SegDesicNet: Lightweight Semantic Segmentation in Remote Sensing with Geo-Coordinate Embeddings for Domain Adaptation

Sachin Verma sachin.verma@ntnu.no
Frank Lindseth frankl@ntnu.no
Gabriel Kiss gabriel.kiss@ntnu.no
Norwegian University of Science and Technology (NTNU), Norway

Abstract

Semantic segmentation is essential for analyzing high-definition remote sensing images (HRSIs) because it allows the precise classification of objects and regions at the pixel level. However, remote sensing data present challenges owing to geographical location, weather, and environmental variations, making it difficult for semantic segmentation models to generalize across diverse scenarios. Existing methods are often limited to specific data domains and require expert annotators and specialized equipment for semantic labeling. In this study, we propose a novel unsupervised domain adaptation technique for remote sensing semantic segmentation by utilizing geographical coordinates that are readily accessible in remote sensing setups as metadata in a dataset. To bridge the domain gap, we propose a novel approach that considers the combination of an image's location-encoding trait and the spherical nature of Earth's surface. Our proposed SegDesicNet module regresses the GRID positional encoding of the geo-coordinates projected over the unit sphere to obtain the domain loss. Our experimental results demonstrate that the proposed SegDesicNet outperforms state-of-the-art domain adaptation methods in remote sensing image segmentation, achieving an improvement of approximately 6% in the mean intersection over union (MIoU) with a $\sim$ 27% drop in parameter count on benchmarked subsets of the publicly available FLAIR #1 dataset. We also benchmarked our method performance on the custom split of the ISPRS Potsdam dataset. Our algorithm seeks to reduce the modeling disparity between artificial neural networks and human comprehension of the physical world, making the technology more human-centric and scalable.

1. Introduction

Remote sensing technologies have gained widespread use and have significantly transformed the field by making high-resolution remote sensing images (HRSIs) more accessible for a broad range of applications. These HRSIs offer detailed information about the Earth's surface, capturing numerous spectral and spatial characteristics at unparalleled resolution levels, expanding the range of observations that can be made, and enabling the observation of a greater variety of objects. Precise land cover information derived from HRSIs is utilized in various planning, monitoring, and management applications across different sectors, for example, urban [47,58,62] and environmental [69]. The foundation for land cover and land use classification is rooted in semantic segmentation (SS) techniques [60]. SS provides a comprehensive understanding of the spatial distribution of objects and their boundaries. This technique assigns class labels to every pixel in an image, specifically identifying the type of class represented by each pixel. Deep Neural Networks (DNNs) [40] have exhibited outstanding performance in tasks concerning scene comprehension, including SS [9, 30, 59, 65], given consistent data distribution across training and test sets. The difficulty in remote sensing lies within the SS task, given the substantial variability in scenes and styles present in remote sensing images. Variations in data distribution can result from factors such as diverse terrains, different weather conditions, various sensor techniques, and cultural and economic disparities [42, 67]. These discrepancies can critically impact the data distribution, and a model trained effectively on the original dataset may not perform optimally when faced with new data, even if the categories remain consistent. To address this issue, a commonly used approach is transfer learning, which involves leveraging the knowledge gained from the source domain to improve the performance in the target domain. By fine-tuning the pre-trained model on a smaller target domain dataset, the model can adapt to the specific characteristics and variations present in the target domain, thus mitigating the effects of the domain shift. However, this step can be challenging owing to the dynamic nature of the target domain. Semantic labeling of data, a crucial aspect of this process, is a resource-intensive operation that requires substantial investment in time, effort, and expertise. Moreover, maintaining an up-to-date and representa-

tive target domain dataset is an ongoing concern because the evolving nature of real-world data necessitates continuous adjustments to ensure the effectiveness of the model across changing scenarios. Given the abundance of remote sensing images that are available without labels, it is impractical to retrain models specifically for each target domain. Consequently, unsupervised domain adaptation (UDA) [17,24,31] has emerged as a solution for generating high-quality segmentation in the target domain without the need for any annotations, thereby offering a practical and efficient solution for adapting models to new domains in the field of remote sensing. The significance of UDA has been widely recognized and explored in the field of computer vision (CV). Generic CV models are frequently employed for remote sensing image analysis without fully considering the unique characteristics and peculiarities of Earth Observation (EO) data. Numerous attempts have been made to utilize this information, particularly the location context, in non-remote sensing setups [6, 35, 50]. In this article, we propose a novel and highly efficient convolutional neural network called SegDesicNet, drawing inspiration from previous works such as [38, 45, 80]. Our approach focuses on UDA for the SS of remote-sensing images.

SegDesicNet introduces a novel algorithm that utilizes geographical coordinates to align the source and target domains in an unsupervised manner. By integrating scene understanding derived from the cosine similarity of grid-based geographical coordinates projected over a unit sphere as the domain adaptation loss in our SegDesicNet model, we try to enhance the model's ability to learn and adapt to its surroundings effectively. our motivation is to achieve rational modeling of human scene understanding, which is an essential aspect of neural networks. The consistency of this algorithm, that is, employing geo-coordinates in a global reference system, increases its versatility and applicability across any region of the world in a remote sensing setup. The contributions of this study are as follows:

- In this study, we introduce an unsupervised domain adaptation technique for HRSI semantic segmentation, aiming to mimic human comprehension effectively. To achieve this, we integrated GRID positional encoding, inspired by [2], and projected it onto a unit sphere to simulate the Earth's surface.
- Our proposed approach outperforms state-of-the-art (SOTA) methods on the widely recognized FLAIR #1 [1] data subset [38], showing a notable mIoU improvement of over 6%. Furthermore, we validated our approach using the ISPRS Potsdam [48] dataset and achieved superior results in the considered target domain.
- Notably, our network architecture exhibits a remarkable reduction in the parameter countcompared with existing state-of-the-art approaches. Specifically, our network has a 27% lower parameter count, enabling more efficient and lightweight models without compromising performance. This efficiency is crucial for practical applications because it reduces the computational requirements and allows for faster and more resource-efficient inferences.

