty. Finally, the dataset is split into 40 domains for the training and 10 for testing, ensuring a comparable distribution of the labels in train and test. The domains are highlighted in Fig. 3. Each patch is enriched with metadata:

- domain and zone label. The zone label is made of two letters, allowing a macro-distinction of the two major types of land cover of the area. The letter U indicates urban, N natural area, A agricultural area, and F forest.
- date and hour at which the aerial image was acquired;
- the geographical coordinates of the centroid and the mean altitude of the patch;
- camera type used during aerial image acquisition [42].

To our knowledge, this is the first time that this dataset has been used for scientific research. Particularly, in our experiments, we used a subset of the whole dataset: D06, D08,

[1] downloadable at https://ignf.github.io/FLAIR/

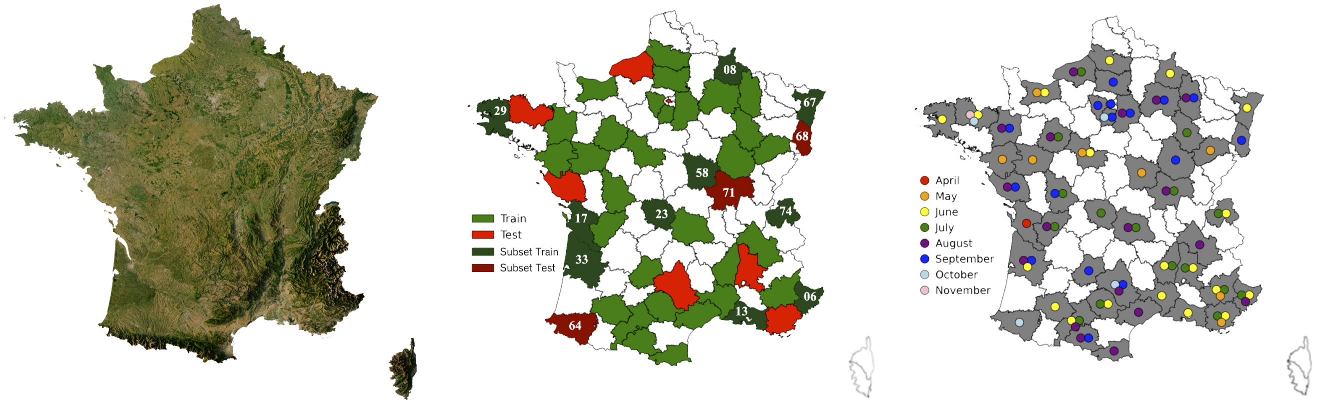

Figure 3. On the left, the ORTHO HR® aerial image cover of France. On the center, the train and test split of the 50 domains, with the domains selected for our experiments highlighted. On the right, the acquisition time of each domain. Figure adapted from [15].

D13, D17, D23, D29, D33, D58, D67, D74 as source domains and D64, D68, D71 as target domains. We ended up with more than 16k images for training and more than 5k for testing.

5. Experimental Setup

As stated, for our experiments, we selected 10 departments as source domains and 3 as target domains. Adopting the same strategy as [15], we considered as *other* all the classes labeled as > 12. These classes are strongly under-represented, being $< 0.2\%$ of all the labels. Thus, we ended up with 13 classes (i.e. *building*, *pervious surface*, *impervious surface*, *bare soil*, *water*, *coniferous*, *deciduous*, *brushwood*, *vine*, *grassland*, *crop*, *plowed land*, *other*). A single Tesla V100-SXM2 32 GB GPU was used for the training phase. Having limited computational power, but still wanting to preserve the high resolution of the dataset (GSD = $0.2\ m$), for the training, we used random crops of 256×256. For the testing stage, we perform inference on four non-overlapping crops of 256×256, for each patch of size 512×512. For the U-Net, we use a ResNet18 [16], pretrained on ImageNet, as encoder and the softmax function as activation on the last layer. For all the experiments, we fix the batch size to 16, the number of epochs to 120 and the learning rate to 0.0001. We used early stopping with a patience of 30 epochs. The semantic segmentation loss is a cross-entropy, ignoring the *other* class. We used Adam as optimizer and RandAugment [11] as the set of augmentations. The mean intersection over union (mIoU) on the first 12 classes is the selected evaluation metric. For the DCS module, we set the parameters to $T = 0.9$ and $\alpha = 0.7$. To assess the performance of our strategy, we selected different methods[2] from the literature for an extensive comparison. We chose: AdaptSegNet [46], which employs an adversarial training approach; ADVENT [48], using an entropy minimization strategy; DAFormer [17], which adopts a transformer with a self-training strategy and UDA_for_RS [24], that optimizes the DAFormer for RS tasks. In Section 6, we present different experimental results, addressing both comparisons and several ablation studies.

