Domain Adaptation for Real-time Semantic Segmentation in Remote Sensing

Borella Simone
Polytechnic University of Turin
s317774@studenti.polito.it

Alfarano Alex
Polytechnic University of Turin
s319818@studenti.polito.it

Arefeh A. Nazari
Polytechnic University of Turin
s329031@studenti.polito.it

Abstract

Semantic segmentation in high spatial resolution remote sensing (RS) images poses significant challenges due to domain gaps between diverse geographic and environmental contexts. This paper investigates domain adaptation techniques to enhance segmentation performance in cross-domain scenarios, leveraging the real-time capabilities of the PIDNet network. Specifically, we explore Unsupervised Domain Adaptation techniques like data augmentation, adversarial strategy, and self-training methods, including Domain Adaptive Contrastive Segmentation (DACS) and an extended variant. This study focuses on combining advanced domain adaptation techniques with lightweight, real-time networks to address domain shift between Urban and Rural environments in high-resolution RS semantic segmentation task. The code is available at https://github.com/SimoneBorella/semantic-segmentation-domain-adaptation

1. Introduction

Semantic segmentation is a fundamental task in computer vision, with the objective of assigning each pixel in an image to a specific class label [1]. It plays a crucial role in various applications like autonomous driving, medical imaging and environmental monitoring, where detailed and accurate pixel-level classification is essential. High spatial resolution remote sensing (RS) technology can help us to better understand the geographical and ecological environment (Figure 1).

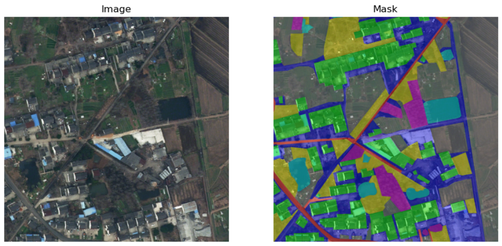

Figure 1. Semantic segmentation in Remote Sensing application.

However, the high variability of RS data, caused by differences in geographic regions and environmental conditions, presents unique challenges. High-resolution data captures finer details, enabling better differentiation between classes. However, it also increases computational complexity and highlights domain discrepancies between datasets collected under different conditions, such as urban versus rural settings. These domain gaps make it difficult for models trained on a source dataset to generalize to another target dataset, and so here comes the need of Domain Adaptation (DA) techniques. Unsupervised Domain Adaptation (UDA) reduces the need for labeling target domain data by leveraging labeled source domain images and unlabeled target domain images during training, enabling models to perform efficiently on the different domains. Among UDA methods, adversarial learning employs a discriminator to distinguish between features extracted from the source and target domains, while the feature extractor aims to generate domain-invariant representations that confuse the discriminator. This adversarial interplay aligns the feature distributions of the two domains. Another UDA approach, Domain Adaptation via Cross-domain Mixed Sampling (DACS), has gained importance using mixed images obtained from source and target domains to enhance feature alignment and robustness. However, standard DACS methods may struggle with class imbalance, a prevalent issue in RS datasets where certain categories dominate while others are underrepresented. To address this limitation, an extension to DACS has been proposed, introducing Gradual Class Weight (GCW) mechanisms to address class imbalance and Local Dynamic Quality (LDQ) as a robust pseudo-labeling strategy.

Traditional architectures such as DeepLab have demonstrated state-of-the-art performance in semantic segmentation tasks. DeepLab's Atrous Spatial Pyramid Pooling (ASPP) and multi-scale feature extraction capabilities have proven their effectiveness for high-resolution data. However, these architectures often come at the cost of computational efficiency, making them less suitable for real-time applications. To address this limitation, real-time networks like PIDNet have emerged. With its lightweight design, PIDNet reaches a balance between accuracy and efficiency, introducing high-speed inference without significantly compromising performance.

This paper explores and compares some UDA strategies in Remote Sensing (RS) semantic segmentation task. By addressing the domain shift between Urban and Rural domains from LoveDA dataset, we aim to evaluate UDA techniques alongside PIDNet real-time architecture to tackle the challenges of high-resolution RS segmentation.

