Real-time Domain Adaptation in Semantic Segmentation

Ali Samimi Fard
S308956@studenti.polito.it

Mahsa Mohammadi
S312251@studenti.polito.it

Mohammadreza Mashhadigholamali
S308499@studenti.polito.it

Abstract

Semantic segmentation is a crucial task in computer vision that involves classifying each pixel in an image into predefined categories, enabling detailed scene understanding and facilitating applications such as autonomous driving, medical imaging, and robotics. This project explores real-time domain adaptation techniques in semantic segmentation networks. Initially, a classic DeepLabV2 network and a real-time BiSeNet network were trained on the Cityscapes dataset, achieving mIoU scores of 53.57% and 34.62%, respectively. The domain shift problem was then assessed by training BiSeNet on the synthetic GTA5 dataset and evaluating it on Cityscapes, resulting in a significant performance drop to 21.79% mIoU. To mitigate this shift, data augmentation techniques were applied, improving the mIoU to 23.60% with Augmentation 1 (Gaussian Blur) and 21.89% with Augmentation 2 (Horizontal Flip). A combination of both augmentations yielded an mIoU of 23.56%. Finally, domain adaptation methods FDA and DACS were implemented, with FDA achieving an mIoU of 29.21%, while DACS results are pending.

1. Introduction

Semantic segmentation can be described as a basic procedure in the field of computer vision and implies the assignment of pixel-level classes to the given image. This task is critical when scrutinizing scenes and has many applications such as in car driving [1] , diagnosis using scans [2] and physical movements using robots [3]. Nevertheless, heterogeneous object matter is one of the key open problems even knowing a great progress in semantic segmentation solutions, one of the main problems being the domain shift problem, in which a model trained on a certain dataset (the source domain) works inefficiently on another dataset (the target domain).

1.1. Problem Statement

The domain shift problem becomes a major challenge towards the actual implementation of semantic segmentation models in practical applications. For example, a model which has been trained using synthetic data such as GTA5 [4] would possibly have a poor performance when tested on Cityscapes [5] datasets. Solving this problem is critical to guaranteeing proper functioning of semantic segmentation models regardless of the environment.

1.2. Approach

In this project, we explore various approaches to mitigate the domain shift problem in semantic segmentation:In this project, we explore various approaches to mitigate the domain shift problem in semantic segmentation:
Model Choice: Cityscapes is described as a dataset used for urban scene understanding where we will be using DeepLabV2 [6], a classic segmentation network, and BiSeNet [7], a real-time segmentation network to evaluate them.
Data Augmentation: To augment our dataset and enhance the performance of our models we perform data augmentation like horizontal flipping and Gaussian blurring.
Our main contributions are:

- Training and evaluating DeepLabV2 and BiSeNet on the Cityscapes dataset.
- Assessing the domain shift problem by training BiSeNet on the GTA5 dataset and evaluating it on Cityscapes.
- Implementing data augmentation techniques to reduce the impact of domain shift.

Domain Adaptation Techniques: Here, we perform FDA for matching the source and target domains and DACS using the cross-domain mixed sampling technique to perform the alignment.
The remainder of the paper is structured as follows: Section 2: Related Work discusses the existing literature on

semantic segmentation, domain adaptation, and data augmentation.
Section 3: Methodology outlines the experimental setup, including datasets, models, and training protocols.
Section 4: Experiments and Results present the experimental results and analyze the effectiveness of the proposed approaches.
Section 5: Conclusion summarizes the findings and suggests directions for future work.

1.3. Related work

1.3.1 Semantic Segmentation

DeepLab Series: The DeepLab series such as DeepLabV2 has certainly progressed the field of semantic segmentation remarkably. The accuracy of the segmentation is improved by atrous convolution and fully connected Conditional Random Fields (CRFs) in DeepLabV2 approach. Atrous convolution helps in creating large receptive field without necessarily increasing the number of parameters while CRFs helps in improving the edges of the segmentation by considering spatial structures. This model is used as a reference point for qualifying new approaches of segmentation and is applied more or less in several fields [6].

1.3.2 Real-Time Segmentation

BiSeNet (Bilateral Segmentation Network): Originally aimed at real-time semantic segmentation, BiSeNet resolves the problem of aspect of time as well as aspect of space. It employs a dual-path architecture: An information preserving path called Spatial Path and a high-level context path for a lightweight model termed as Context Path. This design allows BiSeNet to function well in real-time applications like self-driving cars, for which the segmentation results need to be produced quickly and accurately [7].

1.3.3 Data Augmentation

General Techniques: Horizontal flipping, Gaussian blurring, and color jitter introduction as data augmentation methods improves the model's training set diversification, and, in turn, the model versatility in regards to unseen inputs. They are important for enhancing the efficiency of deep learning models on several computer vision problems [8].
Mixup: Mixup is a more sophisticated data augmentation method where instead of creating completly new samples, it combines two existing samples by taking a convex combination of both. This approach assists the model to learn smoother decision boundaries, and the method has been established as useful in many tasks such as semantic segmentation [9].