6. Experimental Results

GeoMTNet reaches satisfying performance, shown in Tables 1 and 2. As expected, the under-represented classes, such as *coniferous* and *brushwood*, are the most difficult to be correctly predicted. This is due both to the few quantities of data and the radiometric similarity with some more frequent classes. For example, *coniferous* could be easily confused with *deciduous*. At the same time, some errors are due to the fact that images share some similar spatial patterns. This is the case of *vine* and *crop* pixels. Another frequent misclassification error concerns *bare soil*. Even though the performance is satisfying (55% mIoU), we can see that the variance implicit in the definition of this class led to confusion with *herbaceous cover* or *impervious surface*. In the next sections, comparisons (Section 6.1) and ablation studies (Sections 6.2, 6.3 and 6.4) are carried one.

6.1. Comparison

Despite using a smaller number of parameters (33M), GeoMTNet reaches better results than all the other selected architectures (47.22% mIoU). In particular, we can see from Table 1 that there is a deep gap w.r.t. AdaptSegNet [46] (24.97% mIoU) and ADVENT [48] (25.56% mIoU), which are more dated and, probably, properly developed for the synthetic-to-real benchmarks [35, 37]. On the other hand, DAFormer [17] and UDA_for_RS [24], based on a transformer, have comparable performance with the GeoMTNet (respectively 45.61% and 47.02% mIoU). When using strategies properly shaped for RS task, such as in

[2] We tested them through the code in their official GitHub repositories.

UDA_for_RS [24], optimal results are obtained. However, from both Table 1, Table 2 and Fig. 4, we can see that GeoMTNet leads to better results also w.r.t. the aforementioned method with a reduced number of parameters (33M for GeoMTNet vs 85M for UDA_for_RS). Focusing on the detailed performance, reported in Table 2, we can observe that GeoMTNet almost has the best performance on all the classes, except for four of them (that are *pervious surface*, *bare soil*, *brushwood*, and *vine*). This is mainly justified by the fact that each different architecture tends, when deciding among two similar classes, to overestimate one of them and underestimate the other. For example, *vine* is often confused with *plowed land* (and sometimes *crop*, too), due to their similar pattern. DAFormer, still having a gain of more than 10% in IoU over GeoMTNet performance for *vine*, reaches poor results both on *plowed land* (41.83% vs 54.79% in IoU) and *crop* (23.74% vs 35.02% in IoU). This phenomenon could be observed also in Fig. 4, where some predictions of the three best models (namely DAFormer, UDA_for_RS, and GeoMTNet) are reported to draw some qualitative results. We observe that DAFormer performs overall worse, as it often predicts some irrelevant classes, with a poorer texture and shape of the polygons predicted. On the other hand, most of the time UDA_for_RS predicts a smaller number of classes with a wider predicted area for each of them w.r.t. the other methods. This is mainly due to the LDQ module of UDA_for_RS, which bases the pixel prediction on the predictions made on the contiguous pixels. This can be seen both in positive cases, Fig. 4 b), where the land cover prediction of the traffic circle is more consistent than in GeoMTNet, and in negative cases, Fig. 4 d), where the low confidence in predicting *pervious surface* and *building* ends in a uniformed incorrect prediction of *impervious surface*. On the other hand, we can appreciate the consistency in shape reconstructions and boundaries in GeoMTNet more than in the others (see, for example, the building edges in Fig. 4 c)). Moreover, we can see how shadows consist in an important problem for all the architectures (see for example in Fig. 4 a) how the shape of the *plowed land* in the upper right part of the image is badly reconstructed for both UDA_for_RS and GeoMTNet). Another issue to consider is that train patches are of size 512×512 while the model is trained on 256×256 patches. Thus, sometimes, the borders of the predicted tiles to have contrasting predictions, as visible in the central part of Fig. 4 e).