2. Background

LoveDA - LoveDA [3] is an RS Land Cover benchmark dataset designed for adaptive segmentation. It contains high-resolution (1024 x 1024) remote sensing images, obtained from the Google Earth platform, with pixel-level annotations. The dataset includes images from two different domains, Rural and Urban, containing respectively images of geographical areas with lower and higher population density from three different Chinese cities. The dataset encompasses seven distinct classes: *Background*, *Building*, *Road*, *Water*, *Barren*, *Forest*, and *Agriculture*. The training, validation, and test sets are split so that they are spatially independent, thus enhancing the difference between the split sets. The urban areas always contain more artificial structures such as buildings and roads due to their high population density. In contrast, the rural areas have more agricultural land. The inconsistent class distributions between the urban and rural scenes increases the difficulty of model generalization, making the dataset the perfect choice to benchmark domain adaptation strategies.

DeepLabV2 - DeepLab [4] is one of the most popular deep learning architectures for semantic segmentation. It employs atrous convolution (also known as dilated convolution) to extract multi-scale contextual information without reducing the spatial resolution of feature maps. This allows the capture of both very fine-grained details and broader contextual features. DeepLab also introduced the Atrous Spatial Pyramid Pooling (ASPP) module, which ensures multiscale context aggregation, by employing multiple parallel filters with different rates. This allows the spatial resolution of the resulting feature maps to be maintained, which is important for a dense prediction task such as semantic segmentation.

PIDNet - PIDNet [2] is a lightweight and efficient network specifically designed for real-time semantic segmentation tasks inspired by the PID (Proportional Integrative Derivative) controller. It introduces a novel architecture that balances high accuracy with low computational complexity, making it suitable for edge devices and real-time applications. The network is structured into three main branches: the P branch for capturing detailed spatial features using depthwise separable convolutions to reduce computational cost, the I branch for extracting global context using dilated convolutions to maintain a large receptive field without increasing resolution, and the D branch that focuses on detecting object boundaries refining segmentation results, especially around edges. A final feature fusion module combines features from the three branches. This design ensures a trade-off between detail preservation and contextual understanding. The variants of PIDNet (Small, Medium and Large) differ from width and depth, making them suitable for different hardware constraints and efficiency requirements.

Domain Adaptation - Recent deep neural network (DNN)-based methods require large amounts of annotated data and often suffer from performance degradation due to domain gaps, which arise from differences in feature distributions across datasets. Domain adaptation (DA) aims to address this issue by enabling models to transfer knowledge between different domains. DA methods can be classified as supervised, semi-supervised, or unsupervised, based on whether they have access to the labels of the target domain. UDA methods [5] can be grouped into three categories: generative-based, adversarial learning, and self-training (ST) methods. Generative-based methods, such as image translation or style transfer, make source and target images visually similar, allowing segmentation models to be trained with translated images. However, the performance heavily depends on the quality of the translated images, as pixel-level flaws can impact accuracy. Adversarial learning methods introduce a discriminator to minimize the domain gap by aligning feature distributions between source and target domains. These methods, while effective, are often sensitive to hyperparameters and difficult to train. Self-training (ST) methods rely exclusively on the segmentation network. They treat high-confidence predictions on unlabeled target data as pseudo-labels, which are then used along with labeled source data to improve performance in the target domain. ST methods have proved promising, but issues related to pseudo-label accuracy and class imbalance remain a challenge. Data augmentation techniques are widely employed to enhance model generalization and improve alignment with the target domain.

Adversarial learning - The adversarial learning strategy employed is presented in [9] where a segmentation network is trained alongside a discriminator to distinguish segmentation outputs from the source and target domains. This method uses predictions from both source domain, with annotations, and target domain, without annotations, as input to the discriminator. By applying adversarial loss, gradients propagate through the discriminator to the segmentation network, encouraging the segmentation network to generate target domain outputs that resemble source domain predictions. Building on this foundation, multi-level adversarial framework extends adversarial learning by incorporating discriminators at different feature levels within the segmentation network. This multi-level strategy refines alignment across hierarchical representations, improving the quality of segmentation adaptation compared to single-level approaches.

