Segmenting across places: The need for fair transfer learning with satellite imagery

Miao Zhang Harvineet Singh Lazarus Chok Rumi Chunara
New York University
{miaozhng, hs3673, lcc9673, rumi.chunara}@nyu.edu

Abstract

The increasing availability of high-resolution satellite imagery has enabled the use of machine learning to support land-cover measurement and inform policy-making. However, labelling satellite images is expensive and is available for only some locations. This prompts the use of transfer learning to adapt models from data-rich locations to others. Given the potential for high-impact applications of satellite imagery across geographies, a systematic assessment of transfer learning implications is warranted. In this work, we consider the task of land-cover segmentation and study the fairness implications of transferring models across locations. We leverage a large satellite image segmentation benchmark with 5987 images from 18 districts (9 urban and 9 rural). Via fairness metrics we quantify disparities in model performance along two axes – across urban-rural locations and across land-cover classes. Findings show that state-of-the-art models have better overall accuracy in rural areas compared to urban areas, through unsupervised domain adaptation methods transfer learning better to urban versus rural areas and enlarge fairness gaps. In analysis of reasons for these findings, we show that raw satellite images are overall more dissimilar between source and target districts for rural than for urban locations. This work highlights the need to conduct fairness analysis for satellite imagery segmentation models and motivates the development of methods for fair transfer learning in order not to introduce disparities between places, particularly urban and rural locations.

1. Introduction

Satellite imagery is becoming readily available with around 1030 active satellites that are dedicated to earth observation [36]. Out of the different spectra of imagery available from such satellites, visible spectrum imagery is particularly relevant for many applications based on the extremely high resolution and according ability to resolve specific objects of interest [5]. Consequently, satellite images combined with semantic segmentation, the task of clustering parts of an image together which belong to the same object class, can be used to detect objects ranging from natural features (water bodies, forests) to human land-use types (buildings, roads). The extracted information is being applied in a wide range of settings including urban planning [34], modelling disease spread [1], aiding disaster relief efforts [16, 57], and detecting and mapping environmental phenomena [24, 55]. However, because segmentation models employ supervised learning, availability of ground truth data is a major bottleneck for their training. Annotation for the segmentation task is particularly labor intensive as it requires fine-grained labels at the level of pixels which results in incomplete or noisy ground truth data [45]. In such situations, generalizing existing models to non-annotated data by *transfer learning* is a widely applied solution [43, 48].

Transfer learning uses knowledge learnt from the same or related tasks to improve learning on the task at hand (see Pan and Yang [39] for a survey). We will focus on a type of transfer learning setting called *domain adaptation*, where we have a single task but the train and test domains may differ. The key challenge here is the discrepancy in data distributions between domains. In the case of satellite imagery, the discrepancies commonly result from transferring models to new geographies where the landscapes are dissimilar to where the model was trained. For example, Islam [22] finds that a well-trained seagrass detection model from satellite images fails when tested at other locations with different seagrass density. To mitigate the degrading effects of domain discrepancies on segmentation accuracy, previous work has re-designed network architectures [28], loss functions [18, 50], and batch normalization methods [38] to improve model generalization. Other approaches include using labels at a coarser granularity for the target domain (e.g. image-level labels) as weak supervision [21] and learning latent representations shared between source and target do-

mains to help in adaptation [26, 29].

Simultaneously, while machine learning approaches have been used to improve prediction in a variety of tasks, recent studies have highlighted concerns towards model fairness, exhibited by performance disparities across sensitive groups based on geography, demographics, and economic indicators [31, 58]. A push for model fairness aligns with the ideal of equity defined by World Health Organization as "Equity is the absence of unfair, avoidable or remediable differences among groups of people, whether those groups are defined socially, economically, demographically, or geographically or by other dimensions of inequality (e.g. sex, gender, ethnicity, disability, or sexual orientation)." Real-world examples have demonstrated the harmful effects of unfair machine learning models, such as facial recognition software that performs worse on darker women [4] and advertisement systems that deliver economic opportunity-related ads less often to women than men [25]. Indeed, discriminatory issues persist even in state-of-the-art learning methods [32]. Expanding types of data used in machine learning tasks, such as satellite imagery, enables increased use in a wide range of daily-life applications and ever-increasing social impacts. Accordingly, broader aspects and viewpoints of performance, such as fairness, need to be ascertained in multiple machine learning subareas.

In this work we study **the fairness impacts of transfer learning with satellite imagery**. To accomplish this goal, we test multiple semantic segmentation models across different geographies. We then assess if such models made fair predictions on both the source and the target data. We focus on differences between urban and rural areas (i.e. urban/rural categorization is the sensitive attribute) due to persistent and striking disparities between urban and rural areas, especially for poor populations [2, 3]. The unfairness criterion in this work is based on differences in error rates across protected groups where the error rate is computed using Intersection-over-Union (IoU), a standard segmentation metric. We also examine model performance disparity across different land-cover classes. Results show that existing domain adaptation methods do not maintain fairness properties on transfer, either across protected groups or feature classes. This work serves as a valuable demonstration of fairness being an critical issue in transfer learning using a large freely-available satellite imagery dataset.

Important takeaways are as follows.

- Studied models have better overall accuracy (via mean IoU over the 7 classes) on rural districts as compared to urban districts.
- For common unsupervised domain adaptation methods, transfer accuracy is improved, but at the cost of fairness; the performance gap between rural and urban group is enlarged indicating the need to design new methods that transfer well for both the groups.
- Investigating reasons for the above findings, we find that images from rural districts differ more across locations than those of urban districts.