6.2. GeoMultiTask module

To understand the GeoMTNet capabilities, various informative ablation studies have been conducted. To perform these experiments in an easy and rapid way, we used a simple U-Net [36] as CNN, with less than 2M parameters. As mentioned before, the GeoMT takes as input the features provided by the encoder and tries to infer low-frequency encoded coordinates, with a random noise injection. As it could be argued from Tables 3 and 4, both of these strategies improve the net performance.

Architecture	mIoU (%)	params (M)
AdaptSegNet [46]	24.97	99
ADVENT [48]	25.56	99
DAFormer [17]	45.61	85
UDA_for_RS [24]	47.02	85
GeoMultiTaskNet **(ours)**	**47.22**	**33**

Table 1. Our GeoMultiTaskNet outperforms all the other methods on the considered FLAIR target domains. In addition to the improved results in terms of mIoU, the size of the proposed model is also significantly smaller than that of the other selected algorithms.

At first, we focused on the challenge of using coordinates, still feeding the GeoMT with the decoder features. As stated, the goal is to allow the net to capture large-scale patterns, not too specific to the single patch, but rather to areas of France that may eventually even cross the boundaries of individual departments. We tried two strategies: positional encoding and noise injection. Positional encoding [47], used in EO approaches [5], allows the coordinates to be represented with a vector, making it easier to grasp reciprocal phenomena of proximity between patches. Noise injection allows to make net performance more generalizable, avoiding the association of a specific coordinate with a specific patch. In light of these considerations, we have tried different configurations (Table 3). Notably, using a lower frequency (i.e., 1/20000) than the one used in the literature [5, 47], brings greater benefits. In fact, we are interested in large-scale effects, enhanced by a lower frequency. Concerning noise injection, it has been empirically demonstrated that a consistent noise (30 km) compared to the size of a patch (about 100 m) helps the generalization process. However, increasing it too much (about 50 km) leads to excessive network confusion and a consequent drop in performance.

Secondly, we needed to limit the number of parameters, especially w.r.t. the other models in the literature. To do this, we have no longer used as input of the GeoMT the features in output from the decoder, but those in output from the encoder. In fact, the encoder features should already provide the necessary information to perform a correct segmentation. This intuition was supported by the results, shown in Table 4, which shows a slight drop in performance, completely negligible. Notably, the shape of the GeoMT is also slightly different. In the case described so far (input features of the decoder), the module consists of two convolutional layers (to reduce the dimensionality of the features with a limited number of parameters) followed by three linear layers.

Architecture	IoU (%)											
	building	pervious surface	impervious surface	bare soil	water	coniferous	deciduous	brushwood	vine	grassland	crop	plowed land
AdaptSegNet [46]	39.98	20.75	40.23	20.36	15.25	4.93	35.37	10.99	34.51	42.69	11.06	23.47
ADVENT [48]	35.79	24.38	48.82	6.85	31.98	0.00	51.65	11.79	33.33	25.76	11.46	24.29
DAFormer [17]	67.09	45.56	61.99	55.35	65.12	8.91	54.39	**20.31**	**64.39**	38.79	23.74	41.83
UDA_for_RS [24]	66.3	**48.05**	62.36	**59.28**	61.24	9.22	60.02	16.52	57.74	40.12	30.32	54.17
GeoMultiTaskNet **(ours)**	**67.53**	40.86	**63.89**	55.31	**67.02**	**13.85**	**60.97**	14.08	53.09	**40.33**	**35.02**	**54.79**

Table 2. Comparison in the IoU for each class of the considered FLAIR target domains.