DACS - Domain Adaptation via Cross-domain Mixed Sampling (DACS) [8] introduces a self-training approach to improve semantic segmentation performance across different domains. The core idea of DACS is to generate mixed samples by blending images from both the source and target domains, allowing the model to learn better representations by exposing it to both domains simultaneously, bridging the gap between the source and target domains. During the blending operation, pixels belonging to a randomly chosen set of classes from the source image are kept, while remaining pixels are replaced by the target image. A key feature of DACS is its reliance on pseudo-labels for the target domain. Pseudo-labels are generated using confident predictions from the segmentation network itself and are incorporated into the mixed samples during training. This process allows the model to adapt progressively to the target domain without requiring additional labeled data. Furthermore, DACS do not attempt to combine consistency regularization with alignment of image distributions, instead, enforce consistency between predictions of images in the target domain and images mixed across domains.

DACS with GCW and LDQ - The issue of imbalanced category proportions poses a significant challenge to the performance of most standard learning algorithms. Common strategies to address class imbalance often suffer from high computational costs or require extensive hyperparameter tuning. To tackle these challenges, two strategies are proposed to enhance the framework's effectiveness [7]. First, they introduce Gradual Class Weights (GCW), a dynamic method for adjusting class weights in the source domain, to mitigate the class imbalance problem. A previous method assigned a class weight inversely proportional to the statistical frequency of the class in the dataset. This approach ensures that rare classes receive more attention compared to common ones. However, due to the randomness in sampling, the class distribution in the sampled data may differ from the distribution calculated over the entire dataset in advance. For this reason, the weights are updated iteratively after a warmup phase during the initial stages of training, enhancing model stability and robustness.

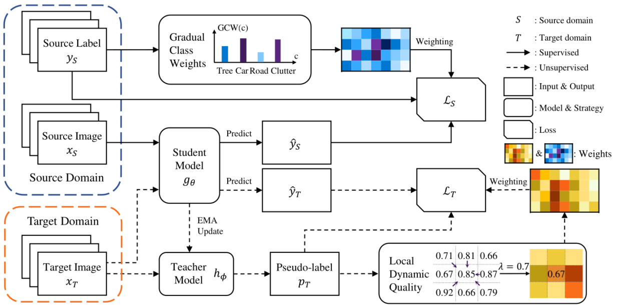

Figure 2. Overview of DACS framework with Gradual Class Weights and Local Dynamic Quality.

For each image n and each class c, the Gradual Class Weights (GCW) are computed as:

$$\mathrm{GCW}(n, c) = \beta \cdot \mathrm{GCW}(n-1, c) + (1-\beta) \cdot W(n, c),$$

$$W(n, c) = \frac{C \cdot \exp\left(\frac{1-f_c}{T}\right)}{\sum_{c'=1}^{C} \exp\left(\frac{1-f_{c'}}{T}\right)}$$

Where:

- C: Cardinality of the classes set.
- f_c: The frequency of class c in the label distribution.
- β: A mixing parameter controlling the influence of previous weights.
- T: A temperature parameter. Higher T leads to a more uniform distribution while a lower one makes the model pay more attention to the rare classes

The second strategy introduces a different approach to evaluate the quality of pseudo-labels. In DACS implementation pseudo-label weights are calculated as the proportion of pixels with confidence above a certain threshold. The proposed Local Dynamic Quality (LDQ) method compute the quality of a pseudo-label based on the ratio of high-quality surrounding pseudo-labels, which can be efficiently calculated through convolution.

$$LDQ(h, w) = \frac{1}{(2K+1)^2} \sum_{i=-K}^{K} \sum_{j=-K}^{K} q(h+i, w+j)$$

Where:

- (h, w): Location in which the weight is calculated.
- K: Depth of neighbors.
- $(2K+1)$: size of the convolution kernel.

Figure 2 shows the integration of GCW and LDQ strategies with DACS framework.

3. Experiments and results

3.1. Overview

The adopted methodology is divided into several key phases, as outlined below:

- **Baseline Evaluation:** Assess the performance of DeepLabV2 and PIDNet models, establishing lower and upper performance bounds to quantify the impact of domain shift.
- **Domain Generalization:** Apply data augmentation techniques to mitigate domain shift and enhance model robustness.
- **Adversarial Learning:** Evaluate the effectiveness of adversarial domain adaptation in aligning source and target distributions.
- **DACS Implementation:** Investigate the impact of Domain Adaptation via Cross-domain Mixed Sampling (DACS) on segmentation performance.
- **DACS Enhancement:** Extend DACS with Gradual Class Weights (GCW) for class balance and Local Dynamic Quality (LDQ) for improved pseudo-labeling reliability.

