UNSUPERVISED DOMAIN ADAPTATION FOR AUTOMATIC ROAD EXTRACTION FROM AERIAL IMAGES

Gustavo Rota Collegio[1], *Antonio Gaudencio Filho*[2], *Aluir Porfírio Dal Poz*[3]

[1,3]Department of Cartography, Faculty of Sciences and Technology, São Paulo State University (UNESP), São Paulo 19060-900, Brazil - (gustavo.collegio, aluir.dal-poz)@unesp.br
[2]DSG, Brazilian Army Geographic Service, Brasília - DF, Brazil - guimaraesfilho.antonio@eb.mil.br

ABSTRACT

Automatic road extraction has been an active research topic for the past few decades, in which methods based on Deep Learning are state-of-the-art. However, to train those models, datasets containing a mass of dense annotations are a prerequisite and do not deal with the different scenarios that the road network has. To address these challenges, we propose an approach based on Unsupervised Domain Adaptation (UDA). Specifically, our UDA-based approach is evaluated on three strategies for Convolution Neural Networks: the U-Net architecture, the use of multi-level and multi-scale information (DeepLabv3+), and the attention mechanism in U-Net. The experiments show that among the strategies, the attention mechanism enhances the connectivity of the road network over some shortcomings, and its contribution against the baseline reaches an increase of 2% and 3% in F1-Score and Intersection over Union metrics, respectively.

Key words – *Road Extraction, Unsupervised Domain Adaptation, Deep Learning, Semantic Segmentation.*

1. INTRODUCTION

Automatic road extraction has been an active research topic for the past few decades. The main reason for its relevance is in order to get the road network updated since it is constantly in change. Moreover, the road network provides several applications for human activities, such as urban planning, economy (e.g., logistics), autonomous navigation, and others.

To keep road networks up-to-date across large-scale areas, remotely sensed data have become the primary source. High-resolution images, particularly, are the main type of data for obtaining information about the roads. In this source, road extraction methods are suitable for exploring the geometrical, radiometrical, topological, functional, scale, and contextual characteristics of the roads [1].

Recently, Convolutional Neural Networks (CNNs), in particular Fully Convolutional Networks (FCNs), opened up a new era for road extraction based on Deep Learning (DL), acquiring promising results and outperforming the handcrafted methods [1]. However, some challenges about the connectivity and topology of the roads still exist and, beyond that, to train those models, datasets containing a mass of dense annotations are a prerequisite. Since data labeling is extremely labor-intensive and usually unavailable, most of the state-of-the-art (SOTA) methods are commonly evaluated on benchmark datasets.

While current methods can achieve accurate results in this scenario, their effectivity might be constrained when applied to real-world practical applications, considering the highly complex context and the dynamism of the roads. For instance, the urban area is more dynamic than the rural area, as well as its context is completely heterogeneous, with different geometries of the roads, similar objects surrounding, many occlusions caused by buildings and vegetation, etc.

This limitation would be mitigated by training the network with the source data and then transferring the knowledge to the target data (e.g., an unknown test set). However, this supposes the availability of labeled data, which is not usual for practical applications. Unsupervised Domain Adaptation (UDA), to deal with this limitation, makes it possible to apply the trained labeled source data in the target data without labels delivering accurate results [2].

On the other hand, although UDA offers a viable approach to tackle the varying conditions of the road network, it remains challenging to handle specific shortcomings of the roads, such as occlusions and their discontinuity. According to the SOTA, exploiting contextual information on the roads, such as the spatial relationships and global and local features between objects in a scene plays a crucial role in addressing these challenges [3].

Based on the foundation work of the U-Net [4], which captures spatial hierarchies and recovers fine details through skip connections, other studies have been proposed to incorporate additional information, such as multi-scale and multi-level features, and the use of attention mechanisms [5]. However, as far as we know, no work has been published to compare how they perform based on the UDA approach for road extraction.

Therefore, in order to fill this gap, this work presents a comparison among the main approaches for UDA-based road extraction. Specifically, our UDA approach leverages the strategy from Domain Adversarial Neural Network [6] and is adapted to efficient but lightweight segmentation models, which are: U-Net (the pioneering architecture), DeepLabv3+ [7], and U-Net based on attention mechanism in order to filter the features propagated through skip connections [8].