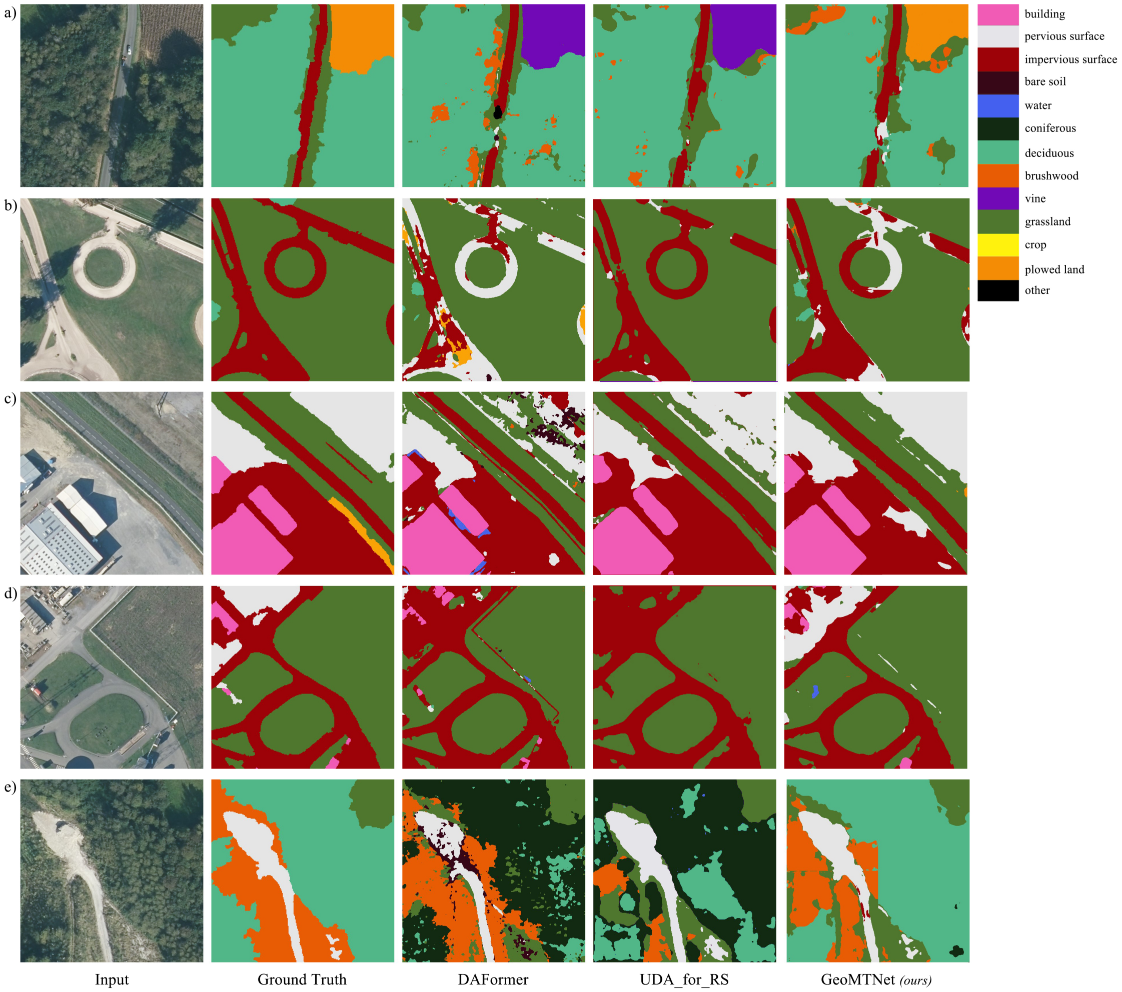

Figure 4. Some examples of predictions for the best performing models. Particularly we can see in order: the input image, the ground truth, the prediction of DAFormer, the prediction of UDA_for_RS and the prediction of our GeoMTNet.

