Bridging the Synthetic-Real Gap: Unsupervised Domain Adaptation for Cross-Domain Image Segmentation

N Abhishek, Akshay Poojary, Varsha Sajjanavar, Apeksha Tangod, Sneha Varur and Channabasappa Muttal

Department of SOCSE, KLE Technological University Hubballi, Karnataka, India

Keywords: Unsupervised Domain Adaptation (UDA), Cross-Domain Image Segmentation, DeeplabV3+, ResNet-101, Synthetic-to-Real Adaptation, Autonomous Driving.

Abstract: It is a challenging problem for cross-domain image segmentation bridging the gap between synthetic and real worlds, which is very relevant given applications in autonomous driving scenarios. This work proposes an effective strategy for solving the problem in unsupervised domain adaptation for cross-domain image segmentation; training the model on the GTA5 dataset and testing it on the Cityscapes. We used the ResNet-101 backbone with DeeplabV3+ and exploited its encoder for feature extraction and an upsampling decoder for effective segmentation. The results show that the approach is quite robust for dealing with domain shifts. Although a domain gap exists between the synthetic and real datasets, it correctly segments complex urban scenes. This work makes segmentation models more accurate and generalizable in real applications by using synthetic training data within an unsupervised learning framework. The two major metrics used to evaluate the work are IoU and mean IoU (mIoU). Our method reached a mIoU of 55.80%, outperforming most state-of-the-art UDA methods for the cross-domain segmentation task.

1 INTRODUCTION

Unsupervised Domain Adaptation (UDA) for Cross-domain image segmentation is a crucial area in the computer vision world, where labeled data are rare or expensive. This problem is becoming more noticeable in regions, such as medical imaging and satellite image analysis, especially when referring to self-driving cars since their data comes from various domains - source and target thus likely with varying appearance, scale, or even texture.(Hoffman et al., 2018) Unsupervised domain adaptation for cross-domain image segmentation refers to the adaptation of a segmentation model, which is trained on labeled data from a source domain, such that it works effectively on an unlabeled target domain without any ground truth available for it.

Traditionally, Image segmentation relies highly on annotated data. However, obtaining labeled datasets for every domain of interest is often infeasible. Thus, UDA has been one of the promising approaches in leveraging labeled data from the source domain and adapting it to an unlabeled target domain, therefore saving the expensive annotations. Current approaches in UDA for image segmentation learn domain-invariant features that can be generalized across domains.(Tsai et al., 2018)

Common application methods include adversarial training with a discriminator used here, which tries to classify features coming from both source and target domains according to whether they belong to specific spaces while the extractor tries not to be detectable as features are constructed to contain rich representation-independent of origin in the original space. There have been positive responses applying adversarial learning between these spaces (Tzeng et al., 2017). Some application methods demonstrated the potential of cycle-consistent adversarial domain adaptation (CyCADA) in bridging the domain gap, particularly in scenarios involving significant visual differences.(Hoffman et al., 2018) Further improvements in such methods include the application of cycle-consistency losses, where it is possible to transform back from the target domain to the source domain to further improve the adaptability of the segmentation model (Zhou et al., 2019). These methods reduce the discrepancy in the distribution of domains and lead to improved performance in the target do-

Abhishek, N., Poojary, A., Sajjanavar, V., Tangod, A., Varur, S. and Muttal, C.
Bridging the Synthetic-Real Gap: Unsupervised Domain Adaptation for Cross-Domain Image Segmentation.
DOI: 10.5220/0013613500004664

In *Proceedings of the 3rd International Conference on Futuristic Technology (INCOFT 2025) - Volume 3*, pages 254-262
ISBN: 978-989-758-763-4

main with better segmentation.

Semantic consistency is the other critical application role for cross-domain segmentation based on UDA. One way to guarantee that this model knows the context of a segmented region in the target domain is by aligning higher semantic features across different domains.

Some further recent works included self-training strategies that produce pseudo-labels during training in the target domain and iterated throughout. The works explored learning from synthetic data and generating pseudo-labels, addressing domain shift challenges (Sankaranarayanan et al., 2018). It is anticipated that using the target domain's unlabeled data will improve the robustness of the segmentation model and achieve performance improvement without manual labeling. Although good progress is being made, the problem remains with domain shift and, more specifically, where domains are quite different to overcome (Tzeng et al., 2017). Soon, this area of UDA for image segmentation offers greater scope in feature alignment and multi-modality to better cope with the real world's complexity (Zhang et al., 2020a).

The objectives of our work are:

- To utilize an encoder-decoder framework with ResNet-101 as the backbone and integrate DeeplabV3+ for feature extraction and segmentation.
- To attain a good mIoU score that outperforms some existing state-of-the-art methods in the "GTA5 to Cityscapes" domain adaptation task.