2. Related Work

2.1. Semantic Segmentation

The advancement of network architecture for SS has seen a notable transition from CNNs [9, 30, 77] to Vision Transformers [10, 15, 65, 70, 77] over time. Another area of research has emerged that focuses on enhancing extracted representations by incorporating attention mechanisms [16, 23, 25, 78] or context representations [29, 59, 70–73] into segmentation models. Within this context, our proposed approach, SegDesicNet, complements existing frameworks and consistently improves several state-of-the-art methods.

2.2. Unsupervised Domain Adaptation

There are multiple approaches to achieving UDA [42]. The following methods can be considered, but they are not limited to: feature alignment [3, 41, 49, 61], discriminate methods [20, 22, 52, 56, 57] and labeling adjustment [66, 76, 82]. The feature alignment approach focuses on aligning the features or characteristics of the source and target domains. The goal was to reduce the distribution discrepancy between the domains. Pseudo-labeling is utilized to adjust or refine the labels of the target domain. The model predictions on the target domain were treated as pseudo-labels, which were then used to enforce consistency in the predictions. Discriminative methods incorporate loss terms or mechanisms that encourage the model to distinguish between source and target domains. These methods aim to learn domain-invariant representations. Instance-based methods [74] adapts instances or samples from the source domain to align them with the target domain. Hybrid approaches [64] combine multiple strategies for a more comprehensive adaptation, integrating features from different domains, combining feature alignment, labeling adjustment, and discriminative methods. Self-supervised learning [5] methods aims to reduce the domain gap between the source and target domains, enhance the adaptability of machine-learning models across different environments or datasets without the need for data labels, and is a widely used technique in computer vision for SS.

2.3. Unsupervised Domain Adaptation for Remote Sensing

Li *et al*. [26] showcased the effectiveness of the transformer [54] in self-training the UDA of remote sensing

images, presenting two strategies, Gradual Class Weights and Local Dynamic Quality, to enhance the performance of the self-training UDA framework. Zhu *et al*. introduced MemoryAdaptNet [79], which constructs an output space adversarial learning scheme and embeds an invariant feature memory module to store invariant domain-level context information, thereby improving the cross-domain semantic segmentation performance. Ma *et al*. [33] aligned domains using centroid and covariance strategies, while Gross *et al*. [18] addressed nonlinear effects and illumination variations by using labeled training spectra to align multiple hyperspectral datasets and navigate towards DNNs. Akiva *et al*. proposed MATTER [4] which learns material and texture representations. Manas *et al*. proposed SeCo [34], which utilizes an unsupervised acquisition procedure and self-supervised learning, whereas Ye *et al*. proposed UDAT [68] which employs a unique object discovery approach to generate training patches. A transformer-based bridging layer was used to align the feature distributions in the nighttime and daytime domains.

2.4. Geographical Metadata

First, geoinformation is used in the classification task [51], where it is concatenated into high-level image features of classification tasks to improve data distinctiveness. Subsequently, several similar methods were developed [7, 46, 55, 63]. Aodha *et al*. [36] modeled the presence of an object category as a function of geolocation and time. Inspired by neuroscience research, [37] introduced Space2Vec to encode the absolute positions and spatial relationships of places. [44] lists a few possibilities for geographic location encoding. To our knowledge, only one specific work has addressed UDA with geo-coordinates for remote sensing [38], in which its DA module regresses the positional embedding of the geo-coordinates. The incorporation of weighted loss in training models involves updating the weights per image using an exponential moving average (EMA), which can lead to computationally intensive training. EMA in certain setups, such as ours, may not align naturally and could introduce unnecessary delays in convergence. This could result in overlooking or underemphasizing crucial updates, ultimately leading to a less accurate final model. The vector incorporates both latitude and longitude values, and the Mean Squared Error (MSE) may not be sufficient to account for the alignment or discrepancy in orientation and direction among the vectors. In contrast to the traditional Euclidean distance or MSE loss, cosine dissimilarity acknowledges that latitude and longitude values should be assessed on a spherical surface rather than in Euclidean space. The spherical nature of the Earth's surface is more accurately represented by cosine dissimilarity than other distance metrics that assume a flat Euclidean space.

Building on this work, we extended the benchmark DA module of [38] such that its learning ability is well utilized in capturing things the way they appear in the physical world, to make it more efficient for DA.

3. Methodology

The datasets used in our experiment contain latitude and longitude information (C_{lon}, C_{lat}) for each image patch.

3.1. Model Architecture

3.1.1 Baseline Semantic Segmentation Architecture

Given a set of source images $\mathbf{X}_S \in \mathbb{R}^{H\times W\times B}$, where H and W are the height and width of the images, respectively, B represents the number of color bands. Semantic labels accompany these images, represented as $\mathbf{Y}_S$ with the same spatial dimension but a single band. We followed the lightweight U-Net [43] as a baseline model to generate a semantic map from images using its RGB channel, similar to its ground truth. To accomplish this, we employed a pre-trained ResNet18 [19] model, which was originally trained on the Imagenet [13] dataset, as the encoder in our U-Net implementation; we manipulated a few last layers in the decoder of this architectures, ending with an lightweight model with merely $\sim$ 24 million parameters. The last layer of the network utilizes softmax activation to produce the final output. This choice of encoder, coupled with softmax activation, allowed us to leverage the powerful feature extraction capabilities of ResNet18 while ensuring the generation of probabilistic predictions for the desired task. Our model is trained in a supervised manner using the renowned cross-entropy [75] loss. In our case, we termed this segmentation loss (L_{seg}), defined as

$$L_{seg} = \frac{-\sum_{h=1}^{H}\sum_{w=1}^{W} y_S^{(h,w)} \cdot \log\left(h_\theta\left(x_S^{(h,w)}\right)\right)}{H\times W}, \quad (1)$$

where h_θ is the U-Net model with weights θ and $x_S \in \mathbf{X}_S$

3.1.2 SegDesic Module for Geographic Integration

The SegDesic module depicted in fig. 1, is composed of six fully connected linear layers, each followed by batch normalization and ReLU activation, with the exception of the last layer. The last encoder layer of the U-Net architecture underwent max pooling to extract features, which were then passed as an input to the SegDesic module. The final output ($\widehat{\mathbf{C}}$) is a vector representing the geo-coordinates of the image patch, which is then normalized and cosine dissimilarity calculated between this final output and the normalized positional encoding ($\mathbf{C}$) of metadata (C_{lon}, C_{lat}) of the corresponding image is calculated, termed here as domain loss.