Each phase addresses specific challenges in adapting semantic segmentation models to the domain shift between Urban and Rural domains of LoveDA dataset. The baseline evaluation provides a clear understanding of the performance gap caused by domain shift, while domain generalization techniques aim to bridge this gap by increasing the diversity of the training data. Adversarial learning and DACS are employed to align the feature distributions of the source and target domains, ensuring that the model generalizes well to unseen data. Finally, the enhancements to DACS, such as GCW and LDQ, are introduced to try to refine the adaptation process, ensuring balanced class representation and reliable pseudo-labeling.

Common implementation details - The different training operations share common characteristics. DeepLabV2 with ResNet101 backbone and PIDNet-S models are employed, both pretrained on ImageNet. All trainings are performed on 20 epochs with a batch size of 6. Optimizations were performed using the Stochastic Gradient Descent (SGD) algorithm with a momentum of 0.9 and a weight decay of 0.0005. The initial learning rate is set to 0.001, decreased using a polynomial decay with exponent 0.9. To evaluate the performance of the different configurations across the various experiments, the Mean Intersection over Union (mIoU) is employed as the evaluation metric.

Source Domain	mIoU (%, Urban)	mIoU (%, Rural)
Rural	22.88	19.93
Urban	26.27	13.91

Table 1. Baseline training results with DeepLabV2 on Urban and Rural domains.

Source Domain	mIoU (%, Urban)	mIoU (%, Rural)
Rural	41.97	33.16
Urban	39.24	25.44

Table 2. Baseline training results with PIDNet-S on Urban and Rural domains.

3.1.1 Baseline training with DeepLabV2 and PIDNet

Implementation Details - The baseline experiments were conducted with two models: DeepLabV2 and PIDNet-S, both pretrained on ImageNet. DeepLabV2 is trained using the Cross-Entropy (CE) loss function to optimize pixel-wise classification accuracy, while PIDNet-S is trained using Online Hard Example Mining (OHEM) Cross-Entropy loss, focusing on hard examples during training. Both models were independently trained on the source domain (Urban) and the target domain (Rural) datasets.

Results - The trained models were evaluated on both domains to assess their performance in classical semantic segmentation tasks for high-resolution remote sensing data.

Results in Table 1 and Table 2 show that both models exhibit a performance drop when tested on a domain different from the one they were trained on, indicating a domain shift between Urban and Rural environments. When trained on Urban data, both models experience a severe drop in Rural mIoU, while training on Rural data allows for relatively better adaptation to Urban settings. Furthermore PIDNet-S, trained on Rural domain, surprisingly results to perform even better with Urban environments, which suggests that the Rural domain contains more diverse or transferable features that generalize well to Urban settings. This could be attributed to the fact that Rural areas in the dataset include a mix of natural and artificial elements, whereas Urban areas predominantly contain artificial ones. In Table 3, the number of model parameters, floating point operations per second (FLOPs), and mean inference times are analyzed to compare the two architectures. The results indicate that PIDNet-S is a lightweight model well-suited for real-time applications, achieving an inference speed over eight times faster than DeepLabV2. Additionally, due to its significantly larger number of parameters, DeepLabV2 requires a greater number of training epochs to achieve meaningful performance in terms of mIoU. Under the given training conditions, PIDNet-S demonstrates to be more efficient.

Model	Images size	Parameters	FLOPs	Mean inference time	Mean FPS
DeepLabV2	512x512	43.016M	185G	128.658 ms	7.773 frames/s
PIDNet	512x512	7.718M	5.933G	9.304 ms	107.476 frames/s
PIDNet	1024x1024	7.718M	23.733G	15.658 ms	63.865 frames/s

Table 3. DeepLabV2 and PIDNet-S models comparison (latencies computed on Google Colab with Tesla T4 GPU).