The paper is structured as follows: Section 1 introduces the motivation and objectives behind leveraging unsupervised domain adaptation (UDA) to tackle the challenges of cross-domain image segmentation. Section 2 reviews existing approaches like adversarial learning and transformer-based architectures, highlighting their limitations. Section 3 details the methodology, including ResNet-101 with DeeplabV3+ architecture and adaptation strategies. Section 4 presents experimental results on GTA5 and Cityscapes datasets using IoU and mIoU metrics. Section 5 concludes with findings and future directions.

2 LITERATURE SURVEY

Several methods address domain shift issues, including adversarial learning, self-supervised learning, transformer-based architectures (Xu et al., 2021), and synthetic-to-real domain adaptation techniques. Among these, adversarial learning stands out with generative adversarial networks (GANs) being used to align feature distributions between the source and target domains. CycleGAN (Zhu et al., 2017), as shown in Fig. 5, employs cycle-consistent adversarial networks for unpaired image-to-image translation, mitigating discrepancies in appearance, such as color, texture, and illumination.

Building upon this, Volpi et al. (Volpi et al., 2018) proposed the MCD (Minimum Class Discrepancy) method, which uses adversarial learning to minimize class discrepancies between domains. By aligning the class-wise predictions of a segmentation model across both domains, MCD improves pixel-wise consistency and enables robust transferability between source and target domains. This method is particularly effective when using synthetic datasets like GTA5 and real-world datasets like Cityscapes, where visual characteristics differ significantly.

Self-supervised learning also plays a critical role by helping learn domain-invariant features. Bousmalis et al. (Bousmalis et al., 2016) introduced domain separation networks to separate domain-specific features from domain-invariant ones. This technique ensures that only relevant features for segmentation are learned, regardless of domain-specific variations. Similarly, Chen et al. (Chen et al., 2020) proposed a contrastive learning framework for domain adaptation that maximizes intra-domain similarity while minimizing inter-domain similarity, enabling robust feature learning.

The DAFormer framework, introduced by Xu et al. (Xu et al., 2021), as shown in Fig. 6, represents a significant leap in cross-domain segmentation. Using a transformer-based architecture, DAFormer leverages self-attention mechanisms to capture both local and global contextual information. These capabilities are especially useful for segmenting complex urban environments, as seen in datasets like Cityscapes. DAFormer's ability to focus on underrepresented classes and refine pseudo-labels during training improves its adaptability to target domain data.

For synthetic-to-real domain adaptation, Tsai et al. (Tsai et al., 2018) introduced AdaptSegNet, combining image-level and feature-level adaptation through adversarial learning. By aligning the structured output predictions of segmentation models, AdaptSegNet reduces domain discrepancies, ensuring that the model learns robust and domain-invariant representations.

Multi-task learning (MTL) has also gained popularity for improving cross-domain segmentation. Zhang et al. (Zhang et al., 2020b) leveraged MTL to optimize segmentation and related tasks, such as

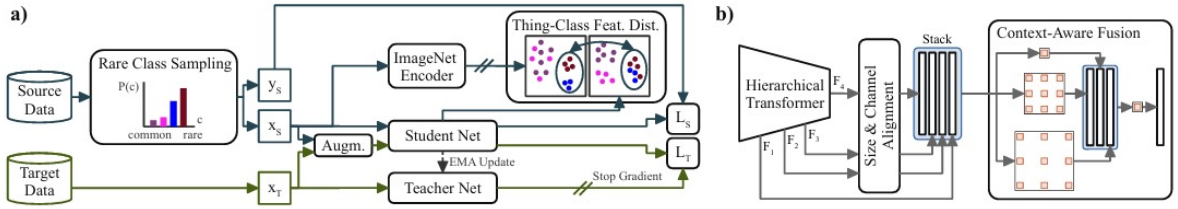


Figure 1: Cross domain adaptation using DAformer framework (Xu et al., 2021)