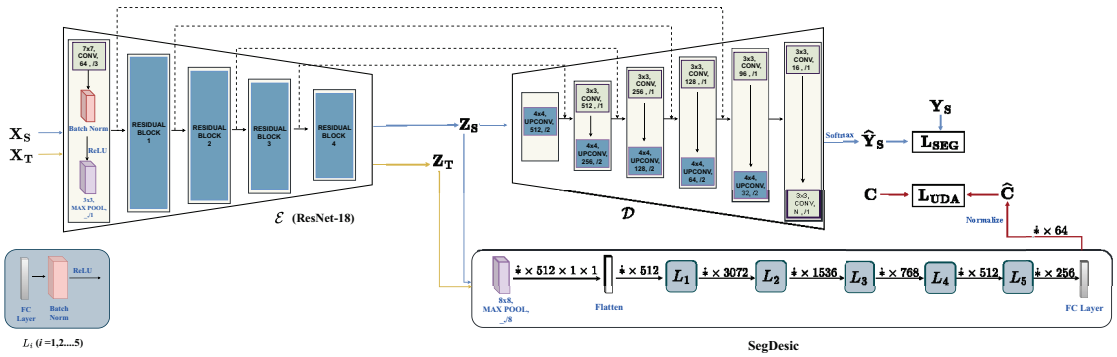

Figure 1. Our model contains ResNet-18 [19] asencoder ($\mathcal{E}$), a decoder ($\mathcal{D}$) and the SegDesic module. The encoder and decoder together form a U-Net [43]. The structure of the SegDesic module with its input and output sizes is shown on the lower block in the figure. In this diagram, $\dashrightarrow$ indicates its skip connections (with feature concatenation).*CONV* represents a 2D convolution layer, *UPCONV* signifies a 2D transposed convolution layer, $* \times *$ refers to the convolution kernel size, $*, /*'$ represents the number of kernels and convolution stride, $\dot{*}$ represents batch size, and N represent number of class.

3.2. Data Representation

3.2.1 Encoding Geographic Coordinates

Considering the multi-scale periodic representation of grid cells in mammals and the established connection between spatial patterns and navigation, as explored by [8] and [12], along with the demonstration of grid-like spatial patterns in trained networks, we propose a novel transformation to generate vectors that serve as supervision for $\widehat{\mathbf{C}}$. Our approach highlights the importance of grid cells for navigation and can be summarized like this:

1. Coordinate centering in the given EPSG:2154 reference system (for FLAIR #1 [1]). This ensures that the median values of the coordinates are $(0, 0)$.

$$\begin{aligned} C'_{\text{lon}} &= C_{\text{lon}} - 489353.59 \text{ m} \\ C'_{\text{lat}} &= C_{\text{lat}} - 6587552.20 \text{ m} \end{aligned} \tag{2}$$

2. Transforming coordinates to the EPSG:4326 coordinate system, from EPSG:2154, a local coordinate system. This transformation aims to expand the algorithm's functionality and improve its ability to recognize geographical features in a standardized configuration. The transformation process is crucial for achieving optimal results in geographic analysis.

$$\lambda, \phi = \text{f}_{2154 \to 4326}(C'_{lon}, C'_{lat}) \tag{3}$$

3. Multi-scale approach adequately captures 2D positions, the positional encoding of coordinates is approached using a multi-scale way. This is necessary because a single scale representation alone is insufficient for accurately representing 2D positions due to the periodic nature of sine and cosine functions [37].

$$\text{GRID}(\lambda, \phi) = \bigcup_{s=0}^{S-1} \left[\sin(\frac{\lambda}{\alpha_s}), \cos(\frac{\lambda}{\alpha_s}), sin(\frac{\phi}{\alpha_s}), \cos(\frac{\phi}{\alpha_s}) \right] \tag{4}$$

where, $\alpha_s = \lambda_{min} \cdot g^{s/(S-1)}$ is a scaling factor that induces increasingly high frequencies through scales s from 0 to $S-1$, $\bigcup$ indicates concatenation and λ_{min}, λ_{max} are the minimum and maximum grid scale with $g = \frac{\lambda_{max}}{\lambda_{min}}$.

4. Projected this encoded vector onto a unit sphere. By mapping the encoded vector onto the unit sphere, we preserve the inherent spherical nature of Earth's observations (EO). When utilized as a ground truth, it allows the model to effectively capture and represent EO in their original spatial context.

$$\mathbf{C} = \frac{\text{GRID}(\lambda, \phi)}{\|\text{GRID}(\lambda, \phi)\|_1} \tag{5}$$

5. Finally, we compute domain loss as

$$L_{\text{UDA}} = 1 - \frac{\mathbf{C} \cdot \widehat{\mathbf{C}}}{\|\mathbf{C}\|_2 \cdot \|\widehat{\mathbf{C}}\|_2} \tag{6}$$