2. Related Work

Before discussing prior work on transfer learning for satellite images, we describe some of the alternative ways to address label scarcity and their shortcomings. Lastly, we summarize work in the nascent area of fair transfer learning.

Approaches to tackle annotation burden for satellite images Given the difficulty in labelling data for semantic segmentation of satellite images, Schmitt, *et al.* [45] developed weakly-supervised learning methods, where noisy, limited, or imprecise data sources are used to provide supervision signal. Previous work has leveraged the spatial context to develop unsupervised losses which, for example, penalize nearby pixels with different predicted labels [35, 50]. In another approach, Castillo-Navarro, *et al.* [6] proposed auxiliary losses based on self-supervised image reconstruction to improve the performance on the main task of image segmentation. To improve efficiency of label-use, Wang *et al.* [53] transferred classification models trained with image-level labels to image segmentation tasks and achieved high accuracy. While these approaches demonstrate successful combination of labeled and unlabeled images, they assume that the images are from the same domain (or distribution). However, the assumption of a consistent domain is not realistic for problems involving satellite images which are often from different geographies. Thus, such approaches are not straight-forwardly applicable in our setting.

Transfer learning for satellite images Transferability of satellite image segmentation models across different geographic locations has been studied in Ghorbanzadeh *et al.* [14]. Using train and test sets across 3 different geographies (Taiwan, China, and Japan), they show consistent decrease in evaluation scores upon transfer. Previous work has incorporated domain adaptation methods to deal with the challenge. For instance, Tran *et al.* [49] proposed a two-stage transfer learning structure which generated pseudo-ground truth segmentation labels for target data. Algorithms to improve the quality of such pseudo labels were studied in [33, 59]. Data augmentation is another strategy for domain adaptation. Ji *et al.* augments images to simulate perturbations due to atmospheric radiation and demonstrate improved generalization of CNN-based models [23]. These studies show the promise of adapting models to data from different locations. But, the transfer is only evaluated based on overall accuracy for the domain, such as using Intersection-over-Union (IoU) to measure the overlap between predicted segmentation maps and ground-truth

masks. Past work does not study fine-grained measures of model performance on transfer, like how is the performance for different subgroups in the domain (based on sensitive attributes or land-use types) impacted. The risk that discrepancy between domains in transfer learning may impact subgroups unfairly remains unexplored.

Fairness in transfer learning Following the work in fair machine learning literature [32], we will narrowly classify the study of performance differences between subgroups as *model fairness* analysis. Compared to fairness analysis within the same domain, little work has studied transfer of fair models across domains. The two objectives–improving transfer accuracy and maintaining fairness–can be at odds with each other [47, 56]. Schumann *et al.* [46] formalized the problem of fair transfer learning which sets the learning objective to improve accuracy as well as fairness in the target domain. Multiple approaches to fair transfer learning have been proposed [8, 27, 30, 37, 41, 42, 47] that make various assumptions on how the domains differ and what data is available. Even when labels are not available for the target domain, like in our setting, methods typically make the *covariate shift* assumption which says that the labeling rule remains the same between the domains and only the feature distribution changes. In this setting, Coston *et al.* [8] propose a method for fair transfer even in cases where sensitive attributes are absent from one of the domains. Other approaches do not require access to target domain data altogether and instead either make causal assumptions on the discrepancy [47] or hypothesize a set of target domains and optimize against them [10, 30]. Finally, Szabó *et al.* [19] conduct a fairness evaluation of segmentation methods assuming a single domain.

However, none of the existing works study fair transfer learning for semantic segmentation models. The task differs considerably from the above settings as the input data is high-dimensional, and the model output and loss function for segmentation are different. We take the first step in this direction by demonstrating the need for such methods via a thorough empirical study on a relevant application.

3. Dataset

We use the publicly available, high spatial resolution land-cover dataset called LoveDA [52] in this study. Compared to other popular satellite image datasets, such as Zurich Summer [50], DeepGlobe [9], and DSTL [20], the recently released LoveDA has more annotated images and includes images from diverse locations. The dataset consists of 5987 images of size 1024×1024 and spatial resolution 0.3m. The images are collected from 18 administrative districts from three cities in China, namely Nanjing, Changzhou, and Wuhan. Out of these, there are 9 urban districts and 9 rural districts, categorized based on their population density and level of economic development. We use

