Real-time Domain Adaptation in Semantic Segmentation

Attrovio Mario, Ghisolfo Giorgia, Russo Michele

January 28, 2025

1 Abstract

Semantic segmentation is a critical task in computer vision, enabling pixel-wise classification of images. However, the performance of segmentation models often degrades when applied to data from different domains, a challenge known as domain shift. This report explores real-time semantic segmentation in the context of domain adaptation using PIDNet as the backbone. We investigate the performance drop caused by domain shift between urban and rural datasets and evaluate mitigation strategies, including data augmentation and advanced domain adaptation techniques like adversarial training and image-to-image translation (DACS). Experimental results on the LoveDA dataset demonstrate that these methods significantly reduce the impact of domain shift while maintaining real-time inference capabilities, achieving a balanced trade-off between accuracy and computational efficiency. The code can be found on our project website: `https://github.com/MichelePoli/AMLProject`.

2 Introduction

Semantic segmentation is a foundational task in computer vision, where each pixel in an image is assigned a label corresponding to a predefined class. It plays a vital role in applications such as autonomous driving, medical imaging, and remote sensing. Recent advancements in deep learning have yielded high-performing models, but these often struggle with domain shift—a phenomenon where a model trained on a source domain (e.g., urban images) performs poorly on a target domain (e.g., rural images) [1]. Addressing this challenge is crucial for real-world deployments where annotated data for all target domains is scarce or unavailable.

Domain adaptation aims to bridge this performance gap by aligning the source and target domains without requiring extensive labeled data from the target domain. While several methods exist, real-time semantic segmentation introduces additional constraints, such as maintaining high inference speed and low computational cost. PIDNet [2], a real-time segmentation network inspired by Proportional-Integral-Derivative (PID) controllers, serves as the backbone for our study due to its efficiency and accuracy in real-time tasks.

This report focuses on evaluating and improving the performance of PIDNet for domain-adaptive semantic segmentation using the LoveDA dataset [1]. We first quantify the performance degradation caused by domain shift. Next, we implement data augmentation techniques and two domain adaptation approaches—adversarial training and image-to-image translation (DACS)—to mitigate this issue. Our findings highlight the potential of these methods to enhance generalization while preserving the real-time capabilities of the model.

3 Related Work

Semantic segmentation has evolved significantly with deep learning. Fully Convolutional Networks (FCNs) [3] pioneered end-to-end segmentation by replacing fully connected layers with convolutional ones, enabling the network to process images of arbitrary sizes and produce pixel-wise predictions. DeepLabV2 [4] introduced atrous convolutions, which expand the receptive field without increasing the number of parameters, allowing for multi-scale context aggregation. Real-time networks like BiSeNet [5] optimized speed-accuracy trade-offs using lightweight backbones and parallel structures, while PIDNet [2] further improved efficiency by mimicking PID controllers to balance high-, mid-, and low-level features.

Domain adaptation techniques address performance degradation across domains. Adversarial methods [6] train a discriminator to distinguish between source and target domain features, encouraging the feature extractor to produce domain-invariant representations. Image-to-image translation approaches like DACS [7] blend domains through mixed sampling, generating pseudo-labeled target domain images to improve model generalization. The LoveDA dataset [1] provided urban/rural splits to benchmark these methods in remote sensing, offering a challenging scenario for domain adaptation research.

4 Methods

4.1 Baseline Model Development

4.1.1 2a Classic Segmentation Architecture

We implemented DeepLabV2 [4] with ResNet-101 backbone, pre-trained on ImageNet, using the LoveDA-urban dataset for training. This architecture leverages atrous spatial pyramid pooling to capture multi-scale contextual information through dilated convolutions. While effective for dense prediction tasks, its computational complexity makes it suboptimal for latency-sensitive applications. The model was initialized with ImageNet weights and trained and evaluated on urban scenes.

4.1.2 2b Efficient Segmentation Framework

To establish real-time capability baselines, we employed PIDNet-S - a streamlined architecture that mimics proportional-integral-derivative control mechanisms through three parallel branches. These branches explicitly manage high-frequency details, mid-level structures, and low-level spatial relationships respectively. The model was similarly initialized with ImageNet weights and trained and evaluated on urban scenes.