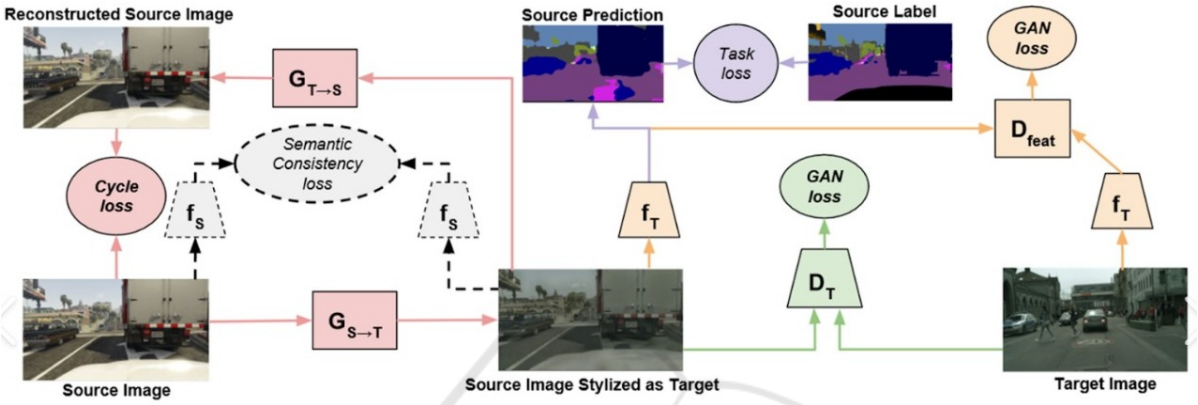


Figure 2: Cross domain adaptation using CycleGAN(Zhu et al., 2017)

object detection or depth estimation, allowing for better generalization across domains. In scenarios such as night-time segmentation or low-light environments, specialized architectures are necessary to handle domain-specific challenges. SEANet, introduced by Zhang et al. (Zhang et al., 2020c), employs a squeeze-enhanced axial attention mechanism to focus on critical spatial features under low-light conditions, enabling better performance in nighttime segmentation tasks. . The literature work points to some gaps such as difficulty in aligning features between source and target domains, minimizing class discrepancies, and learning domain-invariant features. It also highlights issues with segmenting complex environments, adapting to dynamic changes, and making use of multi-modal data. Our approach tackles these by using an encoder-decoder structure with ResNet-101 and DeeplabV3+. This helps to align features better, reduce class discrepancies, and improve segmentation accuracy across domains. The architecture also enables robust learning of domain-invariant features, which enhances performance in complex urban environments. While it doesn't fully address dynamic domain shifts, the model is flexible and could be extended in the future to incorporate multi-modal data, further improving domain alignment.

3 BACKGROUND STUDY

3.1 Resnet-101

ResNet-101 is a deep convolutional neural network (CNN) designed to address the vanishing gradient problem in very deep networks by using residual connections, or shortcuts, that bypass one or more layers. These connections allow smoother gradient flow during training, making it possible to train very deep networks without losing performance. The architecture consists of 101 layers, with residual blocks where the input is added to the output, helping preserve feature identity. This design prevents performance degradation as the network depth increases, making ResNet-101 effective for tasks like classification, detection, and segmentation. In our research, it serves as the backbone for DeeplabV3+, providing hierarchical features crucial for high-quality segmentation.

3.2 Deeplabv3+

DeeplabV3+ is an advanced model for semantic image segmentation, designed to excel in challenging environments like urban scenes. It builds on earlier versions with several key improvements. At its core, DeeplabV3+ uses a fully convolutional network

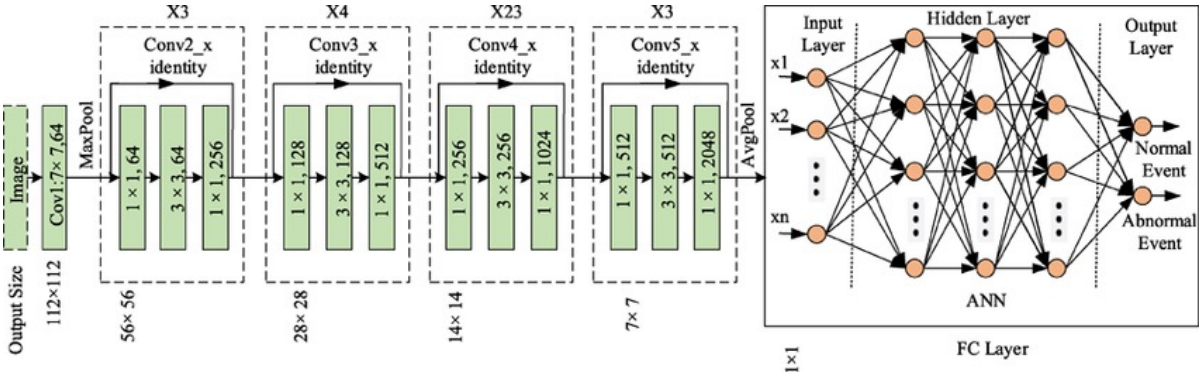


Figure 3: ResNet-101 architecture (Zhang et al., 2020a)