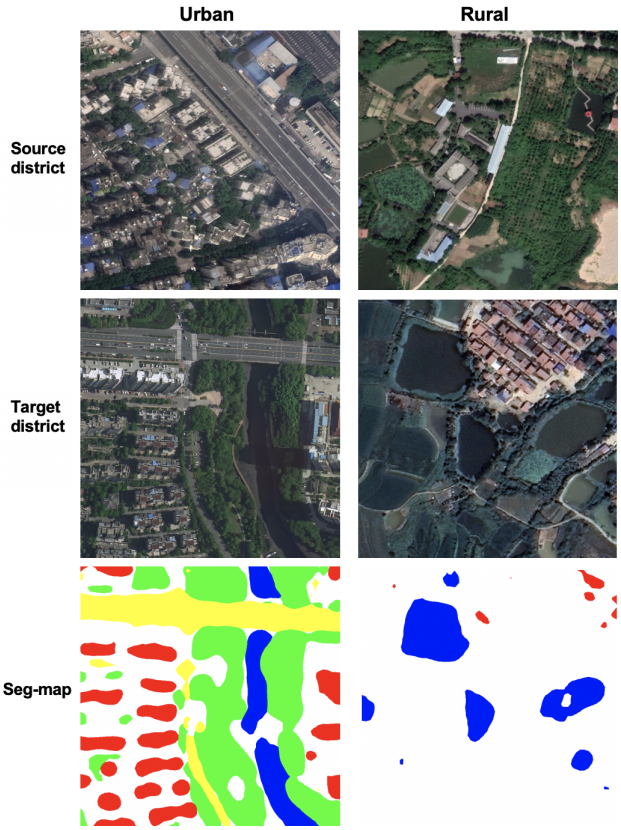

Figure 1. **Sample images from urban and rural scenes.** For each scene, one image from source domain districts, one from target domain districts, and the network's segmentation predictions (Seg-map) for the 7 land-cover classes on that target image are shown.

satellite images from the 12 districts for which ground-truth masks are available: Gulou, Qinhuai, Qixia, Yuhuatai, Jintan, and Jianghan (urban); Pukou, Lishui, Liuhe, Huangpi, Gaochun, and Jiangxia (rural). The remaining 6 districts are not availanble as they are held out for the benchmark challenge. The dataset contains segmentation masks, which are pixel-level labels, for 7 land-cover classes: Background, Building, Road, Water, Barren, Forest, and Agricultural. The Background class consists of any pixel not belonging to the other classes. Statistics of the dataset, given in Figure 3 in [52], show that the pixel counts across the 7 classes are imbalanced. Further, distribution of classes and of building scales differ between urban and rural scenes. Thus the rural and urban groups of images, which we use in our fairness analysis, have different characteristics.

4. Methods

We study three tasks, namely semantic segmentation within the same domain, across districts, and across rural-urban areas. Next, we describe the setup for each task.

4.1. Task A - Semantic segmentation

Our task is to train multi-class semantic segmentation models for detecting the 7 land-cover classes from a given image. Sample images and predicted segmentation maps are shown in Figure 1. Same as the models studied in the LoveDA study [52], we use two commonly used deep learning-based segmentation methods – U-Net [44] and DeepLabV3+ [7] network, both with pre-trained ResNet50 [17] as the backbone model for the encoder [13].

4.1.1 Training-testing details

Images from the 12 districts with labeled data are shuffled and split into training ($\approx$ 80%) and testing sets ($\approx$ 20%). Training set has 3148 images with a mix of 1377 urban and 1771 rural images, and the rest of the images comprise the testing set with a mix of 368 urban images and 473 rural images. Images are augmented during training by mirroring and rotation. Dimension of the input image to the network is 512×512×3 where 3 indicates the RGB bands. The output dimension of the network is 512×512×7, where 7 represents the probability of each pixel belonging to each land-cover class. We use cross-entropy (CE) loss, and stochastic gradient descent (SGD) as the optimizer with a momentum of 0.9 and a weight decay of 10^{-4}. The batch size is set to 16 and the total training iterations are 15000, during which the learning rate is decayed using a polynomial learning rate scheduler implemented in PyTorch [40].

4.1.2 Evaluation metrics

We test the models on either the whole test set or the urban and rural subsets in the test set separately, and evaluate model performance using the following metrics:

Accuracy metrics: Intersection-over-Union (IoU), also called Jaccard index, is used to measure segmentation accuracy which is a common method to evaluate the quality of image segmentation [11, 54]. IoU for a class is defined as the intersection of class-wise ground-truth masks and the predicted segmentation divided by their union,

$$\text{IoU} := \frac{TP}{TP + FP + FN},$$

where TP, FP, and FN are pixel-wise true positives, false positives, and false negatives. We report IoU score of the model on each land-cover class as well as mean over class-wise IoU (referred to as **Mean**) over the 7 classes.

Fairness metrics: Besides looking at IoU on the rural and urban subsets and comparing the two values, we devise three additional metrics that quantify how accuracy is distributed across the classes. The metrics have been used in existing fairness analysis of segmentation models [19]. These are:

1. **Class-std**. Standard deviation of IoUs across the 7 classes;

2. **Worst**. IoU of the worst-performing class (Worst);

3. **Sorted 30% (bottom, top)**. Mean of the bottom 30% and top 30% classes of the sorted class IoUs. In our case, 30% is 2 classes out of 7.

Next, we describe the setup for the two transfer tasks.

4.2. Transfer learning

As mentioned earlier, we consider the setting of unsupervised domain adaption (UDA) that is we have a single task (image segmentation) on the two domains. We assume access to images and labels for the source (train) domain and only the images for the target (test) domain. This is a practical setting in satellite imagery since collecting images is inexpensive due to advancements in remote sensing, however, annotating the segmentation labels is expensive. Thus, we want to be able to use the labelled source images to segment known but as yet unlabelled target domain.

We consider two UDA methods which performed the best on the LoveDA benchmark [52] – class-balanced self-training (**CBST**) [59] and instance adaptive self-training (**IAST**) [33]. CBST optimizes the generation of pseudo-labels used during self-training to be more balanced among the classes by using class-wise confidence scores. IAST adaptively adjusts the pseudo-label generation to improve the diversity of pseudo-labels and saves useful information from hard instances. We also use a natural method which ignores transfer learning and trains only with data from the source domain (**No adaptation**).

For all transfer learning experiments, we use a DeepLabV3+ [7] network with ResNet50 encoder. An Adam optimizer is used with a momentum of 0.9. The batch size is 8 and the total training iterations are 15000. Other experimental setup parameters are the same as in the semantic segmentation task.