4.1.3 PIDNet Architecture Details

The core PIDNet architecture [2] consists of three branches

- **Proportional (P) Branch**: Processes full-resolution inputs using shallow layers to preserve spatial precision. Contains edge-aware convolutions for sharp boundary detection.
- **Integral (I) Branch**: Leverages deep layers with dilated convolutions and spatial pooling to capture multi-scale context. Integrates a Semantic Guidance Module (SGM) to filter low-level noise.
- **Derivative (D) Branch**: Computes high-frequency feature discrepancies between adjacent stages using depthwise separable convolutions. Acts as a boundary corrector by amplifying transitional regions.

The branches are fused through a *Three-Way Attention Fusion (TWAF)* module that dynamically combines features using spatial and channel attention weights.

4.2 3a Domain Adaptation

We investigated cross-domain generalization by evaluating our urban-trained PIDNet-S on the LoveDA-rural dataset without fine-tuning. This experimental design isolates the domain shift problem between man-made urban environments (characterized by geometric regularity and dense infrastructure) and rural landscapes (featuring organic shapes, sparse structures, and varied terrain).

4.2.1 Domain Adaptation challenges

The performance degradation observed could be caused by:

- 1) **Object Scale Variance**: Urban structures maintain relatively consistent scales, while rural elements exhibit greater size variability (trees, water bodies)
- 2) **Contextual Dependencies**: Urban scene semantics rely on positional relationships (e.g., roads between buildings), whereas rural contexts depend more on texture and color patterns
- 3) **Surface Color Distribution**: Artificial materials dominate urban areas (concrete, glass), contrasting with natural materials (soil, vegetation) prevalent in rural regions

This analysis motivates subsequent domain adaptation strategies to bridge the feature distribution gap between structurally distinct environments.

4.3 3b Data Augmentation

Data augmentation was used to improve the model's generalization capability by increasing the diversity of the training data. Two augmentation strategies were applied during training with a probability of 0.5:

- **A1**: Geometric transforms (horizontal/vertical flips, 30° rotation)
- **A2**: Photometric transforms (ColorJitter, GaussianBlur)

Photometric augmentations were more effective because they help the model become invariant to changes in lighting and color, which are common differences between the urban and rural domains.

4.4 Unsupervised Domain Adaptation

This method aligns domains through adversarial training on segmentation outputs [6], comprising three phases:

Phase 1: Source Domain training The segmentation network G processes source images I_s to produce softmax probability maps:

$$P_s = softmax(G(I_s)) \in R^{C \times W \times H} \quad (1)$$

where C denotes semantic classes. Training uses standard cross-entropy loss with source labels Y_s.

Phase 2: Adversarial Alignment The segmentation network now receives gradients to confuse D by:

- Maximizing discriminator uncertainty on P_t (domain label 1)
- Enforcing similar output distributions $P_s \approx P_t$

Phase 3: Discriminator Training A fully-convolutional discriminator D learns to distinguish:

- Source softmax outputs P_s (domain label 1)
- Target softmax outputs $P_t = softmax(G(I_t))$ (domain label 0)

This single-stage approach achieves domain invariance by directly matching the structured output distributions rather than intermediate features.

4.4.1 4b DACS

Domain Adaptation via Cross-domain Mixed Sampling (DACS) [7] addresses unsupervised domain adaptation (UDA) by blending data from the source and target domains through class-level mixing. The main aspects of DACS are summarized below:

- **Augmented Training Samples:** DACS creates hybrid samples by combining:
  - A source domain image, which comes with ground-truth labels.
  - A target domain image, which has pseudo-labels generated by the model.

  During this process, a subset of classes is selected from the source image using its semantic map, and the corresponding pixels are pasted onto the target image.
- **Composite Label Generation:** The labels for the mixed image are produced by merging:
  - Ground-truth labels from the source image.
  - Pseudo-labels predicted for the target image.

  This ensures that the model is trained on hybrid data containing both reliable source annotations and target pseudo-labels, encouraging feature invariance across domains.
- **Training Loss:** The overall loss combines:
  - Supervised learning on the source domain data.
  - Consistency regularization on the mixed-domain samples.