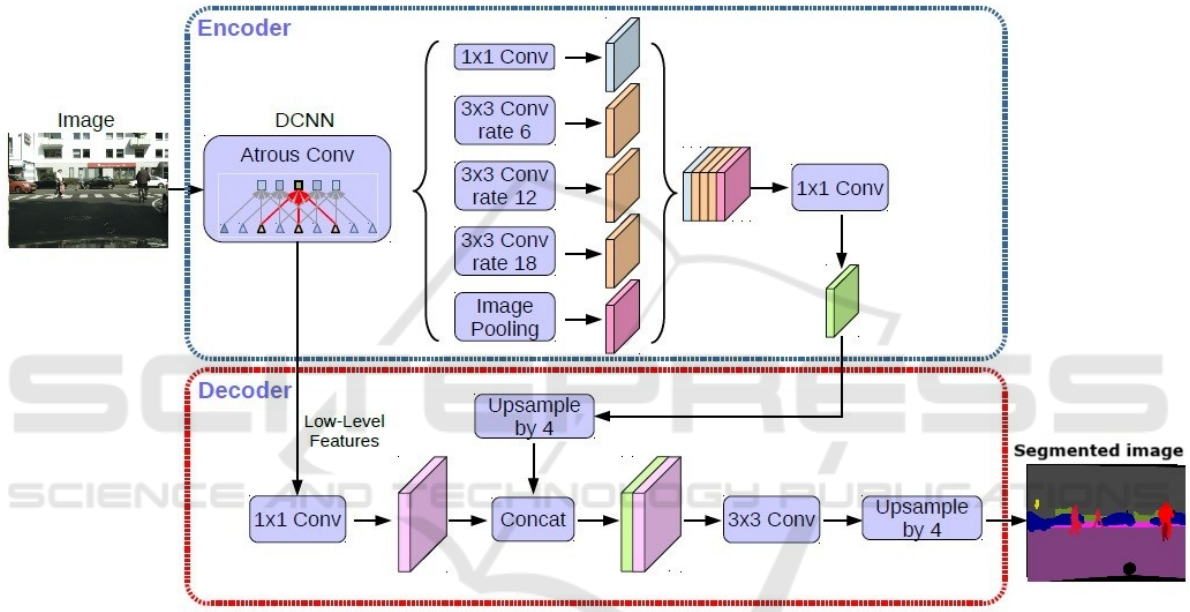


Figure 4: Deeplabv3+ architecture

(FCN) and employs Atrous Convolutions (also called Dilated Convolutions). These help the model capture context at multiple scales by expanding its field of view without adding extra parameters, making it better at identifying objects of different sizes. One of its standout features is the Atrous Spatial Pyramid Pooling (ASPP) module, which uses convolutions with different dilation rates to gather features at various scales. This allows the model to understand both the bigger picture and the finer details. DeeplabV3+ takes it a step further by adding a decoder module, which refines the segmentation results using information from earlier network layers. This added step ensures more accurate and detailed segmentations, making it a powerful tool for a wide range of tasks.

3.3 Integration of ResNet-101 and DeeplabV3+

In our research, combining ResNet-101 for feature extraction with DeeplabV3+ for segmentation allows us to capitalize on the strengths of both architectures. ResNet-101's deep residual learning captures rich, hierarchical features, while DeeplabV3+'s dilated convolutions and ASPP module significantly improve segmentation, especially when dealing with the multi-scale nature of complex scenes. This synergy ensures that your model delivers accurate and robust segmentation results, even when faced with domain shifts between synthetic and real-world datasets.

4 METHODOLOGY

4.1 Proposed Pipeline architecture

Our proposed pipeline architecture for unsupervised cross-domain image segmentation is based on pre-trained DeepLabv3+ as shown in Fig.4, Which incorporates several innovative enhancements to better segmentation and adaptation across domains. Our model starts with an input preprocessing stage where RGB images of size 512×512 are normalized and passed through to an augmentation module, which does domain-specific transformations including color jitter, Gaussian blur, random cropping, and noise injection to make the network robust against domain shifts, which occurs during training

The proposed model leverages pre-trained ResNet-101 3 as its backbone, demonstrating the ability to extract very rich hierarchical features. Using atrous convolutions with dilation rates 6, 12, and 18 effectively captures context at multiple scales. Another important component is the adaptive atrous layers, which dynamically adjust dilation rates to align with domain-specific input statistics, aiming to achieve domain robustness. Furthermore, a global context module is integrated to enhance feature extraction capabilities. This module aggregates domain-invariant global features using enhanced global average pooling with learnable weighting, focusing on the most important channels of the feature maps.

The Atrous Spatial Pyramid Pooling module is central to multi-scale feature extraction. Improvements are introduced with dynamic dilation rates tailored to the target domain and a channel attention mechanism that selectively focuses on relevant feature maps while suppressing irrelevant information. This mechanism effectively accounts for variations in object scales and domain characteristics.

The decoder is designed to restore spatial resolution by incorporating semantic information. The encoder combines high-level semantic features with improved skip connections and domain-specific attention layers, using low-level features extracted from earlier layers of the encoder. These features are further refined with 1×1 convolutions. The decoder progressively upsamples feature maps using bilinear interpolation and convolutional layers to recover fine spatial details, ultimately generating accurate segmentation maps at the input resolution.