4.2.1 Task B - Transfer across districts

First, we consider the scenario where a model is transferred across different geographical locations, which in our case are administrative districts. **Source domain** comprises of 8 districts: Gulou, Qinhuai, Qixia, Jinghan (urban), and Pukou, Gaochun, Lishui, Jingxia (rural); and **Target domain** has 4 districts: Yuhuatai, Jintan (urban), and Liuhe, Huangpi (rural). The "No adaptation" and two UDA methods (CBST, IAST) are applied to train the network on the source domain, and are tested on urban and rural images from the unseen target domain, separately. The same accuracy and fairness metrics listed in Section 4.1.2 are used for evaluation.

Model	Mean			Class-std		
	rural	urban	rural − urban (%)	rural	**urban**	rural − urban (%)
UNet	**0.639**	0.595	0.044 (6.9%)	0.106	0.0946	0.0114 (10.8%)
DeepLabV3+	0.632	0.597	0.035 (5.5%)	0.0982	**0.0896**	0.0086 (8.8%)

Model	Worst			Sorted 30% (bottom, top)		
	rural	**urban**	rural − urban (%)	rural	urban	rural − urban (%)
UNet	0.453	**0.474**	−0.021 (-4.4%)	(0.491, 0.742)	(0.480, 0.705)	(0.011 , 0.037)(2.2%, 5.0%)
DeepLabV3+	0.473	0.473	0 (n/a)	(0.504, 0.740)	(0.489, 0.706)	(0.015, 0.034)(3.0%, 4.6%)

Table 1. **Task A: Evaluation on single-domain semantic segmentation.** Two networks, UNet and DeepLabV3+, are tested on rural and urban districts from the same domain as training set. For metrics, Mean, Class-std, and Worst, the better performing group (between rural and urban) is in bold. The difference in performance between rural and urban is shaded. Typically, performance is better for rural than urban.

Method	Mean			Class-std		
	rural	**urban**	rural − urban (%)	rural	**urban**	rural − urban (%)
No adaptation	0.364	0.486	−0.122 (−25%)	0.200	0.135	0.065 (33%)
CBST	0.374	**0.523**	−0.149 (−28%)	0.215	**0.105**	0.110 (51%)
IAST	0.376	0.493	−0.117(−24%)	0.223	0.135	0.088 (39%)

Method	Worst			Sorted 30% (bottom, top)		
	rural	**urban**	rural − urban (%)	rural	urban	rural − urban (%)
No adaptation	0.0609	0.244	−0.183 (−75%)	(0.098, 0.581)	(0.317, 0.630)	(−0.219, −0.049) (−69%, −7.7%)
CBST	0.0172	**0.362**	−0.345 (−95%)	(0.0943, 0.609)	(0.398, 0.647)	(−0.304, −0.038) (−76%, −5.9%)
IAST	0.0304	0.232	−0.202 (−87%)	(0.0772, 0.598)	(0.327, 0.640)	(−0.250, −0.042) (−76%, −6.6%)

Table 2. **Task B: Evaluation of transfer across districts.** Three methods (No adaptation, CBST, IAST) are trained source districts, and evaluated on target rural and target urban districts. For the metrics, Mean, Class-std, and Worst, the better performing group (between rural and urban) is in bold. The differences in performance between rural and urban are shaded. Models have high unfairness upon transfer.

4.2.2 Task C - Transfer across urban and rural areas

Second, we consider the scenario where the segmentation model is transferred either from urban to rural areas or from rural to urban areas. The source and target domain consists of data from the same set of districts. So for this task, the only source for domain discrepancy is rural and urban discrepancy. The no- adaptation method and two UDA methods are trained on the source domain, and tested on the target domain. Evaluation metrics are the same as earlier.

5. Results

We summarize results for the single-domain in Table 1. Both the networks (UNet, DeepLabV3+) have better overall accuracy, shown with Mean IoU over the 7 classes, for the rural districts compared to the urban districts. Fairness metrics such as IoU of the worst class and mean IoU of 30% bottom classes are comparable between rural and urban. The worst class is Barren for both rural and urban. The 30% bottom classes include Barren and Road for rural, and Barren and Forest for urban (see Table A.1 in Appendix for class-wise results). Rural results show higher mean IoU

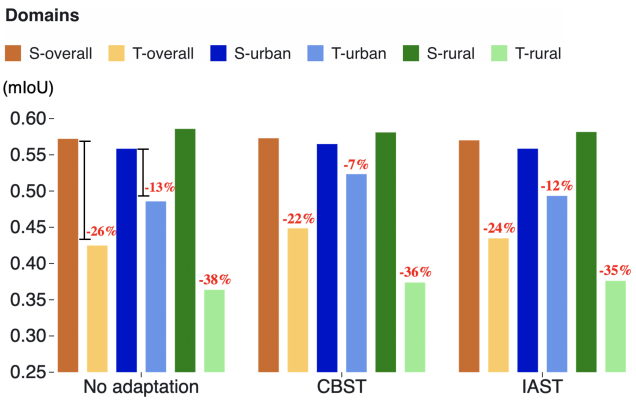

Figure 2. **Task B: Mean IoU upon transfer across districts.** Mean IoU on the union of rural and urban data from the source (S-overall) and target (T-overall), urban data from the source (S-urban) and target (T-urban), rural data from the source (S-rural) and target (T-rural) is plotted when transferring models across districts. No adaptation is the source-only method, CBST and IAST are UDA methods. While overall accuracy drops on transfer, UDA methods have smaller accuracy drop.