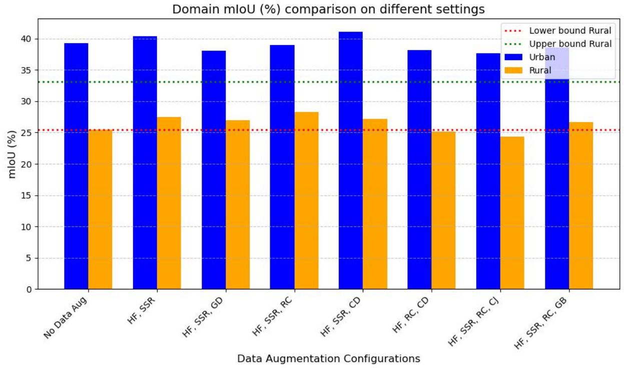

Figure 3. mIoU comparison between the two domains, using different configurations of data augmentation

3.1.2 Reducing Domain Shift with Data Augmentation

Implementation details - Taking the same PIDNet-S training settings used in the first experiment, a comprehensive data augmentation pipeline is applied to modify the visual appearance of source images and enhance the model's robustness to variations in lighting, texture, and structural patterns. Augmentation techniques were combined together in different settings to enhance the diversity of training data. The transformations involved are *Horizontal Flip* (HF), *Shift Scale Rotate* (SSR), *Grid Distortion* (GD), *Random Crop* (RC), *Brightness Contrast* (BC), *Coarse Dropout* (CD), *Color Jitter* (CJ), *Gaussian Blur* (GB). The probability to perform each transformation was set to 0.5.

Results - Data augmentation strategies improve generalization capability across diverse remote sensing environments, as evidenced by a consistent increase in mIoU scores when evaluated on both Urban and Rural domains. Figure 3 presents the results obtained by evaluating various data augmentation strategies on both Urban and Rural environments. Remarkable results are achieved by the configuration composed by HF, SSR and RC reaching the highest performance in the target domain (28.30% mIoU).

3.1.3 Adversarial Learning

Implementation Details – Inspired by [9], the proposed adversarial learning approach has been adapted for the PIDNet-S architecture by incorporating discriminator networks in two distinct configurations: single-level and multi-level.

In the single-level configuration, an additional discriminator network is introduced in the output space, aiming to directly align the predicted label distributions between the source and target domains. Moreover, the multi-level configuration extends adversarial learning across multiple feature levels, enabling adaptation at lower-level feature representations. This approach ensures that even early-stage features, which are distant from the final high-level output labels, undergo domain adaptation, thereby improving the overall robustness of the model across different domains. In multi-level configuration, the additional discriminator is attached to the existing output of the PIDNet-S from branch P, aiming not only to employ adversarial learning at different feature levels but also across different conceptual branches. This strategy tries to align both detailed and contextual information extracted from the source and target domains, enabling the model to adapt at both fine-grained and high-level feature representations.

The training setup for PIDNet-S follows the same configuration as in previous experiments, utilizing the OHEM Cross-Entropy loss function and an SGD optimizer with a momentum of 0.9, weight decay of 0.0005, and an initial learning rate of 0.001. The learning rate is progressively reduced using a polynomial decay strategy with an exponent of 0.9. In contrast, the additional discriminator networks are trained using the Binary Cross-Entropy (BCE) loss function and the Adam optimizer with β parameters set to (0.9, 0.99). The initial learning rate for the discriminator networks is set to 0.0005 and follows the same polynomial decay schedule with an exponent of 0.9.

Both the single-level and multi-level configurations are initially trained without data augmentation, followed by training with data augmentation applied.

Results - The results in Table 4 suggest that adversarial learning, in this case, may not effectively bridge the domain gap and it appears to have introduced more drawbacks than benefits. The decline in performance when applying adversarial learning suggests that the model struggles to balance the adversarial objective. This highlights that PIDNet faces difficulties in effectively aligning the feature space between the source and target domains, making it harder to generalize across domain shifts.

Mode	Data augmentation	mIoU (%) Urban	mIoU (%) Rural
Single level	-	34.15	21.56
Single level	HF, SSR, RC	34.19	20.14
Multi level	-	33.68	17.74
Multi level	HF, SSR, RC	35.41	20.59

Table 4. Adversarial training results on Urban and Rural domains.

3.1.4 DACS

Implementation details – This implementations relies on the same PIDNet-S training settings used in the first experiment. The PIDNet-S model is trained over both source domain images and mixed images between the two domains, employing the actual image-to-image strategy.

Domain alignment is performed, as suggested in [8], with ClassMix mixing strategy [6] by randomly choosing one half of the classes present in the source image, keeping all the pixels which belong to those classes and replacing remaining ones with the corresponding target image pixels.