Domain adaptation techniques are employed to handle domain shifts. A domain-specific batch normalization layer computes its statistics dynamically at training time to adapt to the characteristics of the target domain. Moreover, an auxiliary domain discriminator network introduces an adversarial loss penalizing large domain discrepancies to align the feature distributions of the source and target domains.

Resnet-101 involves an iterative quality improvement in a pseudo-label self-training loop; thereby, it exploits the model with high-confidence predictions to its greater advantage. Furthermore, the method is trained by multiple resolutions, since the resolution dealt with differs for different domains. The final classifier produces a semantic segmentation map of 19 distinct classes, which is optimized using both cross-entropy losses on the source domain's data and domain alignment loss such that the synthetic-real gap is bridged. Semantic segmentation requires classifying each pixel of an image into one of the C semantic classes. The cross-entropy loss computes the difference between what the model predicts as the class probabilities for each pixel and the ground truth labels for each pixel. The model outputs a tensor of shape [B,C,H,W] where B: Batch size, C: Number of classes, H,W: Spatial dimensions of the image. Ground Truth Labels (Y) as a tensor of shape [B,C,W] where each pixel is assigned an integer label $0 \leq l < C$, denoting the class. Predicted probabilities for each class at pixel i, obtained by applying the softmax function.

$$p_{i,c} = \frac{\exp(\hat{y}_{i,c})}{\sum_{c=1}^{C} \exp(\hat{y}_{i,c})} \tag{1}$$

For each pixel i, the cross-entropy loss compares the predicted probability distribution across classes to the true label and is given by:

$$\ell_i = -\sum_{c=1}^{C} y_{i,c} \cdot \log(p_{i,c}) \tag{2}$$

The aggregated cross-entropy loss over all N pixels in the image is typically averaged and defined as:

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \cdot \log(p_{i,c}) \tag{3}$$

Where:

1. N: Total number of pixels in the input batch, computed as $N = B \times H \times W$

4.2 Implementation Details

Hardware and Framework: The model is implemented in TensorFlow and trained on a GPU with 48 GB of NVIDIA L40S (Lightning AI). The training utilizes a learning rate of 0.001 with exponential decay, the Adam optimizer, and cross-entropy loss for image segmentation tasks.

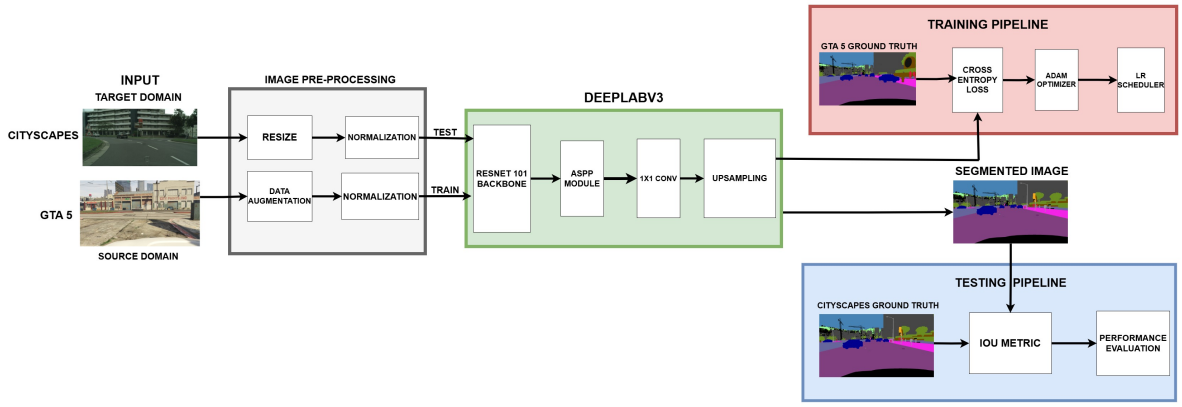


Figure 5: Proposed pipeline architecture

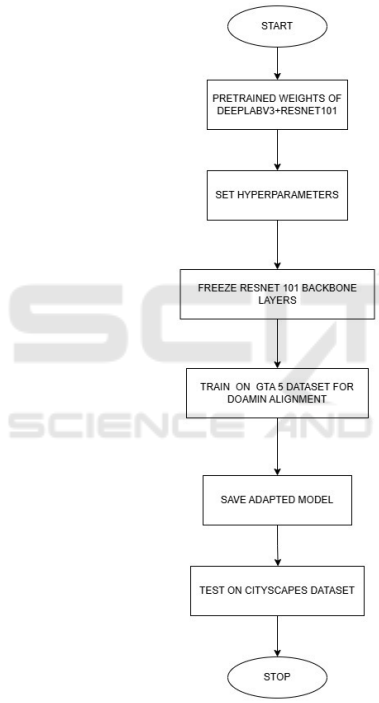


Figure 6: schematic representation of system model

The input image is downsampled to 1280x640 and randomly cropped to 512x512 for training. The discriminator also uses the Adam optimizer, with a learning rate set to 0.0001. Both the segmentation model and the discriminator use polynomial learning rate decay by multiplying the learning rate by a factor

$$\left(1 - \frac{\text{iter}}{\text{total iter}}\right)^{0.9} \tag{4}$$

. We define the total iteration as Run for 1000 iterations with an early-stop policy.