6.3. GeoTimeMultiTask experiments: using the temporal information

We tried to include temporal information as well, also inspired by other works [4, 30]. In particular, we inserted both month and time of day information, discarding the year. In fact, the month impacts the seasonality of some classes (e.g., the vegetative ones), and the hour the acquisition conditions [10]. As previously, we tried to inject some noise, so that the features could generalize better. Specifically, the time information was circle encoded (i.e., arranged equally spaced on a circle) and, when used, a random noise of ±1 was added. Finally, these experiments were carried out using either the encoder or the decoder features. In both circumstances, the TimeMultiTask module (TimeMT) has been defined similarly to the GeoMT, but smaller in size. For example, in the case of using the encoder features, the TimeMT consists of one max-pooling layer and two linear layers. We refer to these experiments as GeoTimeMT, being characterized by both GeoMT and

noise (km)	1/frequency (-)	mIoU (%)	params (M)
-	-	42.05	1.9
-	-	41.69	270
-	10,000	43.57	270
±30	10,000	43.69	270
±30	20,000	44.83	270
±50	20,000	42.68	270

Table 3. Ablation studies, showing the behavior of GeoMultiTaskModule under different noise injections and encoding.

input features	mIoU (%)	params (M)
-	42.05	1.9
output of the decoder	44.83	270
output of the encoder	44.70	11.2

Table 4. Ablation studies, showing the behavior of GeoMultiTaskModule when using different input features.

input features	time used	time noise	mIoU (%)	params (M)
-	-	-	42.05	1.9
decoder	both	no	38.19	405
encoder	both	no	42.72	11.9
encoder	month	yes	43.77	11.9

Table 5. Ablation studies, showing the behavior of GeoTimeMultiTaskModule under different conditions.

TimeMT modules. Also for this set of experiments, a simple U-Net was used as the backbone, without ResNet as the encoder. The results are shown in Table 5. Two behaviors can be observed immediately: using temporal metadata leads to limited improvements (+1.72% mIoU w.r.t. the baseline); GeoTimeMT, which combines geographical and temporal information, does not improve results obtained using only GeoMT (43.77% vs 44.70% mIoU). For these reasons, our GeoMTNet makes only use of geographical coordinates.

Analyzing the detailed results, we can observe that using the features outputted by the encoder is more beneficial than the ones outputted by the decoder, both from a performance (38.19% vs 42.72% mIoU) and size point of view (405M vs 11.9M parameters). In fact, the benefits derived from temporal metadata are more related to the features directly encoded from the images, such as radiometric information of the images, more than to the decoded representations of the patches, such as the one connected to land cover. In addition, we observe again that using less precise information, thus with noise injection, leads to better results (42.72% vs 43.77% mIoU). Finally, we observe that the hour information is less relevant than the month information. In fact, the large variance of the dataset and the large amount of images, make it more important and beneficial to have representations from different seasons of same classes more than under different light conditions. In fact, classes such as *brushwood* or *crop* really vary their radiometric information depending on the seasonality.

6.4. Comprehensive baselines

We considered important to evaluate the impact of each component of the GeoMTNet. The results of these experiments are shown in Table 6.

net	mIoU (%)	params (M)
baseline	42.51	25
+GeoMT	46.68	32.7
+DCS	43.25	25
GeoMTNet	**47.22**	32.7

Table 6. Ablation studies about the component of the GeoMultiTaskNet. As stated, both components lead to better results than the baseline, even though the GeoMultiTask module performs better.