3.3. Domain Adaptation Strategy

3.3.1 Domain Alignment Techniques

A domain denoted by $\mathbf{D}$ encompasses both the data ($\mathbf{X}$) and its corresponding distribution ($\mathbf{P}(\mathbf{x})$). Thus, the domain can be represented as $\mathbf{D} = \{\mathbf{X}, \mathbf{P}(\mathbf{x})\}$, where $\mathbf{x}$ is an element of $\mathbf{X}$. Conversely, a task ($\mathbf{T}$) characterized by the label space ($\mathbf{Y}$) and prediction function ($\mathbf{f}(\mathbf{x})$) can be interpreted as the posterior probability $\mathbf{p}(\mathbf{y}|\mathbf{x})$. Hence, the task can be expressed as $\mathbf{T} = \{\mathbf{Y}, \mathbf{P}(\mathbf{y}|\mathbf{x})\}$, where $\mathbf{y} \in \mathbf{Y}$.
In domain adaptation (DA), it is typically assumed that the

source domain ($\mathbf{D_S}$) and target domain ($\mathbf{D_T}$) share the same task ($\mathbf{T_S} = \mathbf{T_T}$) but exhibit different distributions ($\mathbf{P}(\mathbf{x_s}) \neq \mathbf{P}(\mathbf{x_t})$), where $\mathbf{x_s} \in \mathbf{X_S}$ and $\mathbf{x_t} \in \mathbf{X_T}$. This indicates variations in the statistical properties between the data in the two domains, such as feature distributions or data generation processes. The objective of DA is to train a model that effectively transfers knowledge from the source domain to the target domain, despite these distributional differences, to enhance performance on the target domain task. Supervised and semi-supervised DA require labeled data in the target domain, whereas unsupervised DA operates under the assumption that there is no knowledge of the label space in the target domain ($\mathbf{Y_T}$).

3.3.2 Utilization of Geographic Features for Adaptation

We adopted a similar architecture setup used by [38] in our work, in which the input data $\mathbf{X}_S$ and $\mathbf{X}_T$ go into the encoder of U-Net processes denoted as $\mathcal{E}$ to produce feature maps $\mathbf{Z}_S \in \mathbb{R}^{H' \times W' \times C}$ and $\mathbf{Z}_T \in \mathbb{R}^{H' \times W' \times C}$, where H', W', and C represent the height, width, and number of channels of the feature maps, respectively. Subsequently, decoder $\mathcal{D}$ of U-Net is fed with only $\mathbf{Z}_S$ to generate segmentation map $\hat{\mathbf{Y}}_S$. In addition, because the geo-coordinates are utilized for domain adaptation, both $\mathbf{Z}_S$ and $\mathbf{Z}_T$ are fed as inputs to the SegDesic module, which predicts a vector $\hat{\mathbf{C}} \in \mathbb{R}^D$ containing the localization information of the image and is then normalized (in the same way as depicted in eq. (5)). The loss obtained from this module was used to fine-tune the encoder features. The final loss of SegDesicNet takes the following form:

$$L = L_{seg} + \alpha \times (L^S_{UDA} + L^T_{UDA}) \tag{7}$$

where $L^S{}_{UDA}$ and L^T_{UDA} are the loss obtained from the SegDesic module for source and target data. To ensure that the contribution of the loss from different domains remains proportional to the segmentation loss, we performed scaling on its value using a hyperparameter α.

4. Experiment

4.1. Datasets

FLAIR #1: It consists of 50 spatial domains, each depicting the diverse landscapes and climates of metropolitan France, corresponding to specific French departments [1]. Each image patch measures 512 × 512 pixels with a ground sampling distance (GSD) of 0.2 meters, and each domain contains between 1,725 and 1,800 patches. Based on the implementation in [38], we used 12 major semantic classes and subsets for the source and target domain data in our experiments. In addition, we evaluated the performance of our model using different custom splits for the source and target domains.

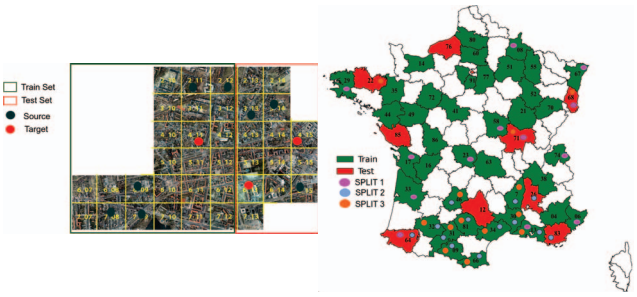

Figure 2. Shown are the training and test set splits of the ISPRS Potsdam dataset (on the left) and the FLAIR #1 dataset (on the right), with custom splits of the source and target domains indicated by colored dots.

ISPRS Potsdam: This dataset [48] consists of 38 labeled tiles, each covering an area of 6000 × 6000 pixels, with every pixel categorized into one of the six semantic classes. The dataset achieves a GSD of approximately 5 cm per pixel. Following standard practice mentioned in [11], the tiles were partitioned into non-overlapping images of 512 × 512 pixels.

The details of custom data splits are listed in Tab. 1 and the domain gap of in the data is presented in fig. 2. SPLIT1 is maintained as used in [38].

5. Result and Ablations

5.1. Baselines

Unfortunately, we were unable to compare the performance of the proposed method with any UDA method based on geographical considerations because the code for the previous SOTA method for geographic SS [38] is unavailable. Nevertheless, we compared our approach to the results reported by four state-of-the-art UDA methods for land cover mapping: AdaptSegNet [53], CLAN [32], IAST [39], and ADVENT [56]. Additionally, we evaluated our results for SPLIT1 (Tab. 1) against SOTA GeoMultiTaskNet [38] on the FLAIR #1 dataset.

5.2. Experimental Setup

For our implementation, we applied the settings common to most works [14, 28, 73, 81] with batch size 16, 0.001 initial LR, and poly learning decay with Adam during training. In addition, during training, the images are randomly cropped from 512×512 to 256×256, while inference is performed over all patches that are merged back to their original shape. The training iteration was set to 200 iterations unless otherwise specified. We used early stopping with a patience of 30 epochs. Additionally, we set the constant value of random seed as 18 for different modules/packages used during implementation. A single NVIDIA GeForce

Data Set	SPLIT	Source Domain	Target Domain	# Train Image	# Test Image
FLAIR #1	SPLIT1 [38]	D06, D08, D13, D17, D23, D29, D33, D58, D67, D74	D64, D68, D71	16050	5350
	SPLIT2	D07, D09, D013, D30, D31, D32, D34, D46, D66, D81	D26, D64, D83	15325	5625
	SPLIT3	D07, D09, D13, D30, D31, D32, D34, D46, D66, D81	D22, D68, D71	16272	5300
ISPRS Potsdam	SPLIT4	2_11, 2_12, 3_13, 3_14, 5_11, 6_15, 6_9, 7_8, 7_9	4_11, 6_13, 4_15	1296	432