5 RESULTS AND ANALYSIS

5.1 Dataset description

There are two datasets used in our work, One is the training dataset which is the GTA5 dataset and another is the testing dataset which is the CITYSCAPES dataset.

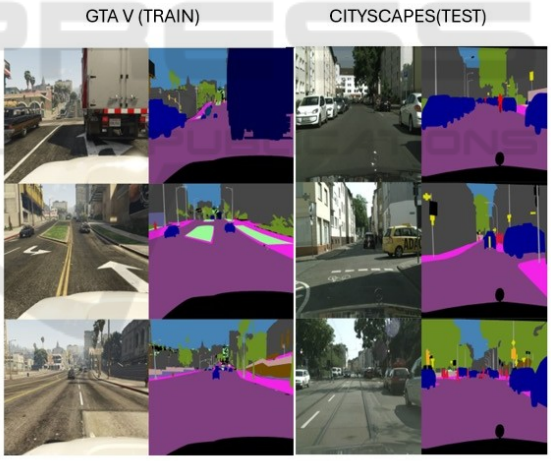


Figure 7: Training and testing datasets

GTA5 Dataset(Richter et al., 2016): This is a synthetic dataset from the GTA5 game with 24,966 high-resolution images (1914×1052 pixels). It consists of pixel-perfect annotations of 19 semantic classes that reproduce the real-world urban scene, such as roads, vehicles, and pedestrians. Its virtual environment ensures that it is a large, cost-effective, and diverse data source, but it creates a domain gap when applied to real-world tasks.

Cityscapes Dataset(Cordts et al., 2016): Real-world dataset of 5,000 images divided into the train

set (2,975), validation set (500), and test set (1,525). Captured in German cities with a consistent resolution of 2048×1024 pixels. It offers dense annotations for 19 semantic classes, with a focus on urban landscapes and high-quality benchmarks for real-world applications.

These datasets have been combined to allow the researcher to study these challenges of domain adaptation. The models are first trained on synthetic data from the GTA5 source domain and then tested on real-world data from the Cityscapes dataset., which is considered a target domain. This approach here further pushes efforts made for unsupervised domain adaptation techniques.

5.2 Evaluation

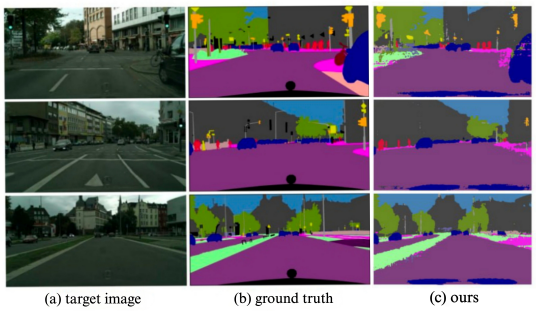
This section assesses the performance of the proposed system using standard statistical metrics. To gauge the system's effectiveness, we conducted a thorough comparison of per-class Intersection over Union (IoU) scores and the mean IoU (mIoU) percentage for the "GTA5 to Cityscapes" domain adaptation task. These metrics are essential for measuring how well our model adapts from the synthetic GTA5 dataset to the real-world Cityscapes dataset, ensuring it performs well across diverse domains. Our evaluation highlights the system's ability to generalize across domains, with detailed analysis showcasing performance across individual semantic classes. The per-class IoU scores provide insights into the system's strengths and limitations for specific object categories, while the overall mIoU serves as a robust metric for overall segmentation quality. These results underline the model's capability to achieve competitive cross-domain segmentation performance.

We now present a comparison of the performance of various methods on the semantic segmentation task for the "GTAV to Cityscapes" dataset. The methods are evaluated based on per-class IoU scores for individual classes, as well as the overall mean Intersection over Union (mIoU), which combines the IoUs across all classes.

The key takeaways from the results are summarized as follows:

Overall Performance (mIoU):

The method Ours achieved the highest mIoU score at **55.80** compared to all the methods in the table 1. Thus, it proves that Our approach outperforms several other state-of-the-art techniques, namely DPR (Ding et al., 2019) with 46.5, DISE (Zhao et al., 2019) with 45.4, and AdvEnt (Vu et al., 2019a) with 45.5, proving the efficiency of our model concerning adaptation to the "GTAV to Cityscapes" domain shift. Our



(a) target image (b) ground truth (c) ours

Figure 8: Results showing target image, Ground truth and segmented image

model outperforms several of the existing approaches, including SIBAN (Sibi et al., 2019) which reported mIoUs of 42.6.