The component that leads to the greatest improvement is the GeoMT which leads to a gain of about 4% in mIoU, while DCS does not go beyond a percentage point. This is due to the fact that GeoMT is properly shaped to enhance RS metadata, to empower the architecture on which it is applied. In contrast, unlike other approaches such as [24], in our GeoMTNet, the weighting module, namely DCS, does not impact the pseudo-labels, but the predicted labels directly. Therefore, its effectiveness on images from target domains is influenced directly by the source images.

7. Conclusions

In light of major technological innovations, more and more RS images are available. However, the annotation of these images is not progressing at the same rate, leading to a vast amount of unlabelled data. Most of the time, these images carry metadata, which are often simply discarded for CV tasks. In this work, we showed that the use of architectures specifically designed to exploit such metadata in an EO context can lead to excellent results. To this end, we proposed GeoMultiTaskNet, which outperforms other models in the literature, despite being a lightweight network, on the FLAIR dataset. This real-world scenario-oriented dataset presents a great variety of information and is well-suited for this type of experiments. In this context, this work only presents itself as a first step in a line of research that is as important as it is still under-investigated: remote sensing unsupervised domain adaptation. Future steps include the extension of GeoMultiTaskNet over the entire FLAIR dataset. In addition, the intention is to probe this model on other datasets [51], where domain shift is more important, and to find new ways to integrate geo-metadata into already performing models, such as transformers.

References

[1] Institut national de l'information géographique et forestière. https://www.ign.fr, 2022. [Accessed: 02 February 2023]. 4

[2] Alexey Abramov, Christopher Bayer, and Claudio Heller. Keep it simple: Image statistics matching for domain adaptation. *arXiv preprint arXiv:2005.12551*, 2020. 2

[3] Peri Akiva, Matthew Purri, and Matthew Leotta. Self-supervised material and texture representation learning for remote sensing tasks. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8203–8215, June 2022. 2

[4] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *Proc. of the IEEE/CVF International Conference on Computer Vision*, pages 10181–10190, 2021. 2, 3, 7

[5] Luc Baudoux, Jordi Inglada, and Clément Mallet. Toward a yearly country-scale corine land-cover map without using images: A map translation approach. *Remote Sensing*, 13(6), 2021. 2, 3, 6

[6] Michael Max Bühler, Christoph Sebald, Diana Rechid, Eberhard Baier, Alexander Michalski, Benno Rothstein, Konrad Nübel, Martin Metzner, Volker Schwieger, Jan-Albrecht Harrs, et al. Application of copernicus data for climate-relevant urban planning using the example of water, heat, and vegetation. *Remote sensing*, 13(18):3634, 2021. 1

[7] Wei-Lun Chang, Hui-Po Wang, Wen-Hsiao Peng, and Wei-Chen Chiu. All about structure: Adapting structural information across domains for boosting semantic segmentation. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1909, 2019. 2

[8] Mu Chen, Zhedong Zheng, Yi Yang, and Tat-Seng Chua. Pipa: Pixel-and patch-wise self-supervised learning for domain adaptative semantic segmentation. *arXiv preprint arXiv:2211.07609*, 2022. 2

[9] Gong Cheng, Xingxing Xie, Junwei Han, Lei Guo, and Gui-Song Xia. Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:3735–3756, 2020. 1

[10] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David B. Lobell, and Stefano Ermon. SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 7

[11] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition workshops*, pages 702–703, 2020. 5

[12] data.gouv.fr. Orthophotographie Haute Résolution - ORTHO® HR 2020 (reformatée) - Calvados. https://www.data.gouv.fr/fr/datasets/orthophotographie-haute-resolution-ortho-r-hr-2020-reformatee-calvados/, 2020. [Accessed: 23 January 2023]. 4

[13] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 1

[14] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 2

[15] Anatol Garioud, Stéphane Peillet, Eva Bookjans, Sébastien Giordano, and Boris Wattrelos. Flair #1: semantic segmentation and domain adaptation dataset. *arXiv preprint arXiv:2211.12979v4*, 2022. 1, 2, 4, 5