Sub-task	Method	Mean	Class-std	Worst	Sorted 30% (bottom, top)
Rural→Urban	No adaptation	0.437	0.108	0.301	(0.322, 0.566)
	CBST	**0.469**	0.123	**0.326**	(**0.332**, 0.617)
	IAST	0.443	0.175	0.205	(0.211, **0.638**)
Urban→Rural	No adaptation	0.426	0.108	0.226	(0.271, 0.531)
	CBST	**0.467**	0.129	0.228	(0.283, **0.599**)
	IAST	0.454	0.120	**0.229**	(**0.307**, 0.592)

Table 3. **Task C: Evaluation of urban-rural transfer.** Three methods (No adaptation, CBST, IAST) are trained on rural districts and evaluated on urban districts, and vice versa. Results with the most improvements are marked in bold. UDA methods improve Mean IoU compared to No adaptation but increase standard deviation of IoUs across classes.

from top 30% classes than urban results, but higher class-wise standard deviation. Overall, we observe rural-urban disparities in all four metrics.

For Task B on transfer learning across districts, we summarize the results in Figure 2 and Table 2. Figure 2 visualizes the mean IoU metric for both source and target districts. Based on the first two bar plots (dark and light orange) for No adapation, CBST, and IAST, we conclude that UDA methods improve overall segmentation accuracy on the target domain (T-overall) compared to No adaptation (a decrease of 22% and 24% vs that of 26%). Similar trend is observed for each of the source-target pairs for rural and urban separately. However, the performance gap *between* the rural and urban data from target (T-rural and T-urban) remains large. For instance, from Table 2 we observe that CBST obtains mean IoU of 0.523 on urban area which is better than the "No adaptation" 0.486, and IAST obtains 0.376 on rural area better than the "No adaptation" 0.364. However, the networks remain unfair across rural-urban groups after the transfer (large values in the rural − urban columns). UDA methods further lower fairness: CBST increases the difference of mean IoU between urban and rural by 22% (-0.122 to -0.149), increases the difference of standard deviation by 69% (0.065 to 0.11), and increases the difference of worst-performing class' IoU by 89% (-0.183 to -0.345).

Next, we examine Task C on transfer learning from urban domain to rural domain and vice versa. Results are summarized in Table 3. For both the transfer directions, the two UDA methods improve the overall accuracy, shown as higher mean IoU, higher IoU on the worst-preforming class, and higher mean IoU on bottom and top 30% classes. However, compared to "No adaptation", UDA methods disperse model performance across the classes, measured by higher standard deviation. For example, CBST increases Class-std from 0.108 to 0.123 on rural to urban transfer and from 0.108 to 0.129 on urban to rural transfer. Looking more closely into the CBST method which obtains the best overall transfer accuracy (0.469 and 0.467), its performance on each class is visualized in Figure 3. We observe that the IoU changes for each class upon transfer are highly unequal. For example, in transferring from Rural→Urban, the network retains accuracy on the Water and Road classes, but lost significant accuracy (53%) on the Forest class. In transferring from Urban→Rural, accuracy drops significantly on Road, Water, and Barren classes, but increases by 33% on the Forest class.

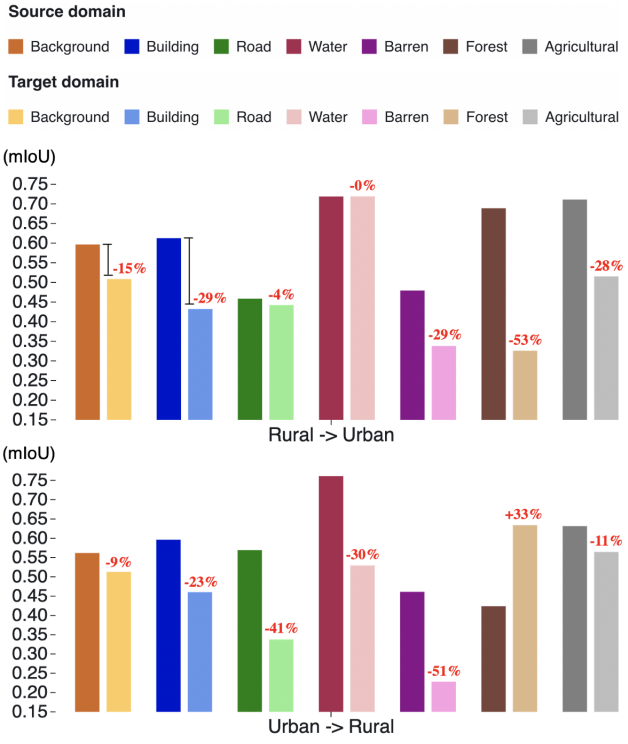

Figure 3. **Task C: Mean IoU upon transfer across rural-urban.** Mean IoU for 7 landscape classes on source and target domain when transferring from rural area to urban area, and from urban area to rural area, with the UDA method CBST. Performance changes vary substantially by class.