Table 1. Custom Source and Target Domain Splits

Method	Building	Pervious	Impervious	Bare Soil	Water	Coniferous	Deciduous	Brushwood	Vine	Grassland	Crop	Plowed Land	mIoU
CLAN [32]	6.24	13.66	17.09	1.50	12.99	1.29	27.22	3.36	30.69	27.34	7.69	18.42	13.96
AdaptSegNet [53]	39.98	20.75	40.23	20.36	15.25	4.93	35.37	10.99	34.51	**42.69**	11.06	23.47	24.97
ADVENT [56]	35.79	24.38	48.82	6.85	31.98	0.00	51.65	11.79	33.33	25.76	11.46	24.29	25.51
IAST [39]	55.67	36.43	53.71	26.95	53.33	0.00	50.67	11.56	43.24	26.28	26.31	44.27	35.70
DAFormer [56]	67.09	45.56	61.99	55.35	65.12	8.91	54.39	**20.31**	64.39	38.79	23.74	41.83	45.62
UDA_for_RS [27]	66.30	**48.05**	62.36	**59.28**	61.24	9.22	60.02	16.52	57.74	40.12	30.32	54.17	47.11
GeoMultiTaskNet [38]	67.53	40.86	63.89	55.31	67.02	**13.85**	60.97	14.08	53.09	40.33	35.02	**54.79**	47.23
Ours	**67.65**	45.18	**64.13**	55.92	**71.37**	11.62	**62.65**	16.65	**74.50**	40.90	**37.83**	53.06	**50.12**

(a) SPLIT1

Method	Building	Pervious	Impervious	Bare Soil	Water	Coniferous	Deciduous	Brushwood	Vine	Grassland	Crop	Plowed Land	mIoU
CLAN [32]	24.27	21.27	41.19	8.89	26.57	4.22	28.26	24.70	54.84	31.00	18.80	1.28	23.77
AdaptSegNet [53]	60.07	24.06	54.22	21.29	24.31	11.90	39.02	26.31	47.25	31.56	19.95	10.56	30.88
ADVENT [56]	56.21	29.82	55.47	27.52	44.73	15.83	41.38	22.77	55.18	41.47	12.52	30.26	36.10
IAST [39]	68.84	**42.05**	61.19	**61.86**	**79.38**	0.00	46.27	**35.69**	69.40	35.38	34.06	**36.71**	47.57
ours	**69.60**	38.63	**65.31**	46.16	73.54	**25.52**	**46.89**	35.19	**74.00**	**44.87**	**34.65**	26.67	48.42

(b) SPLIT2

Method	Building	Pervious	Impervious	Bare Soil	Water	Coniferous	Deciduous	Brushwood	Vine	Grassland	Crop	Plowed Land	mIoU
CLAN [32]	28.86	18.38	37.78	5.31	17.92	7.78	45.82	10.69	66.38	40.61	14.67	7.81	25.17
ADVENT [56]	49.15	30.39	50.13	18.23	43.70	15.75	59.95	**15.37**	49.89	**53.45**	9.18	22.87	34.84
AdaptSegNet [53]	57.95	27.68	50.17	11.78	35.25	17.96	63.73	14.28	55.64	40.86	28.26	17.81	35.11
IAST [39]	62.31	**52.13**	**64.67**	15.04	**75.28**	0.00	61.19	12.70	69.00	30.85	**39.85**	49.28	44.36
ours	**63.33**	46.61	60.32	**18.52**	51.09	**31.25**	**66.32**	14.12	**74.58**	47.30	34.25	**52.94**	**46.72**

(c) SPLIT3

Table 2. Class-wise comparison of IoU(%) for for different methods across various data domains of FLAIR #1 dataset.

Architecture	#Parameters(M)
AdaptSegNet [53]	99
ADVENT [56]	99
DAFormer [21]	85
UDA_for_RS [27]	85
CLAN [32]	44
IAST [39]	45
GeoMultiTaskNet [38]	33
Ours	~ 24

Table 3. Comparing the parameter count of the baseline models used in SPLIT1 to that of ours

Method	Impervious	Building	Low Vege	Tree	Car	mIoU
IAST [39]	68.42	75.86	64.77	64.46	43.72	63.46
CLAN [32]	64.67	68.71	58.72	47.28	42.63	56.40
ADVENT [56]	72.18	78.33	69.65	67.65	68.69	71.30
AdaptSegnet [53]	73.40	78.89	68.14	67.69	70.65	71.75
Ours	**74.88**	**81.17**	**70.49**	**69.47**	**72.26**	**73.65**

Table 4. Comparison of IoU (%) across classes for various methods on custom data splits SPLIT4 of the ISPRS Potsdam dataset.

RTX 4090 GPU (24 GB) was used in the training phase. Similar to the works discussed above, we chose the standard mean intersection over union (mIoU) as our evaluation metric.

In Tab. 2, we present a comparison of class-wise Intersection over Union (IoU) between the baseline models and our model. All the metrics reported in Tab. 2 and Tab. 4 are either obtained directly from the previous published sources or are obtained by training their official implementation along using their pretrained weight. To evaluate the performance of our model across different domain gaps, we extended our experiment by creating 2 additional custom data splits of FLAIR#1 dataset and 1 from ISPRS Potsdam dataset (mentioned in Tab. 1) and obtained the results listed in Tab. 2b–Tab. 2c and Tab. 4. Experiments were conducted to investigate the behavior of SegDesicNet across different values (λ_{min}, λ_{max}) and numbers of grid scales (S) alongside the domain contribution factor α. The results for SPLIT1 are presented in Tab. 5. We used this specific value for all of our experiments. Overall, the results on the target domain test set of FLAIR #1 and ISPRS Potsdam demonstrate that the identified lightweight networks possess the UDA-oriented capability to transfer knowledge from a source domain to a target domain. Our results demonstrate a considerable improvement in the mean IoU (mIoU) by approximately 6% compared with state-of-the-art performance (refer to Tab. 2a). Similar findings were obtained for the custom domain split of the ISPRS Potsdam dataset. Our pro-

posed algorithm outperformed the other baseline methods. It is also worth noting that ranking of the different methods varies between the various domain splits, and some models failed to recognize certain classes entirely, our methods consistently outperformed them and maintained a very consistent final value. Notably, unlike the approach described in [38], we did not employ any specific weighting mechanism or any data augmentation and designed a very lightweight method with approximately 27% fewer model parameters, despite this, our proposed method achieved robust performance across nearly all underrepresented classes (including *coniferous, deciduous, brushwood, vine, grassland, crop, and plowed land*) among every data splits.