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[17] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9924–9935, June 2022. 1, 2, 5, 6, 7

[18] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In *Proc. of European Conference on Computer Vision, Part XXX*, pages 372–391. Springer, 2022. 1, 2

[19] Wei Huang, Yilei Shi, Zhitong Xiong, Qi Wang, and Xiao Xiang Zhu. Semi-supervised bidirectional alignment for remote sensing cross-domain scene classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 195:192–203, 2023. 2

[20] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4893–4902, 2019. 1

[21] Ronald Kemker, Carl 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

[22] Konstantin Klemmer, Nathan Safir, and Daniel B Neill. Positional encoder graph neural networks for geographic data. *arXiv preprint arXiv:2111.10144*, 2021. 2

[23] Wenyuan Li, Keyan Chen, Hao Chen, and Zhenwei Shi. Geographical knowledge-driven representation learning for remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022. 2

[24] Weitao Li, Hui Gao, Yi Su, and Biffon Manyura Momanyi. Unsupervised domain adaptation for remote sensing semantic segmentation with transformer. *Remote Sensing*, 14(19):4942, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[25] Shuai Liao, Xirong Li, Heng Tao Shen, Yang Yang, and Xiaoyong Du. Tag features for geo-aware image classification. *IEEE transactions on multimedia*, 17(7):1058–1067, 2015. 2

[26] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. *Advances in neural information processing systems (NeurIPS)*, 29, 2016. 1

[27] Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. In *Proc. of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 9596–9606, 2019. 2

[28] Gengchen Mai, Krzysztof Janowicz, Yingjie Hu, Song Gao, Bo Yan, Rui Zhu, Ling Cai, and Ni Lao. A review of location encoding for geoai: methods and applications. *International Journal of Geographical Information Science*, 36(4):639–673, 2022. 2

[29] Gengchen Mai, Krzysztof Janowicz, Bo Yan, Rui Zhu, Ling Cai, and Ni Lao. Multi-scale representation learning for spatial feature distributions using grid cells. *arXiv preprint arXiv:2003.00824*, 2020. 2

[30] Oscar Manas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodriguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *Proc. of the IEEE/CVF International Conference on Computer Vision*, pages 9414–9423, 2021. 2, 3, 7

[31] Valerio Marsocci and Simone Scardapane. Continual barlow twins: continual self-supervised learning for remote sensing semantic segmentation. *arXiv preprint arXiv:2205.11319*, 2022. 1

[32] Valerio Marsocci, Simone Scardapane, and Nikos Komodakis. MARE: Self-supervised multi-attention resu-net for semantic segmentation in remote sensing. *Remote Sensing*, 13(16):3275, 2021. 1

[33] Jaemin Na, Dongyoon Han, Hyung Jin Chang, and Wonjun Hwang. Contrastive vicinal space for unsupervised domain adaptation. In *Proc. of the European Conference on Computer Vision (ECCV)*, pages 92–110. Springer, 2022. 2

[34] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3764–3773, 2020. 2

[35] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Proc. of the European Conference on Computer Vision (ECCV), Part II 14*, pages 102–118. Springer, 2016. 5

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2, 6

[37] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3234–3243, 2016. 5

[38] Franz Rottensteiner, Gunho Sohn, Jaewook Jung, Markus Gerke, Caroline Baillard, Sebastien Benitez, and Uwe Breitkopf. The ISPRS benchmark on urban object classification and 3D building reconstruction. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences I-3 (2012), Nr. 1*, 1(1):293–298, 2012. 1

[39] Sancho Salcedo-Sanz, Pedram Ghamisi, María Piles, Martin Werner, Lucas Cuadra, A Moreno-Martínez, Emma Izquierdo-Verdiguier, Jordi Muñoz-Marí, Amirhosein Mosavi, and Gustau Camps-Valls. Machine learning information fusion in Earth observation: A comprehensive review of methods, applications and data sources. *Information Fusion*, 63:256–272, 2020. 1