Class-wise IoU Analysis:

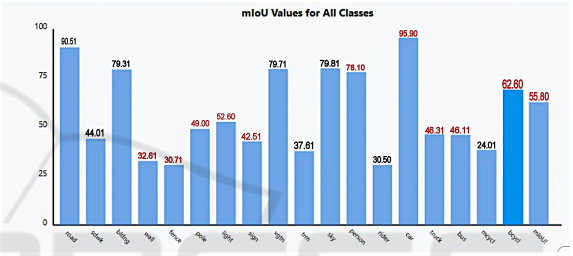


Figure 9: Graph showing class-wise IoU values and mIoU

High-performing classes: In some classes, our model works well. For example, IoU for the car class is **95.90**, which is the highest in the table 1 and beats other methods. This means that our model has learned to segment cars in different contexts and domain shifts. Besides, our model achieved a pretty high bus IoU score of **46.31** compared to other methodologies and indicates robustness for the vehicle detection task, especially larger vehicles like buses. Person and truck classes: Our proposed method has a person IoU of 78.10, far ahead of most of the methods, thereby indicating the competency of our method to adequately segment humans. Truck IoU of 46.11 is also competitive and well shows the segmentation ability of large vehicles.

Comparison with state-of-the-art methods:

DPR (Ding et al., 2019), using ResNet-101 as a backbone, achieves a mIoU of 46.5 which is acceptable but seriously lags behind our method (9.3 percentage points). Methods DISE (Zhao et al., 2019) and AdvEnt (Vu et al., 2019a)] which achieved smaller mIoUs (45.4 and 45.5 respectively) also support the conclusion drawn here that our approach does better for this adaptation task. CLAN (Liu et al., 2020a) and SIBAN (Sibi et al., 2019) outperform some but still lag behind our method with mIoUs of

Table 1: **Comparison on "GTAV to Cityscapes" in terms of per-class IoUs and mIoU (%).**

Method	Base Model	road	sdwk	bldng	wall	fence	pole	light	sign	vegtn	trm	sky	person	rider	car	truck	bus	train	mcycl	bcycl	mIoU
DPR (Ding et al., 2019)	ResNet-101	92.3	51.9	82.1	29.2	25.1	24.5	33.8	33.0	82.4	32.8	82.2	58.6	27.2	84.3	33.4	46.3	2.2	29.5	32.3	46.5
SIBAN (Sibi et al., 2019)	ResNet-101	88.5	35.4	79.5	26.3	24.3	28.5	32.5	18.3	81.2	40.0	76.5	58.1	25.8	82.6	30.3	34.4	3.4	21.6	21.5	42.6
AdaptSeg (Tsai et al., 2018)	ResNet-101	86.5	36.0	79.9	23.4	23.3	23.9	35.2	14.8	83.4	33.3	75.6	58.6	27.6	73.7	32.5	35.4	3.9	30.1	28.1	42.4
CLAN (Liu et al., 2020a)	ResNet-101	87.0	27.1	79.6	27.3	23.3	28.3	35.5	24.2	83.6	27.4	74.2	58.6	28.0	76.2	33.1	36.7	6.7	31.9	31.4	43.2
DISE (Zhao et al., 2019)	ResNet-101	91.5	47.5	82.5	31.3	25.6	33.0	33.7	25.8	82.7	28.8	82.7	62.4	30.8	85.2	27.7	45.0	6.4	25.2	24.4	45.4
AdvEnt (Vu et al., 2019b)	ResNet-101	89.4	33.1	81.0	26.6	26.8	27.2	33.5	24.7	83.9	36.7	78.8	58.7	30.5	84.8	38.5	44.5	1.7	31.6	32.4	45.5
MSL (Liu et al., 2020b)	ResNet-101	89.4	43.0	82.1	30.5	21.3	30.3	34.7	24.0	85.3	39.4	78.2	63.0	22.9	84.6	36.4	43.0	5.5	34.7	33.5	46.4
DLOW (Shaban et al., 2018)	ResNet-101	87.1	33.5	80.5	24.5	13.2	29.8	29.5	26.6	82.6	26.7	81.8	55.9	25.3	78.0	33.5	38.7	6.0	22.9	34.5	42.3
Ours	ResNet-101	90.51	44.01	79.31	**32.61**	**30.71**	**49.00**	**52.60**	**42.51**	79.71	37.61	79.81	**78.10**	30.50	**95.90**	**46.31**	**46.11**	**38.40**	24.01	**62.60**	**55.80**

43.2 and 42.6, respectively, and relatively lower performance on specific classes like bus and truck.