  Importantly, this does not require any annotations from the target domain.

4.4.2 Extensions

We explored three alternative real-time semantic segmentation architectures:

- **BiSeNetV1** [5]: Employs a dual-path design with a *spatial path* (high-resolution stream for fine details) and a *context path* (fast downsampling with global average pooling for long-range dependencies). The features are fused using a specialized Feature Fusion Module (FFM).
- **LinkNet** [8]: Uses a lightweight encoder-decoder structure with residual skip connections. The encoder reduces spatial dimensions through cascaded convolutional blocks, while the decoder employs transposed convolutions for upsampling, with direct additive links between corresponding encoder-decoder layers.
- **STDC n** [9]: Features a Short-Term Dense Concatenate backbone that progressively aggregates multi-scale features through dense connections in early stages, followed by a Context Path with Attention Refinement Modules (ARMs) to enhance contextual awareness.

These models emphasize distinct architectural strategies: BiSeNetV1's parallel multi-resolution processing, LinkNet's symmetrical skip connections, and STDC-Net's dense feature aggregation.

5 Experimental Results

5.1 LoveDA Dataset

The LoveDA dataset [1] is a high-resolution (0.3 m) remote sensing dataset designed for domain-adaptive semantic segmentation in urban and rural environments. It contains 5,987 images across three Chinese cities (Nanjing, Changzhou, Wuhan), annotated with seven classes: *background*, *building*, *road*, *water*, *barren*, *forest*, and *agriculture*. The dataset is explicitly divided into urban (2,522 images) and rural (3,465 images) domains to study domain shift challenges.

Key characteristics of LoveDA include:

- **Multi-scale Objects**: Urban scenes feature densely packed buildings and structured roads, while rural areas contain scattered agricultural plots and irregular water bodies. Buildings in urban regions exhibit larger size variance compared to rural regions (Figure **??**).
- **Complex Backgrounds**: The *background* class dominates both domains (Figure **??**), encompassing diverse elements like vehicles and undeveloped land, which introduces high intra-class variance.
- **Domain Shift**: Urban and rural domains exhibit divergent class distributions (e.g., urban has 32% buildings vs. rural's 8%) and spectral properties (lower variance in rural areas due to homogeneous landscapes).

For domain adaptation experiments, LoveDA provides two tasks:

1. **Urban → Rural**: Train on urban data (Qinhuai, Qixia, Jianghan, Gulou) and test on rural (Jiangning, Xinbei, Liyang).
2. **Rural → Urban**: Train on rural data (Pukou, Lishui, Gaochun, Jiangxia) and test on urban (Jiangye, Wuchang, Wujin).

The dataset's inherent challenges—scale variation, background complexity, and domain-specific class imbalances—make it a rigorous benchmark for evaluating real-time domain adaptation methods.

5.2 Performance on Source Domain

Table 5.2 shows the performance of DeepLabV2 and PIDNet-S on the LoveDA-urban dataset (source domain). PIDNet-S, designed for real-time performance, achieved a significantly higher mIoU than DeepLabV2 while also providing latency, FLOPs, and parameter count.

Model	mIoU (%)	Latency (ms)	FLOPs	Params
DeepLabV2	34.40	-	-	-
PIDNet-S	48.67	288.70	6.35G	7.72M

5.3 Domain Shift Evaluation

Table 5.3 quantifies the domain shift from LoveDA-urban to LoveDA-rural. The baseline PIDNet-S model, trained on urban data, experienced a significant performance drop when tested on rural data. Data augmentations (A1 and A2) improved the mIoU, with A2 (photometric transforms) being the most effective. Adversarial training and DACS further mitigated the domain shift, achieving similar performance.