[40] Antoine Saporta, Arthur Douillard, Tuan-Hung Vu, Patrick Pérez, and Matthieu Cord. Multi-head distillation for continual unsupervised domain adaptation in semantic segmentation. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3751–3760, 2022. 2

[41] Andrii Shelestov, Hanna Yailymova, Bohdan Yailymov, Leonid Shumilo, and AM Lavreniuk. Extension of copernicus urban atlas to non-european countries. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 6789–6792. IEEE, 2021. 1

[42] Jean-Philippe Souchon, Christian Thom, Christophe Meynard, and Olivier Martin. A large format camera system for national mapping purposes. *Revue Francaise de Photogrammétrie et de Télédétection*, page 48–53, avr. 2014. 4

[43] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Proc. of the European Conference on Computer Vision (ECCV)*, pages 443–450. Springer, 2016. 2

[44] Kevin Tang, Manohar Paluri, Li Fei-Fei, Rob Fergus, and Lubomir Bourdev. Improving image classification with location context. In *Proc. of the IEEE international conference on computer vision (ICCV)*, pages 1008–1016, 2015. 2

[45] Onur Tasar, Yuliya Tarabalka, Alain Giros, Pierre Alliez, and Sebastien Clerc. Standardgan: Multi-source domain adaptation for semantic segmentation of very high resolution satellite images by data standardization. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 2

[46] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker. Learning to adapt structured output space for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 5, 6, 7

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3, 6

[48] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019. 2, 5, 6, 7

[49] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Mathieu Cord, and Patrick Pérez. Dada: Depth-aware domain adaptation in semantic segmentation. In *ICCV*, 2019. 2

[50] Qin Wang, Dengxin Dai, Lukas Hoyer, Luc Van Gool, and Olga Fink. Domain adaptive semantic segmentation with

self-supervised depth estimation. In *Proc. of the IEEE/CVF International Conference on Computer Vision*, 2021. 2

[51] Junshi Xia, Naoto Yokoya, Bruno Adriano, and Clifford Broni-Bediako. Openearthmap: A benchmark dataset for global high-resolution land cover mapping. In *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6254–6264, 2023. 8

[52] Binhui Xie, Shuang Li, Mingjia Li, Chi Harold Liu, Gao Huang, and Guoren Wang. Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation. *arXiv preprint arXiv:2204.08808*, 2022. 1, 2

[53] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proc. of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020. 2

[54] Naisen Yang and Hong Tang. Semantic segmentation of satellite images: A deep learning approach integrated with geospatial hash codes. *Remote Sensing*, 13(14):2723, 2021. 3

[55] Junjie Ye, Changhong Fu, Guangze Zheng, Danda Pani Paudel, and Guang Chen. Unsupervised domain adaptation for nighttime aerial tracking. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8896–8905, 2022. 2

[56] Bing Zhang, Yuanfeng Wu, Boya Zhao, Jocelyn Chanussot, Danfeng Hong, Jing Yao, and Lianru Gao. Progress and challenges in intelligent remote sensing satellite systems. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:1814–1822, 2022. 1

[57] Tong Zhang, Peng Gao, Hao Dong, Yin Zhuang, Guanqun Wang, Wei Zhang, and He Chen. Consecutive pre-training: A knowledge transfer learning strategy with relevant unlabeled data for remote sensing domain. *Remote Sensing*, 14(22), 2022. 2

[58] Jingru Zhu, Ya Guo, Geng Sun, Libo Yang, Min Deng, and Jie Chen. Unsupervised domain adaptation semantic segmentation of high-resolution remote sensing imagery with invariant domain-level context memory. *arXiv preprint arXiv:2208.07722*, 2022. 2

[59] Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE geoscience and remote sensing magazine*, 5(4):8–36, 2017. 1

[60] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proc. of the European Conference on Computer Vision (ECCV)*, pages 289–305, 2018. 2, 4