6. Discussion

For the locations included in this study, segmentation results showed a disparity in performance between rural and urban areas. Though the two groups obtain similar accuracy on the respective bottom 30% performing land-cover classes, rural areas obtain better accuracy on the top 30% performing classes. Moreover, performance distribution across classes are different between rural and urban images. Specifically, the segmentation model detects Forest and Agriculture classes well in their rural form, and detects Road and Water classes well in their urban form (detailed results are reported in Table 5 in the Appendix). Due to urbanization, rural and urban areas have clear landscape differences. For example, roads are typically wider in the urban scenes and narrower in rural scenes and water takes on larger shapes like lakes in urban scenes, and smaller shapes like ditches in rural scenes [52]. This may explain why the networks show advantages in urban images on Road and Water classes. Moreover, agricultural land covers large area and is continuously distributed in rural scenes. The percentage of pixels with Agriculture and Forest elements is also higher [52]. This can facilitate learning on these two classes in rural areas as compared to urban areas.

We considered two practical transfer learning tasks with satellite images and assess network fairness while transferring across geographical locations, and across rural and urban areas. Broadly, we observed that when transferring across districts, networks made more unfair predictions on data from the new domain than data from the same domain as training. For the network trained without any adaptation, the mean IoU accuracy difference between rural and urban images on the target domain is around 64% higher than the difference reported in the single-domain task. Similarly, all other fairness metrics show much higher differences between groups on transferring to the target domain. Notably, when applying UDA methods, CBST and IAST, transfer accuracy was improved, but at the cost of fairness damage. These methods further enlarged the performance gap between rural and urban groups measured in all four metrics. These findings indicate a need for new domain adaptation methods that tackle the challenge of fair transfer learning.

One of the possible reasons why the network can better adapt to urban images than rural images is the unequal domain discrepancy. To estimate how similar the source and target images are in the rural and urban groups, we use two metrics – Proxy-A-distance (PAD) [12] and Maximum mean discrepancy (MMD) [15]. Both measure the dissimilarity between data distributions of different domains. We randomly sample 100 images from each domain at a time and ran 30 trials to compute the two measures. The mean and standard deviation of distance across the trials are reported in Table 4. We observe that the raw satellite images

Group	Source	Target	PAD	MMD
Urban	Gulou Qinhuai Qixia Jinghan	Yuhuatai Jintan	0.26 (±0.12)	0.207 (±0.0235)
Rural	Pukou Gaochun Lishui Jingxia	Liuhe Huangpi	0.64 (±0.21)	0.262 (±0.0437)

Table 4. **Shift in raw image distribution.** Measurement of source-target domain distance using two metrics – Proxy-A-distance (PAD) and Maximum mean discrepancy (MMD). The implementation is based on the online codebase in [51]. Rural images shift more than urban images.

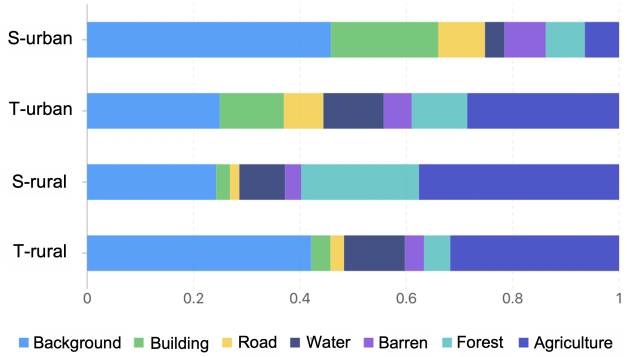

Figure 4. **Shifts in class distribution.** Class distribution in terms of proportion of pixels per class is plotted for urban images from source domain (S-urban), urban images from target domain (T-urban), rural images from source domain (S-rural), and rural images from target domain (T-rural). Class distribution is substantially different for all subsets.

are overall more dissimilar between source and target districts for rural than for urban, which is a likely cause of the unequal transfer learning performance between the two groups. Figure 1 illustrates example images from both domains and the segmentation predictions our network made on the target. We see that Buildings across source and target districts are of similar shape and arrangement for urban, but they are disordered and dissimilar for rural. Accordingly, the model segmentation map shows that the model segments urban buildings well but fails to detect most of the rural buildings. This observation indicates the importance of checking class differences besides overall differences across the whole image.

Along these lines, we define pixel-wise class distribution as the proportion of pixels belonging to each class. We assess shifts in the class distribution between source and target for both rural and urban images. Class distributions are shown in Figure 4. For urban locations, the Water and

Agriculture classes have the largest shifts from source to target. For rural, the Background and Forest classes have the largest shifts. Indeed there are large class shifts between source and target for both the urban and rural data. For example, the class distribution of urban target data seems more rural-like with an increased proportion in the Agriculture class. This emphasizes the internal variation in rural and urban categories. Moreover, as our data consists of images from just one set of locations, data from different locations are needed for more generalizable conclusions. However, based on the selection of classes examined in our data which are common land-use classes globally, some results (such as in Table 4 indicate common threads that can be applicable to rural-urban disparities in general.

In the second transfer learning task, the networks show unfairness on rural to urban domain transfer. Differences between rural and urban scenes provide explanation for why almost all classes lost accuracy on the target domain. Some classes show opposite transfer performance in the two sub-tasks of Task C. The Road class lost only 4% accuracy on Rural→Urban transfer but lost 41% accuracy on Urban→Rural transfer. The Water class shows similar patterns, whereas the Forest class lost 53% accuracy on Rural→Urban transfer but gains 33% accuracy on Urban→Rural transfer. These observations indicate that for some classes, the features learnt by the networks from rural scenes can be easily adapted to interpret urban scenes, and some classes have the opposite case. From this perspective, features of different classes can have very different generalization ability, which will cause the transfer performance to be highly unequal across classes. This feature-specific characteristic may be leveraged in the design of future transfer learning methods.