The same transformation is applied also on labels, overlapping source labels with computed target preudo-labels. In order to compute pseudo-labels, target images are processed by an Exponential Moving Average (EMA) model. This technique is employed to stabilize the training process and enhance the generalization ability of the model by maintaining a smoothed version of the model's weights.

Only pseudo-labels with probabilities greater than a fixed threshold λ, chosen as 0.968, are used for training, and they are known as high-quality pseudo-labels. Data augmentation transformations were applied first only on mixed images and then on both source and mixed images with a probability of 0.5.

Results - Table 5 clearly indicates that applying DACS to PIDNet-S does not yield performance improvements over the lower-bound mIoU. This outcome may be attributed to several factors. First, the quality of the pseudo-labels generated during the mixing process may have been insufficient, leading to unreliable supervision. Additionally, class conflation, where distinct classes become confused due to visual similarities, may have hindered the effectiveness of the adaptation, as mixed samples could introduce ambiguous class boundaries. Finally, the lightweight nature of PIDNet-S may limit its capacity to capture complex cross-domain feature representations.

In the following section, we tried to improve the performances of DACS trying to face some intrinsic problems.

Data augmentation	mIoU (%) Urban	mIoU (%) Rural
-	35.50	17.79
RC-ALL	35.26	20.83
RC-ALL, CJ-MXD, GB-MXD	36.20	21.48
RC-ALL, HF-MXD, SSR-MXD	35.82	21.82
RC-ALL, CJ-ALL, GB-ALL	31.39	17.54
RC-ALL, HF-ALL, SSR-ALL	33.19	19.72

Table 5. DACS training results on Urban and Rural domains.
”ALL”: augmentation performed on both source and mixed data.
”MXD”: augmentation performed only on mixed data

3.1.5 DACS with GCW and LDQ

Implementation details - To address some limits within the DACS framework, we introduced two key modifications: Gradual Class Weights (GCW) to enhance the representation of minority classes during training and so overcome the class imbalance problem, and Local Dynamic Quality (LDQ) to improve the reliability of pseudo-labels, basing on local information. In our experiments, we first conducted trials with fixed GCW suggested parameters T set to 0.1 and β set to 0.99. Trainings were conducted without applying extensive data augmentation, except for Random Crop (RC), which was consistently applied to images from both domains. Subsequently, we incorporated Horizontal Flip (HF) and Shift Scale Rotate (SSR), applying these augmentations exclusively to the mixed images. Furthermore, hyperparameter tuning was performed on the GCW strategy to determine the optimal parameter set, represented by T temperature parameter and β mixing parameter, that enhances the mIoU of underrepresented classes.

Results - The results obtained from integrating Gradual Class Weights (GCW) and Local Dynamic Quality (LDQ) are summarized in Table 6 and Table 7. Figure 4 shows the behaviour of Gradual Class Weights during training, which, after a warmup phase, result to be consistent with the LoveDA dataset statistics. Overall, the best model performance, achieving a mean Intersection over Union (mIoU) of 20.03%, does not overcome the performance of the original DACS strategy. However, as shown in Figure 5, the IoU for underrepresented classes, such as Barren, Forest, and Agriculture, is improved. This improvement, though, comes at the expense of the more represented classes, such as Background, Building, and Road, where the IoU has decreased. This trade-off highlights the challenge of balancing performance across classes with varying levels of representation, which is a key consideration in domain adaptation tasks.

Techniques	Data augmentation	mIoU (%) Urban	mIoU (%) Rural
GCW	RC	34.23	18.98
GCW	RC + MXD	34.13	19.11
LDQ	RC	33.44	17.44
LDQ	RC + MXD	32.34	16.04
GCW & LDQ	RC	32.81	17.42
GCW & LDQ	RC + MXD	34.11	17.48

Table 6. Results in training DACS with GCW and LDQ on Urban and Rural domains. Here ”MXD” stands for HF and SSR applied on the mixed images.

Techniques	T	β	mIoU (%) Urban	mIoU (%) Rural
GCW	0.1	0.9	34.63	18.73
GCW	0.3	0.9	33.29	17.36
GCW	0.6	0.9	34.45	17.51
GCW	0.9	0.9	34.25	18.90
GCW	0.1	0.99	36.10	19.81
GCW	0.3	0.99	34.52	20.03
GCW	0.6	0.99	35.42	19.48
GCW	0.9	0.99	33.47	19.19

Table 7. Hyperparameter tuning results for Gradual Class Weights (GCW) with different values of T and β.