The remaining of this work is organized as follows. Section 2 presents the dataset used and the proposed approach. Section 3 shows the experimental protocol and discusses the findings of the current study. Finally, Section 4 concludes the investigation.

2. PROPOSED APPROACH

2.1. Dataset

The proposed approach is evaluated on the benchmark dataset available by [9]. The set contains high-resolution

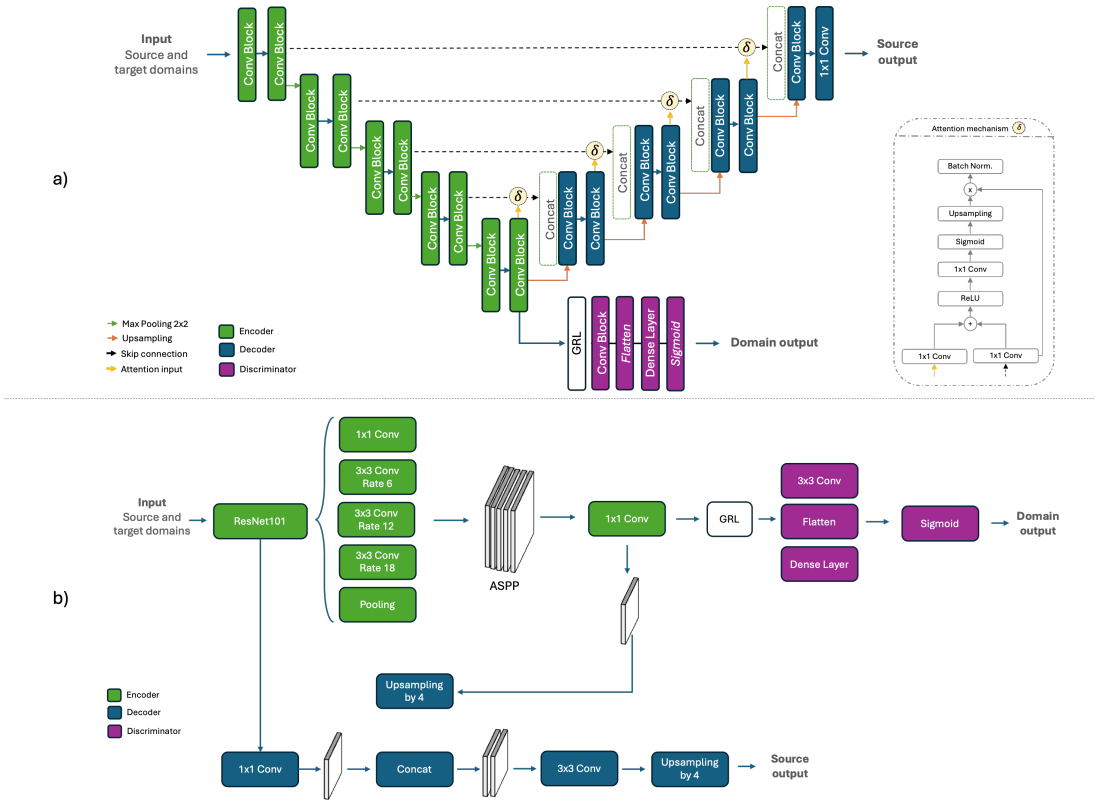


Figure 1: UDA-based model: a) U-Net with attention mechanism; and b) DeepLabv3+. GRL stands for Gradient Reversal Layer.

aerial images, whose road segmentation reference maps were manually labeled and the road width is about 12-15 pixels. The images have a Ground Sample Distance (GSD) of 1.2 m, radiometric resolution of 8 bits, and a total of 224 samples, with dimensions of at least 600x600 pixels.

This dataset was selected for several reasons: it includes different types of road pavements and presents regions with differing levels of complexity (e.g., residential areas whose roofs have similar radiometric characteristics to roads), including a few occlusion cases. In addition, it supports comparisons among diverse works in the SOTA.