6 CONCLUSION

We present a novel approach to unsupervised domain adaptation by leveraging the encoder-decoder framework with a memory-based regularization technique. Our method utilizes intra-domain knowledge to reduce uncertainty during model learning, without introducing additional parameters or external modules. By using the model itself as a memory module, we achieve an elegant and efficient regularization of the training process. Despite its simplicity, our approach complements existing methods and delivers competitive performance on two prominent synthetic-to-real benchmarks: GTA5 to Cityscapes.

Our results demonstrate that the proposed model effectively addresses challenges in domain adaptation, achieving robust segmentation performance by reducing the domain gap. The integration of memory-based regularization highlights the potential for leveraging inherent model properties to improve training stability and accuracy.

Future enhancements could focus on designing models that are inherently robust to environmental variations, such as changes in lighting, texture, and adverse conditions. Additionally, advancements in adversarial learning techniques, such as improved methodologies inspired by CycleGAN, may further enhance domain correspondences. Self-supervised learning approaches could also play a significant role in reducing dependency on annotated datasets while fostering the extraction of domain-invariant features. Finally, exploring segmentation models based on transformers and expanding testing across diverse datasets, including scenarios with low lighting and adverse weather, can provide deeper insights into the adaptability of the proposed system.

REFERENCES

Bousmalis, K., Frosio, N. D., Bursuc, D., Hays, J., Tsoumakas, G., and Metaxas, D. N. (2016). Domain separation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1709–1717.

Chen, W., Wei, Y., Yang, Y., Wang, Z., Li, W., and Wang, X. (2020). Contrastive learning for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7568–7577.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., F., R., Y., A., and Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223.

Ding, Z., Wang, Q., Huang, J., Zhang, K., and Xie, L. (2019). Dpr: Domain propagation network for cross-domain semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3330–3340.

Hoffman, J., Tzeng, E., Park, T., Saenko, K., and Darrell, T. (2018). Cycada: Cycle-consistent adversarial domain adaptation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1996–2005.

Liu, L., Lu, H., Lee, L., Yang, M., Wong, T.-L., Wu, D., and Lin, Y.-W. (2020a). Clan: Class-wise alignment for unsupervised domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2249–2257.

Liu, W., Chen, B., Zhang, Z., Li, X., and Li, X. (2020b). Msl: Multi-scale learning for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4706–4715.

Richter, S., D., V. W. W., M., R., R., A., A., M., and A., M. (2016). Playing for data: Ground truth from computer games. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 102–118. Springer.

Sankaranarayanan, S., Balaji, Y., Jain, A., Kumar, S. R., and Gupta, A. (2018). Learning from synthetic data: Addressing domain shift for semantic segmentation.

In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2701–2710.

Shaban, A., Zhang, T., Fang, W., Zhang, Y., Xiang, T., and Li, Z. (2018). Dlow: Domain-learning by optimizing the distribution of labels in unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5367–5376.

Sibi, P., Viswanath, V., Bhat, P., Mottaghi, R., and Farhadi, A. (2019). Siban: Segmentation-invariant batch normalization for cross-domain semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4550–4559.

Tsai, Y.-H., Chiu, W.-C., Hsu, H.-W., Sun, M., and Yang, M.-H. (2018). Adaptsegnet: Adversarial adaptation for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6956–6965.

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. (2017). Adversarial discriminative domain adaptation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7167–7176.

Volpi, M., Mancini, M., Bria, A. M., Massemini, R., and Perona, P. (2018). Minimum class discrepancy for domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 602–617.

Vu, D., Cheng, W.-L., Yang, Y., Kumar, N., Lee, H., and Lin, M.-H. (2019a). Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2226–2235.

Vu, T.-H., Jain, H., Bucher, M., Goldluecke, B., and Gool, L. V. (2019b). Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2517–2526.

Xu, Y., Shen, X., Li, Y., Duan, X., Zhang, Z., and Jia, J. (2021). Daformer: Domain-adaptive transformer for cross-domain semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8983–8992.

Zhang, L., Wen, L., Ji, R., and Luo, P. (2020a). Aligning higher semantic features for cross-domain image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 49–65.

Zhang, Y., Wei, Z., Liu, Z., Yang, S., and Wei, Z. (2020b). Multi-task learning for cross-domain semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4503–4512.

Zhang, Y., Wei, Z., Liu, Z., Yang, S., and Wei, Z. (2020c). Seanet: Squeeze-enhanced axial attention network for nighttime segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4535–4544.

Zhao, H., Li, X., Zhang, Z., Wang, H., and Wang, X. (2019). Dise: Domain invariant semantic embeddings for cross-domain semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3341–3349.

Zhou, L., Di, X., and Yao, J. (2019). Cycle-consistent generative adversarial networks for unsupervised domain adaptation in image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232.