λ_{min}	λ_{max}	S	α	$mIoU$
–	–	–	0	45.6
0.01	0.00001	8	0.5	47.9
0.01	0.000001	8	1.0	45.52
0.01	**0.00001**	**16**	**0.5**	**50.12**
0.1	0.00001	16	0.5	49.46
0.01	0.000001	16	1	45.52
0.01	0.00001	32	0.5	48.75
0.1	0.00001	32	0.5	49.07
0.01	0.000001	32	1	45.48
0.01	0.00001	64	0.5	48.60
0.01	0.000001	64	1	46.66
0.1	0.0001	64	1	48.6

Table 5. Observations depicting the performance of SegDesicNet across various hyperparameters for SPLIT1

Qualitative Evaluation : In fig. 3, examples of land cover mapping results on the target domain of FLAIR #1 SPLIT1. In this figure, we observe that while GeoMultiTaskNet comes close to our results for classes such as *plowed land* and *brushwood*, our proposed method outperforms it considerably. Our model successfully captured distinctive features that might vary in style (visual features), as seen in the second image. This indicates our model's ability to handle style differences in mobility infrastructure, which can occur owing to large domain differences, a key challenge in domain adaptation. However, our model has shown some indistinctness in differentiating *pervious surfaces* from *grasslands*. Despite this minor shortcoming, it still performs exceptionally well across other categories. Notably, DAFormer and UDA_for_RS, despite their strong performance in the impervious surface class, exhibited misclassification issues. This highlights the robustness of our model in maintaining accuracy across a broader range of land cover types.

6. Limitation and Future Works

Although the proposed method demonstrated encouraging performance, it has some limitations. First, there are several classes, particularly the *pervious surface* class, which consistently underperformed throughout the experiment. As shown in fig. 3, the model performed well when the *pervious surface* class was surrounded by a *deciduous* class. However, its performance degrades when *pervious surface* class is surrounded by *grassland*. All the baselines including our methods has shown low accuraccy over the pixels affected by shadows. This observation suggests that there is room to improve the proposed model by incorporating a class-conditional style in which the probability of a class can be further analyzed based on its neighboring classes. In addition, during training, we chopped the input images to a size of 256×256 in an attempt to decrease the network parameter, but we can see in the fig. 3 top right that our impervious surface is misjudged due to non overlapping splitting. Hence, better conditional cropping should be performed to address it in better terms. Second, to narrow the domain gap between the neural network's style of understanding and the artificial neural network we used, we implemented GRID positional encoding. However, this approach may not be optimal. There is a pressing need to explore more efficient methods for modeling geocoordinates that account for the Earth's curved surface to make our approach more human-centric. In addition, one constraint of lightweight modeling in the context of remote sensing data is the necessity for real-time execution across various applications. we addressed this by manipulating our network layers, there may be numerous possibilities to address this more effectively. Also more features can be added to this framework, which includes enabling the model to utilize on cross-sensor data, such as by predicting the height of a pixel along with its classes. Thus, we can enhance the practicality and applicability of the proposed method in real-world scenarios.

7. Conclusion

Our work has explored unsupervised domain adaptation methods using the geographical coordinates of image patches. Our findings demonstrate that incorporating geo-coordinates into neural network training enhances the model's ability to adapt to target domains without ground truth label, leading to improved performance. This innovative method has pivotal implications for various diverse environmental applications, as it enables accurate and efficient adaptation due to incorporation geo-coordinates. By mimicking human perception of the physical environment, our approach advances the understanding of how neural networks perceive their surroundings. When considering latitude and longitude values on a spherical surface, cosine dissimilarity is a more effective approach compared to Euclidean distance or mean squared error (MSE) loss. This is because cosine dissimilarity takes into account the Earth's curved surface, unlike other distance metrics that assume a flat Euclidean space, cosine dissimilarity captures the spher-