Beyond those aspects, our unsupervised adaptation strategy also considers the adaptation of different types of data: from grayscale[1] (source domain) to RGB images (target domain). Even though the target domain labels are available, the appeal of UDA does not use them, except for evaluation.

Lastly, the pre-processing steps crop images into patches of 512x512 px, and, for training, the samples are padded with mirror mode. Data augmentation operations are also applied, such as random flips from left to right and from up to down, image rotation, and image transposing.

2.2. UDA-based models

The UDA-based models (i.e., U-Net, Attention U-Net, and DeepLabv3+) are implemented according to the Domain Adversarial Neural Network (DANN) [6], as depicted in Figure 1. Particularly, the network can minimize the divergence between two probability distributions (source and target data) parameterized by an encoder module. To achieve this, DANN is implemented with other two modules: a decoder, which acts as a supervised classifier for the source data, and a domain discrimination that aims to distinguish from which domain the input belongs. The three mentioned models have the same setting for UDA, and the implementation of them are detailed in the following.

The first one is the U-Net architecture, presented in Figure 1(a), but without the attention mechanism. We use the standard layers in the encoder (VGG-16 backbone [10]) and decoder, except by adding Batch Normalization. U-Net has the strategy of skipping connections to recover spatial information from different levels. However, in order to highlight salient features that are passed through them, we adopt the attention mechanism proposed by [8]. The adaptation for UDA is made prior to the concatenation steps to merge only relevant activations. As illustrated in Figure 1(a), this represents our second approach, Attention U-Net, allowing the model to filter out irrelevant or noisy responses within the skip connections.

The final implementation introduces the concept of multi-scale and multi-level features. DeepLabv3+ employs Atrous Spatial Pyramid Pooling (ASPP) in the encoder, which adjusts the filter's field-of-view in order to capture multi-scale information. Additionally, the decoder module focuses on recovering lower-level spatial information extracted from the earlier layers. To do so, the encoder incorporates the ResNet-101 [11] as the backbone before the ASPP to enhance the feature extraction. Although U-Net and DeepLabv3+ have different backbones (i.e., VGG-16 and ResNet-101, respectively), we maintain this configuration to support the inherent design of each model.

[1]The available RGB images converted into grayscale.

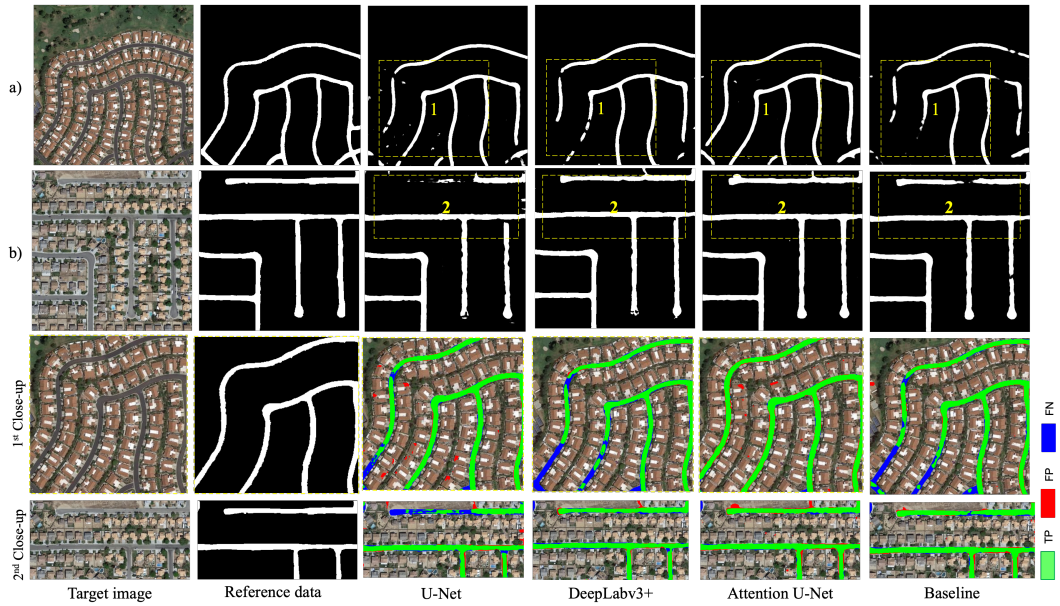


Figure 2: UDA predictions on the target domain. The 3rd and 4th rows correspond to the highlighted close-ups in a) and b). The last column corresponds to the Attention U-Net baseline. TP, FP, FN: true positives, false positives, and false negatives, respectively.