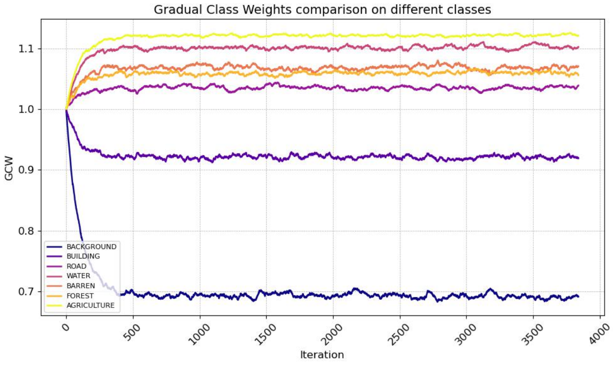

Figure 4. Behaviour of Gradual Class Weights during training.

4. Conclusions

To summarize, the objective was to benchmark various techniques aimed at mitigating the domain gap between urban and rural environments using the LoveDA HSR land-cover dataset in the semantic segmentation task, with a focus on the real-time PIDNet-S network.

The investigation encompassed data augmentation strategies and multiple domain adaptation approaches, including adversarial learning, self-learning via DACS, and an extended version of DACS designed to address some of its limitations.

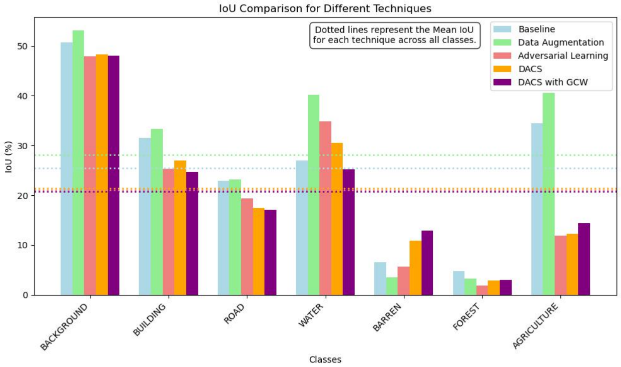

Figure 5. Comparison of IoU values per class for the top-performing models across the various techniques evaluated.

Figure 5 presents a detailed comparison of the Intersection over Union (IoU) per class, along with the mean IoU (mIoU), for the best-performing models across different domain adaptation strategies. This visualization highlights the strengths and weaknesses of each approach in handling specific semantic categories, providing insights into how different techniques impact both well-represented and underrepresented classes.

Figure 6 and Figure 7 shows qualitative results respectively testing on Urban and Rural images.

The results highlight that while PIDNet-S is optimized for efficient segmentation, it does not inherently accommodate domain shifts effectively. The model struggles to achieve robust feature space alignment between source and target domains, primarily due to its lightweight design, which constrains its ability to capture complex domain-specific variations necessary for successful adaptation. A larger model, such as DeepLabV2 with a ResNet101 backbone, when properly trained, could better satisfy domain adaptation requirements by leveraging its deeper architecture and greater capacity to learn transferable features across domains, without downsampling them excessively and preserving fine details. Atrous convolutions employed in Atrous Spatial Pyramid Pooling (ASPP) allow DeepLabV2 to capture multi-scale contextual information which is particularly powerful in Remote Sensing applications.

Future work could focus on exploring Domain Adaptation strategies suitable for lightweight architectures like PIDNet, as well as investigating alternative real-time architectures that achieve an optimal balance between performance, latency, and adaptability to domain shifts in remote sensing applications.