Model	Road	Building	Water	Barren	Forest	Agric.	mIoU
PIDNet	16.34	23.99	35.46	3.12	8.83	31.82	23.98
+ A1	29.33	40.77	36.71	9.17	9.53	33.13	29.91
+ A2	31.41	38.37	31.33	10.26	15.10	37.51	30.80
+ A1 + A2	27.92	32.97	33.98	10.50	10.69	37.45	29.35
+ Adv.	0.36	13.41	32.72	8.28	49.73	12.26	30.59
+ DACS	31.32	36.60	42.70	4.49	2.70	40.60	30.63

5.4 Extensions

Table 5.4 presents the results of the extensions on the LoveDA-rural dataset. Style transfer preprocessing improved the mIoU slightly compared to the baseline but was less effective than data augmentation or domain adaptation techniques. BiSeNetV1 and LinkNet showed competitive performance, with BiSeNetV1 outperforming PIDNet-S on the target domain.

Model	mIoU (%)
PIDNet + Style Transfer	25.46
BiSeNetV1	32.21
LinkNet	30.43

5.5 Comparison with Original UDA Paper Results

Table 5.5 compares our domain adaptation results with those reported in the original DACS paper [7] and LoveDA benchmarks [1]. While DACS achieved 39.10% mIoU on LoveDA-rural in the original implementation, our PIDNet-S adaptation reached only 30.63%. Similarly, adversarial training underperformed compared to Tsai et al. [6] (30.59% vs. 35.20%). Three key factors explain these differences:

- **Model Architecture**: The original DACS paper used DeepLabV2 with ResNet-101, which has significantly higher capacity (44.5M params) compared to our real-time PIDNet-S (7.72M params). This architectural difference directly impacts feature representation power.
- **Training Constraints**: Our experiments used a fixed 20-epoch training schedule to maintain real-time deployment capabilities, whereas the original works employed longer training (50+ epochs) with extensive hyperparameter tuning.
- **Latency-Accuracy Trade-off**: PIDNet-S prioritizes inference speed (288ms) over pure accuracy, while DeepLabV2-based implementations ignore latency constraints (typically ¿1,000ms).

These results highlight the inherent challenge of balancing domain adaptation performance with real-time requirements—a critical consideration for edge deployment scenarios.

Method	Model	mIoU (%)	Latency (ms)
DACS (Original) [7]	DeepLabV2 + ResNet-101	39.10	1,200
DACS (Ours)	PIDNet-S	30.63	288
Adv. Training (Original) [6]	DeepLabV2 + VGG16	35.20	850
Adv. Training (Ours)	PIDNet-S	30.59	288

6 Conclusion

This work demonstrates PIDNet's effectiveness for real-time semantic segmentation on the LoveDA-urban dataset (288ms latency, 48.67% mIoU) and quantifies the domain shift effect when applied to LoveDA-rural (23.98% mIoU). Data augmentation, particularly photometric transforms (A2), provided the most consistent improvement (+6.82% mIoU), while adversarial training and DACS showed moderate gains (+6.61% and +6.65% mIoU, respectively). Extensions with alternative models highlighted PIDNet's superior speed-accuracy balance, although BiSeNetV1 performed slightly better on the target domain. Future work could explore hybrid adaptation strategies, optimized style transfer pipelines, and the combination of DACS and adversarial training for further performance enhancements.

References

[1] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation," *CoRR*, vol. abs/2110.08733, 2021. [Online]. Available: https://arxiv.org/abs/2110.08733

[2] J. Xu, Z. Xiong, and S. P. Bhattacharyya, "Pidnet: A real-time semantic segmentation network inspired by pid controllers," 2023. [Online]. Available: https://arxiv.org/abs/2206.02066

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

[4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[5] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," *Proceedings of the European conference on computer vision (ECCV)*, pp. 325–341, 2018.

[6] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7472–7481, 2018.

[7] W. Tranheden, V. Olsson, J. Pinto, and L. Svensson, "Dacs: Domain adaptation via cross-domain mixed sampling," *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1479–1489, 2021.

[8] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," *2017 IEEE Visual Communications and Image Processing (VCIP)*, pp. 1–4, 2017.

[9] M. Fan, S. Lai, J. Huang, X. Wei, Z. Chai, J. Luo, and X. Wei, "Rethinking bisenet for real-time semantic segmentation," *CoRR*, vol. abs/2104.13188, 2021. [Online]. Available: https://arxiv.org/abs/2104.13188