In all models, the discriminator is implemented at the feature level, receiving the deepest features from the encoder (indicated by purple blocks in Figure 1). Consequently, the forward propagation has two input batches: source and target domains. The encoder convolutes both batches, but only the source batch is passed to the decoder classification. The discriminator, with domain labels respectively of 0 and 1, receives both of them. In the backward pass, gradients of the decoder and the discriminator modules are calculated with respect to the same module: the encoder (θ_f weights). Thus, the total loss function is written as:

$$E(\theta_f, \theta_y, \theta_d) = \mathcal{L}_y(\theta_f, \theta_y) - \lambda \mathcal{L}_d(\theta_f, \theta_d) \quad (1)$$

where $\mathcal{L}_y$ is the loss for the decoder (θ_y weights); $\mathcal{L}_d$ is the loss for the domain discriminator (θ_d weights); and λ is a weighting scalar. This last negative term corresponds to the operation of the Gradient Reversal Layer (GRL), which acts as an identity mapping during the forward pass but reverses the gradient during the backward pass. In our approach, we implement GRL in the backward pass. Additionally, we adopt an alternate update strategy to stabilize the adversarial training process. Specifically, the discriminator is optimized separately using another learning rate and optimizer to address its faster convergence compared to the encoder-decoder modules.

3. EXPERIMENTAL ANALYSIS

3.1. Experimental protocol

The experiments were conducted with an NVIDIA RTX 3080 10GB graphics processing unit (GPU). The proposed approach was implemented using the TensorFlow framework (2.15 version). The loss function Focal Tversky Loss [12] which controls the classes' unbalancing of the roads was used in the decoder, and the Binary Cross-Entropy (BCE) for the domain discriminator. The encoder-decoder module was optimized using the Stochastic Gradient Descent (SGD) with a learning rate decaying exponentially from $1e{-}2$ up to $1e{-}4$. The discriminator was updated with the Adam optimizer with a learning rate of $1e{-}4$. The best performances were achieved by setting the term λ at 0.001 for both U-Net models and 0.0009 for DeepLabv3+ throughout the training.

To evaluate the contribution of UDA for road extraction, the UDA-based models are compared against a baseline and in fully supervised training. Both approaches adopt the same model settings but without UDA. The baseline is trained on the source domain and tested directly on the target domain, while the fully supervised training is trained and tested on the target domain. The metrics F1-Score and Intersection over Union (IoU) are calculated to carry this out.

3.2. Results and discussion

This section presents the analyses of the trained models' target predictions. As shown in Table 1, the metrics outperform the baseline for our three UDA-based models, but as expected, they have an inferior performance compared to fully supervised training. The U-Net-based model outperforms the baseline around 5% and 6% of F1-Score and IoU, respectively. DeepLabv3+, for each metric, increases them by 1%. The Attention U-Net is the approach that achieves the highest metrics for road extraction. For UDA, this approach contributes, respectively, around 2% and 3% for F1-Score and IoU against the baseline, and these metrics (88.66% and 79.87% of F1-Score and IoU) are superior to the supervised training on DeepLabv3+. Besides, the fully supervised training also outperforms the results reached by [9].

Qualitatively, Figure 2 presents the performance of UDA-based models. Specifically, Figure 2a) compared to Figure 2b) shows that the U-Net model increases the probability of

Model		F1-Score	IoU
U-Net	Baseline	80.02	67.36
	UDA	<u>85.04</u>	<u>74.31</u>
	Supervised*	**89.56**	**81.86**
DeepLabv3+	Baseline	84.3	72.58
	UDA	<u>84.41</u>	<u>73.57</u>
	Supervised	**85.79**	**76.17**
Attention U-Net	Baseline	86.25	76.39
	UDA	<u>88.66</u>	<u>79.87</u>
	Supervised	**89.97**	**81.98**

The best results: bold; the second-best results: underlined.
*U-Net modified by [9].