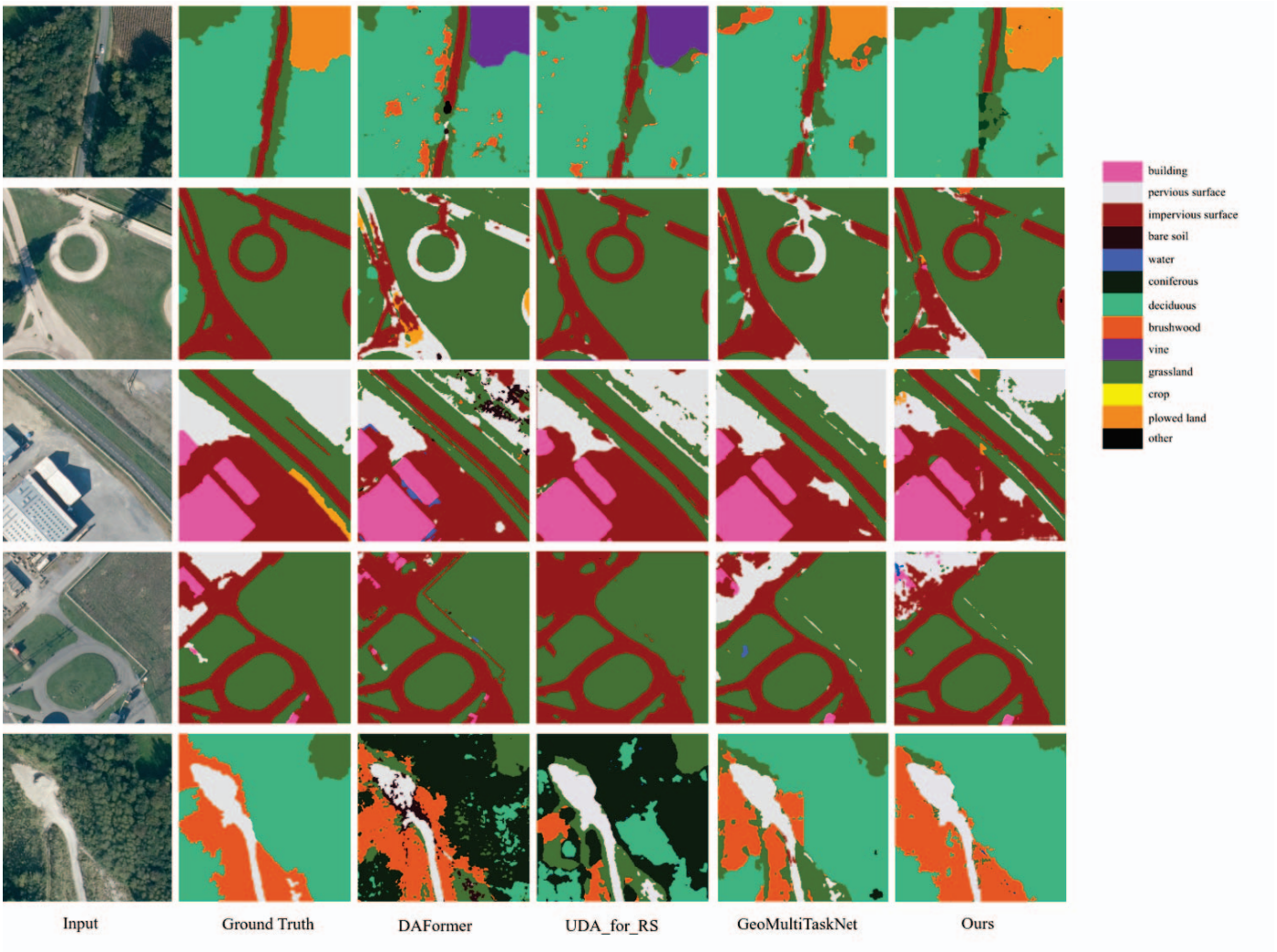

Figure 3. Comparison of segmentation map prediction on the target domain images of SPLIT1 by baseline models to our best model.

ical nature of the Earth, resulting in a more comprehensive evaluation. Although our research showed promising results, it is crucial to acknowledge some limitations. Future research could investigate alternative methods for incorporating geographical information into neural network architectures for unsupervised domain adaptation. Furthermore, assessing the scalability and computational efficiency of the proposed method across various domains and datasets can enhance its practical application. Our study demonstrates the potential of utilizing geographical coordinates to enhance unsupervised domain adaptation in neural networks. By learning from the human perception of their physical surroundings, our approach presents a promising avenue for addressing the challenges in adapting models to new environments. By embracing interdisciplinary approaches and incorporating insights from fields such as geography and cognitive science, we can continue to advance machine learning and deepen our understanding of how neural networks interact with the physical world.

Acknowledgements

This research received funding from the PERSEUS project, a European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 101034240. This paper is supported by the MoST (MobilitetsLab Stor-Trondheim) project (https://www.mobilitetslabstortrondheim.no/en/)

References

[1] Flair n. 1: Semantic segmentation and domain adaptation. https://codalab.lisn.upsaclay.fr/competitions/8769, 2022. [Accessed: 31 January 2024]. 2, 4, 5

[2] Alison Abbott and Ewen Callaway. Nobel prize for decoding brain's sense of place. *Nature News*, 514(7521):153, 2014. 2

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

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

[5] Nikita Araslanov and Stefan Roth. Self-supervised augmentation consistency for adapting semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15384–15394, 2021. 2

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

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

[8] Andrea Banino, Caswell Barry, Benigno Uria, Charles Blundell, Timothy Lillicrap, Piotr Mirowski, Alexander Pritzel, Martin J Chadwick, Thomas Degris, Joseph Modayil, et al. Vector-based navigation using grid-like representations in artificial agents. *Nature*, 557(7705):429–433, 2018. 4

[9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 1, 2

[10] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022. 2

[11] MMSegmentation Contributors. Dataset preparation for mmsegmentation: Isprs potsdam, 2024. Accessed: 2024-07-15. 5

[12] Christopher J Cueva and Xue-Xin Wei. Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. *arXiv preprint arXiv:1803.07770*, 2018. 4

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 3

[14] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Semantic correlation promoted shape-variant context for segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8885–8894, 2019. 5

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[16] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019.

[17] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 2

[18] Wolfgang Gross, Devis Tuia, Uwe Soergel, and Wolfgang Middelmann. Nonlinear feature normalization for hyperspectral domain adaptation and mitigation of nonlinear effects. *IEEE Transactions on Geoscience and Remote Sensing*, 57(8):5975–5990, 2019. 3

[19] 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. 3, 4

[20] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9924–9935, 2022. 2

[21] 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. 6

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

[23] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 603–612, 2019. 2

[24] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4893–4902, 2019. 2

[25] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. *arXiv preprint arXiv:1805.10180*, 2018. 2

[26] Weitao Li, Hui Gao, Yi Su, and Biffon Manyura Momanyi. Unsupervised domain adaptation for remote sensing semantic segmentation with transformer. *Remote Sensing*, 14(19), 2022. 2

[27] 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. 6

[28] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9167–9176, 2019. 5

[29] Di Lin, Yuanfeng Ji, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Multi-scale context intertwining for semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 603–619, 2018. 2

[30] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1, 2

[31] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. *Advances in neural information processing systems*, 29, 2016. 2

[32] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2507–2516, 2019. 5, 6

[33] Li Ma, Melba M. Crawford, Lei Zhu, and Yong Liu. Centroid and covariance alignment-based domain adaptation for unsupervised classification of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(4):2305–2323, 2019. 3