7. Conclusion

Transfer learning models for semantic segmentation are often evaluated based on overall accuracy metrics. Here, we expand the scope of their evaluation by conducting a systematic fairness evaluation when models are transferred across domains. We examine the performance of two unsupervised domain adaptation methods on a large-scale public satellite imagery dataset. Model fairness is evaluated between rural and urban locations as the models are trained and tested across administrative districts. Based on the experiments, we conclude that the domain adaptation methods we study can be improved in terms of retaining model fairness across rural and urban data. Domain adaptation improves overall accuracy at the cost of decreasing fairness on test domain. Further, more shifts in the raw image distribution and pixel-wise class distribution result in more performance drop. Broadly, our findings demonstrate potential fairness problems when working with satellite image data sourced from different locations. Also, the findings indicate the need for developing methods focused on fair transfer learning, such as through new model architectures or loss functions.

References

[1] Nabeel Abdur Rehman, Umar Saif, and Rumi Chunara. Deep landscape features for improving vector-borne disease prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 44–51, 2019. 1

[2] Michiel Bakker, Humberto Riverón Valdés, Patrick D Tu, Krishna P Gummadi, Kush R Varshney, Adrian Weller, and AS Pentland. Fair enough: Improving fairness in budget-constrained decision making using confidence thresholds. 2020. 2

[3] Ritika Brahmadesam and Kush R Varshney. Fairly estimating socioeconomic status under costly feature acquisition, 2021. 2

[4] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 23–24 Feb 2018. 2

[5] AP Carleer, Olivier Debeir, and Eléonore Wolff. Assessment of very high spatial resolution satellite image segmentations. *Photogrammetric Engineering & Remote Sensing*, 71(11):1285–1294, 2005. 1

[6] Javiera Castillo-Navarro, Bertrand Le Saux, Alexandre Boulch, Nicolas Audebert, and Sébastien Lefèvre. Semi-supervised semantic segmentation in earth observation: The minifrance suite, dataset analysis and multi-task network study. *Machine Learning*, pages 1–36, 2021. 2

[7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 4

[8] Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. Fair transfer learning with missing protected attributes. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 91–98, 2019. 3

[9] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 172–181, 2018. 3

[10] Wei Du and Xintao Wu. *Fair and Robust Classification Under Sample Selection Bias*, page 2999–3003. Association for Computing Machinery, New York, NY, USA, 2021. 3

[11] Mark Everingham, SM Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The

pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 4

[12] 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. 7

[13] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, Pablo Martinez-Gonzalez, and Jose Garcia-Rodriguez. A survey on deep learning techniques for image and video semantic segmentation. *Applied Soft Computing*, 70:41–65, 2018. 4

[14] Omid Ghorbanzadeh, Alessandro Crivellari, Pedram Ghamisi, Hejar Shahabi, and Thomas Blaschke. A comprehensive transferability evaluation of u-net and resu-net for landslide detection from sentinel-2 data (case study areas from taiwan, china, and japan). *Scientific Reports*, 11(1):1–20, 2021. 2

[15] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012. 7

[16] Ananya Gupta, Elisabeth Welburn, Simon Watson, and Hujun Yin. Cnn-based semantic change detection in satellite imagery. In *International Conference on Artificial Neural Networks*, pages 669–684. Springer, 2019. 1

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[18] Corentin Henry, Seyed Majid Azimi, and Nina Merkle. Road segmentation in sar satellite images with deep fully convolutional neural networks. *IEEE Geoscience and Remote Sensing Letters*, 15(12):1867–1871, 2018. 1

[19] Szabó, A., Jamali-Rad, H. & Mannava, S. Tilted cross-entropy (TCE): Promoting fairness in semantic segmentation. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 2305-2310 (2021) 3, 4

[20] Vladimir Iglovikov, Sergey Mushinskiy, and Vladimir Osin. Satellite imagery feature detection using deep convolutional neural network: A kaggle competition, 2017. 3

[21] Javed Iqbal and Mohsen Ali. Weakly-supervised domain adaptation for built-up region segmentation in aerial and satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 167:263–275, 2020. 1

[22] Kazi Aminul Islam. *Deep Learning Approaches for Seagrass Detection in Multispectral Imagery*. PhD thesis, Old Dominion University, 2021. 1

[23] Shunping Ji, Shiqing Wei, and Meng Lu. A scale robust convolutional neural network for automatic building extraction from aerial and satellite imagery. *International journal of remote sensing*, 40(9):3308–3322, 2019. 2

[24] Marios Krestenitis, Georgios Orfanidis, Konstantinos Ioannidis, Konstantinos Avgerinakis, Stefanos Vrochidis, and Ioannis Kompatsiaris. Oil spill identification from satellite images using deep neural networks. *Remote Sensing*, 11(15):1762, 2019. 1

[25] Anja Lambrecht and Catherine Tucker. Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Management Science*, 65(7):2966–2981, 2019. 2

[26] Ke Li, Mingju Wang, Yixin Liu, Nan Yu, and Wei Lan. A novel method of hyperspectral data classification based on transfer learning and deep belief network. *Applied Sciences*, 9(7), 2019. 2

[27] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *ICML*, pages 6357–6368, 2021. 3

[28] Yikun Li and Timo R Bretschneider. Semantic-sensitive satellite image retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 45(4):853–860, 2007. 1