Table 1: Quantitative performance of predictions on the target domain (%).

false positives and DeepLabv3+ has more false negatives. The Attention U-Net, on the other hand, reduces the probability of noises in the predictions.

In detail, the close-ups from Figure 2 emphasize two complex cases of road extraction. The first case focuses on a residential area whose roofs of the houses have a radiometric aspect similar to the roads. While U-Net and DeepLabv3+ are more sensible, the Attention U-Net model is not. The second close-up shows a partial occlusion by trees over the roads. In this case, models that leverage scene context are the approaches that reach the best performance, that is, DeepLabv3+ and Attention U-Net. Specifically, both models better understand spatial relationships between objects, especially those surrounding the roads, such as trees.

Finally, the last column of Figure 2 compares UDA and the baseline for Attention U-Net, the UDA-based model that achieves the highest performance in the experiments. When dealing with similar radiometric information such as buildings and roads, only the grayscale information is not sufficient for testing in the target domain. However, incorporating target data information, the RGB information acts as an additional cue in the inference, improving the distinguishing among classes. This demonstrates how UDA enhances the capacity for generalization of the model.

4. CONCLUSIONS

This work aimed to investigate which strategy based on CNNs can improve the performance of UDA to contribute to road extraction, evaluating different data types and scenarios of the road network.

In this UDA case of grayscale-to-RGB, we demonstrated that information from the target domain, even unsupervised, can contribute as complementary information for the model, thereby distinguishing the classes from the roads more accurately. In addition, UDA approaches based on the strategy of scene context improve the geometric delineation of the roads, as well as their connectivity.

Despite this can be analyzed visually, the next step of this work aims to use other tools to provide a better comprehension of how this works in the predictions. Furthermore, other studies are suggested from the attention mechanism approach, such as the use of transformers and fusion with multimodal data.

5. ACKNOWLEDGEMENTS

This work was supported by the São Paulo Research Foundation (FAPESP) under Grant 2021/03586-2 and Grant 2024/02040-4.

6. REFERENCES

[1] Antonio Filho, Milton Shimabukuro, and Aluir Dal Poz. Deep learning multimodal fusion for road network extraction: Context and contour improvement. *IEEE Geoscience and Remote Sensing Letters*, 2023.

[2] Hongyu Chen, Hongyan Zhang, Guangyi Yang, Shengyang Li, and Liangpei Zhang. A mutual information domain adaptation network for remotely sensed semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 2022.

[3] Xianhong Zhu, Xiaohui Huang, Weijia Cao, Xiaofei Yang, Yunfei Zhou, and Shaokai Wang. Road extraction from remote sensing imagery with spatial attention based on swin transformer. *Remote Sensing*, 16(7), 2024.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv:1505.04597 [cs]*, sep 2015.

[5] Jinpeng Li, Jun He, Weijia Li, Jiabin Chen, and Jinhua Yu. Roadcorrector: A structure-aware road extraction method for road connectivity and topology correction. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–18, 2024.

[6] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-Adversarial Training of Neural Networks. *Advances in Computer Vision and Pattern Recognition*, 17(9783319583464):189–209, may 2015.

[7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation, 2018.

[8] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. 4 2018.

[9] Guangliang Cheng, Ying Wang, Shibiao Xu, Hongzhen Wang, Shiming Xiang, and Chunhong Pan. Automatic Road Detection and Centerline Extraction via Cascaded End-to-End Convolutional Neural Network. *IEEE Transactions on Geoscience and Remote Sensing*, 55(6):3322–3337, jun 2017.

[10] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, sep 2014.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem:770–778, dec 2015.

[12] Nabila Abraham and Naimul Mefraz Khan. A Novel Focal Tversky loss function with improved Attention U-Net for lesion segmentation. *Proceedings - International Symposium on Biomedical Imaging*, 2019-April:683–687, oct 2018.