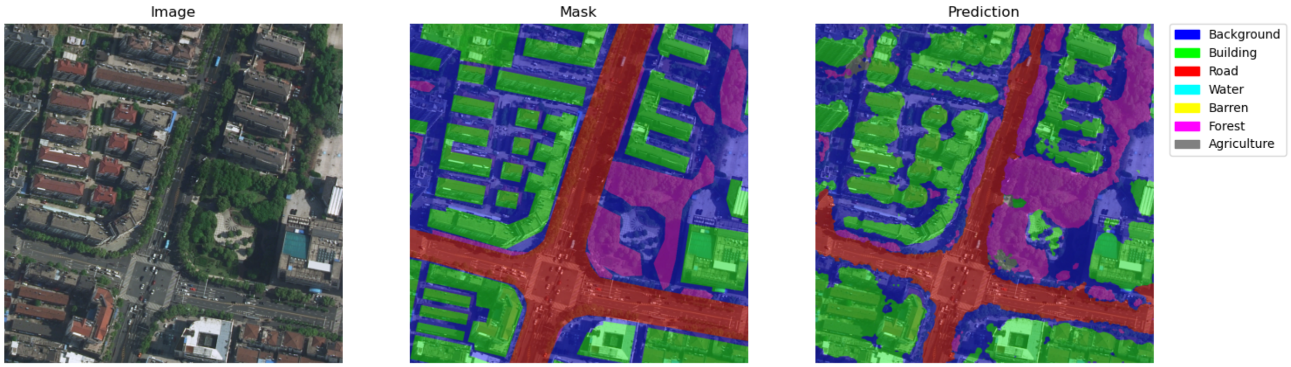

Figure 6. Labels prediction on Urban image using PIDNet-S trained on Urban domain.

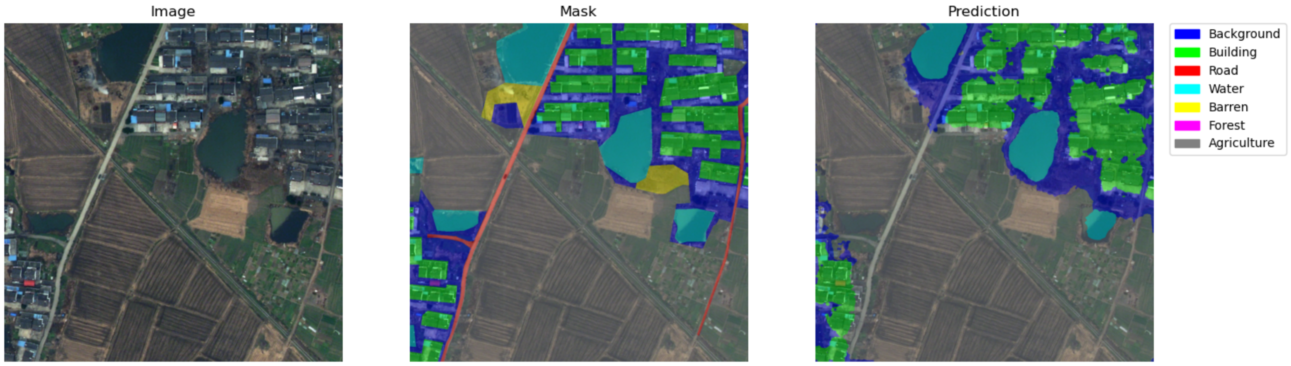

Figure 7. Labels prediction on Rural image using PIDNet-S trained on Urban domain.

References

[1] Erdem Akagunduz Irem Ulku. A survey on deep learning-based architectures for semantic segmentation on 2d images. 2022. 1

[2] Shankar P. Bhattacharyya Jiacong Xu, Zixiang Xiong. Pidnet: A real-time semantic segmentation network inspired by pid controllers. 2021. 2

[3] Ailong Ma Xiaoyan Lu Junjue Wang, Zhuo Zheng and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. 2022. 2

[4] Iasonas Kokkinos Kevin Murphy Liang-Chieh Chen, George Papandreou and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. 2017. 2

[5] Shanghang Zhang Bo Li Han Zhao Bichen Wu Ravi Krishna Joseph E. Gonzalez Alberto L. Sangiovanni-Vincentelli Sanjit A. Seshia Sicheng Zhao, Xiangyu Yue and Kurt Keutzer. A review of single-source deep unsupervised visual domain adaptation. 2020. 2

[6] Juliano Pinto Viktor Olsson, Wilhelm Tranheden and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. 2020. 6

[7] Yi Su Weitao Li, Hui Gao and Biffon Manyura Momanyi. Unsupervised domain adaptation for remote sensing semantic segmentation with transformer. 2022. 3

[8] Juliano Pinto Wilhelm Tranheden, Viktor Olsson and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. 2020. 3, 6

[9] Samuel Schulter Kihyuk Sohn Ming-Hsuan Yang Yi-Hsuan Tsai, Wei-Chih Hung and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. 2018. 2, 5