[34] Oscar Mañas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodríguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9414–9423, October 2021. 3

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

[36] Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 3

[37] 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. 3, 4

[38] V. Marsocci, N. Gonthier, A. Garioud, S. Scardapane, and C. Mallet. Geomultitasknet: remote sensing unsupervised domain adaptation using geographical coordinates. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2075–2085, Los Alamitos, CA, USA, jun 2023. IEEE Computer Society. 2, 3, 5, 6, 7

[39] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. *CoRR*, abs/2008.12197, 2020. 5, 6

[40] Risto Miikkulainen, Jason Liang, Elliot Meyerson, Aditya Rawal, Dan Fink, Olivier Francon, Bala Raju, Hormoz Shahrzad, Arshak Navruzyan, Nigel Duffy, et al. Evolving deep neural networks. In *Artificial intelligence in the age of neural networks and brain computing*, pages 269–287. Elsevier, 2024. 1

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

[42] Jiangtao Peng, Yi Huang, Weiwei Sun, Na Chen, Yujie Ning, and Qian Du. Domain adaptation in remote sensing image classification: A survey. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:9842–9859, 2022. 1, 2

[43] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*, volume 18, pages 234–241, Munich, Germany, October 5-9 2015. Springer International Publishing. 3, 4

[44] M. Rußwurm, K. Klemmer, E. Rolf, R. Zbinden, and D. Tuia. Geographic Location Encoding with Spherical Harmonics and Sinusoidal Representation Networks. *arXiv preprint arXiv:2310.06743*, Oct 10 2023. 3

[45] S. Salcedo-Sanz, P. Ghamisi, M. Piles, M. Werner, L. Cuadra, A. Moreno-Martínez, E. Izquierdo-Verdiguier, J. Muñoz-Marí, Amirhosein Mosavi, and G. Camps-Valls. Machine learning information fusion in earth observation: A comprehensive review of methods, applications and data sources. *Information Fusion*, 63:256–272, 2020. 2

[46] Tawfiq Salem, Scott Workman, and Nathan Jacobs. Learning a dynamic map of visual appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[47] 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

[48] Ahram Song and Yongil Kim. Semantic segmentation of remote-sensing imagery using heterogeneous big data: International society for photogrammetry and remote sensing potsdam and cityscape datasets. *ISPRS International Journal of Geo-Information*, 9(10):601, 2020. 2, 5

[49] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 443–450. Springer, 2016. 2

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

[51] Kevin Tang, Manohar Paluri, Li Fei-Fei, Rob Fergus, and Lubomir Bourdev. Improving image classification with location context. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. 3

[52] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7472–7481, 2018. 2

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

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[55] V. Vivanco Cepeda, G.K. Nayak, and M. Shah. GeoCLIP: Clip-Inspired Alignment between Locations and Images for Effective Worldwide Geo-localization. *arXiv e-prints*, pages arXiv–2309, 2023. 3

[56] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2517–2526, 2019. 2, 5, 6

[57] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Dada: Depth-aware domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7364–7373, 2019. 2

[58] Haoyu Wang, Xiuyuan Zhang, Shihong Du, Lubin Bai, and Bo Liu. Mapping annual urban evolution process (2001–2018) at 250 m: A normalized multi-objective deep learning regression. *Remote Sensing of Environment*, 278:113088, 2022. 1

[59] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 1, 2

[60] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. Understanding convolution for semantic segmentation. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 1451–1460. Ieee, 2018. 1

[61] Qin Wang, Dengxin Dai, Lukas Hoyer, Luc Van Gool, and Olga Fink. Domain adaptive semantic segmentation with self-supervised depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8515–8525, 2021. 2

[62] Qihao Weng. Remote sensing of impervious surfaces in the urban areas: Requirements, methods, and trends. *Remote Sensing of Environment*, 117:34–49, 2012. 1

[63] Scott Workman and Hunter Blanton. Augmenting depth estimation with geospatial context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4562–4571, October 2021. 3

[64] Binhui Xie, Shuang Li, Mingjia Li, Chi Harold Liu, Gao Huang, and Guoren Wang. Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[65] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 1, 2

[66] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020. 2

[67] Qingsong Xu, Yilei Shi, Xin Yuan, and Xiao Xiang Zhu. Universal domain adaptation for remote sensing image scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023. 1

[68] Junjie Ye, Changhong Fu, Guangze Zheng, Danda Pani Paudel, and Guang Chen. Unsupervised domain adaptation for nighttime aerial tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8896–8905, June 2022. 3

[69] Bo Yu, Lu Yang, and Fang Chen. Semantic segmentation for high spatial resolution remote sensing images based on convolution neural network and pyramid pooling module. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(9):3252–3261, 2018. 1

[70] Changqian Yu, Jingbo Wang, Changxin Gao, Gang Yu, Chunhua Shen, and Nong Sang. Context prior for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12416–12425, 2020. 2

[71] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 173–190. Springer, 2020. 2

[72] Yuhui Yuan, Lang Huang, Jianyuan Guo, Chao Zhang, Xilin Chen, and Jingdong Wang. Ocnet: Object context for semantic segmentation. *International Journal of Computer Vision*, 129(8):2375–2398, 2021. 2

[73] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7151–7160, 2018. 2, 5

[74] Jing Zhang, Zewei Ding, Wanqing Li, and Philip Ogunbona. Importance weighted adversarial nets for partial domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8156–8164, 2018. 2

[75] Zhengming Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. 3

[76] D. Zhao, S. Wang, Q. Zang, D. Quan, X. Ye, R. Yang, and L. Jiao. Learning pseudo-relations for cross-domain semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19191–19203, 2023. 2

[77] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 2

[78] Zilong Zhong, Zhong Qiu Lin, Rene Bidart, Xiaodan Hu, Ibrahim Ben Daya, Zhifeng Li, Wei-Shi Zheng, Jonathan Li, and Alexander Wong. Squeeze-and-attention networks for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13065–13074, 2020. 2

[79] 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 prototype memory. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–18, 2023. 3

[80] 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. 2

[81] Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 593–602, 2019. 5

[82] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018. 2