[29] Bing Liu, Xuchu Yu, Anzhu Yu, and Gang Wan. Deep convolutional recurrent neural network with transfer learning for hyperspectral image classification. *Journal of Applied Remote Sensing*, 12(2):1 – 17, 2018. 2

[30] Debmalya Mandal, Samuel Deng, Suman Jana, and Daniel Hsu. Ensuring fairness beyond the training data. *Advances in neural information processing systems*, 2020. 3

[31] Amy McGovern, Imme Ebert-Uphoff, David John Gagne II, and Ann Bostrom. The need for ethical, responsible, and trustworthy artificial intelligence for environmental sciences. *arXiv preprint arXiv:2112.08453*, 2021. 2

[32] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021. 2, 3

[33] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *European conference on computer vision*, pages 415–430. Springer, 2020. 2, 4

[34] Khaled Moghalles, Heng-Chao Li, Zaid Al-Huda, and Essa Abdullah Hezzam. Multi-task deep network for semantic segmentation of building in very high resolution imagery. In *2021 International Conference of Technology, Science and Administration (ICTSA)*, pages 1–6. IEEE, 2021. 1

[35] Anton Obukhov, Stamatios Georgoulis, Dengxin Dai, and Luc Van Gool. Gated crf loss for weakly supervised semantic image segmentation. *arXiv preprint arXiv:1906.04651*, 6, 2019. 2

[36] Union of Concerned Scientists. Ucs satellite database. https://www.ucsusa.org/resources/satellite-database, 2022. [Accessed 23-March-2022]. 1

[37] Luca Oneto, Michele Donini, Massimiliano Pontil, and Andreas Maurer. Learning fair and transferable representations with theoretical guarantees. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 30–39, 2020. 3

[38] Anthony Ortiz, Caleb Robinson, Dan Morris, Olac Fuentes, Christopher Kiekintveld, Md Mahmudulla Hassan, and Nebojsa Jojic. Local context normalization: Revisiting local normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11276–11285, 2020. 1

[39] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. 1

[40] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 4

[41] Ashkan Rezaei, Anqi Liu, Omid Memarrast, and Brian D. Ziebart. Robust fairness under covariate shift. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(11):9419–9427, May 2021. 3

[42] Yuji Roh, Kangwook Lee, Steven Whang, and Changho Suh. Sample selection for fair and robust training. *Advances in Neural Information Processing Systems*, 34, 2021. 3

[43] Esther Rolf, Jonathan Proctor, Tamma Carleton, Ian Bolliger, Vaishaal Shankar, Miyabi Ishihara, Benjamin Recht, and Solomon Hsiang. A generalizable and accessible approach to machine learning with global satellite imagery. *Nature communications*, 12(1):1–11, 2021. 1

[44] 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. 4

[45] Michael Schmitt, Jonathan Prexl, Patrick Ebel, Lukas Liebel, and Xiao Xiang Zhu. Weakly supervised semantic segmentation of satellite images for land cover mapping–challenges and opportunities. *arXiv preprint arXiv:2002.08254*, 2020. 1, 2

[46] Candice Schumann, Xuezhi Wang, Alex Beutel, Jilin Chen, Hai Qian, and Ed H Chi. Transfer of machine learning fairness across domains. *arXiv preprint arXiv:1906.09688*, 2019. 3

[47] Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 3–13, New York, NY, USA, 2021. Association for Computing Machinery. 3

[48] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010. 1

[49] An Tran, Ali Zonoozi, Jagannadan Varadarajan, and Hannes Kruppa. Pp-linknet: Improving semantic segmentation of high resolution satellite imagery with multi-stage training. In *Proceedings of the 2nd Workshop on Structuring and Understanding of Multimedia heritAge Contents*, pages 57–64, 2020. 2

[50] Michele Volpi and Vittorio Ferrari. Semantic segmentation of urban scenes by learning local class interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–9, 2015. 1, 2, 3

[51] Jindong Wang et al. Everything about transfer learning and domain adapation. http://transferlearning.xyz. 7

[52] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021. 3, 4, 7

[53] Sherrie Wang, William Chen, Sang Michael Xie, George Azzari, and David B Lobell. Weakly supervised deep learning for segmentation of remote sensing imagery. *Remote Sensing*, 12(2):207, 2020. 2

[54] Zhaobin Wang, E Wang, and Ying Zhu. Image segmentation evaluation: a survey of methods. *Artificial Intelligence Review*, 53(8):5637–5674, 2020. 4

[55] Renzhe Wu, Guoxiang Liu, Rui Zhang, Xiaowen Wang, Yong Li, Bo Zhang, Jialun Cai, and Wei Xiang. A deep learning method for mapping glacial lakes from the combined use of synthetic-aperture radar and optical satellite images. *Remote Sensing*, 12(24):4020, 2020. 1

[56] Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. To be robust or to be fair: Towards fairness in adversarial training. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11492–11501. PMLR, 18–24 Jul 2021. 3

[57] Fei Zhao and Chengcui Zhang. Building damage evaluation from satellite imagery using deep learning. In *2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 82–89. IEEE, 2020. 1

[58] Yuyin Zhou, Shih-Cheng Huang, Jason Alan Fries, Alaa Youssef, Timothy J Amrhein, Marcello Chang, Imon Banerjee, Daniel Rubin, Lei Xing, Nigam Shah, et al. Radfusion: Benchmarking performance and fairness for multimodal pulmonary embolism detection from ct and ehr. *arXiv preprint arXiv:2111.11665*, 2021. 2

[59] 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, 4