1.3.4 Domain Adaptation

Unsupervised Domain Adaptation: In the case of unsupervised domain adaptation, techniques are involved in the problem of the shift of the domain, for instance, from synthetic images to real-world images. Such methods employ adversarial learning to match the distributions of features extracted from the source and target domain; it helps to use the former for the target task, as the latter contains domain-invariant features [10].
Image-to-Image Translation: FDA and DACS are recognized as well-known image-to-image translating methods. However, FDA transform the style of the low-frequency spectrum images to fix the target domain while enhancing the segmentation retention without requiring labels from the target domain [11].
DACS generates a training set of images of both domains and this makes the model to learn the features that are insensitive to the variations within the domains directly leading to improved cross-domain performance [12].

1.3.5 Benchmark Datasets

Cityscapes: The Cityscapes dataset and Cityscapes contains urban street scenes, which is huge and containing pixel-level annotations. Cityscapes is used often in the context of semantic segmentation since the dataset contains a variety of images and many of them are difficult, therefore, it constitutes a good test for checking the performance of the models [5].
GTA5: The GTA5 dataset consists of synthetic images coming from a video game named Grand Theft Auto V which is annotated in the same way as the Cityscapes dataset. Attributed to its realistic features and scenarios, it is suitable for training models in domain adaptation investigations for use in more realistic scenarios [4].

2. Methodology

2.1. Cityscapes Dataset

Overview: The Cityscapes dataset is a large-scale benchmark suite designed for semantic urban scene understanding. It is widely used in the computer vision community, particularly for tasks such as semantic segmentation. The dataset consists of high-resolution images captured from diverse urban environments across 50 different cities.
Subset Description: In this project, we utilized a subset of the Cityscapes dataset, focusing on 19 essential classes relevant to urban scene understanding. This subset comprises a

total of 2,072 images, with 1,572 images allocated for training and 500 images for testing. Each image in the dataset is accompanied by a corresponding ground truth annotation, providing pixel-level labels for precise semantic segmentation.
Classes: The 19 classes included in this subset are:

- Flat: road, sidewalk
- Construction: building, wall, fence
- Object: pole, traffic light, traffic sign
- Nature: vegetation, terrain
- Sky: sky
- Human: person, rider
- Vehicle: car, truck, bus, train, motorcycle, bicycle

these classes cover a comprehensive range of objects and structures commonly found in urban environments, providing a robust foundation for semantic segmentation tasks.
Annotations: Each image in the subset is annotated at the pixel level, offering detailed semantic labels for each of the 19 classes. The annotations are divided into:

- Fine Annotations: High-quality, detailed pixel-level annotations.
- Coarse Annotations: Less detailed annotations that are still valuable for training models leveraging large volumes of weakly-labeled data.

Image Specifications

- **Resolution:** Each image is provided at a high resolution of 2048x1024 pixels.
- **Variability:** The images were captured under various weather conditions, times of day, and seasons to ensure diversity and robustness in the dataset.

Example Images: For each image in the dataset, there are three associated files:

1. Original Image: The raw, high-resolution image captured from the urban environment.
2. Annotation (Label): The pixel-level annotated image indicating different classes with unique colors.
3. Color-Coded Annotation: A color-coded version of the annotated image for better visual understanding.

2.2. GTA5 Dataset

Overview: The GTA5 dataset is derived from the game Grand Theft Auto V (GTA5) and is used for training and evaluating semantic segmentation models. This dataset provides high-quality, pixel-accurate semantic annotations by leveraging the photorealistic and diverse environments within the game. It is particularly valuable for tasks in urban scene understanding, as it simulates real-world conditions with high fidelity.
Data Acquisition: The dataset was created by extracting frames from the game and annotating them with semantic labels. The approach involved capturing images from the game, then using advanced techniques to assign semantic labels to each pixel in the images.
Key Features:

- **Images and Annotations:** The full dataset includes 25,000 images with dense pixel-level semantic annotations. For this project, a subset of 2,500 images was used.
- **Resolution:** Each image has a resolution of 1914x1052 pixels.
- **Annotation Speed:** The labeling process was completed in 49 hours, achieving a much faster annotation speed compared to traditional methods.

Classes: The dataset includes 19 classes compatible with other semantic segmentation datasets for urban scenes, such as the Cityscapes dataset. These classes cover a wide range of objects and structures found in urban environments.
Methodology:

- Data Collection:
  - Frames were extracted from the game at regular intervals during gameplay.
  - RenderDoc was used to record every 40th frame during gameplay, which was then processed to extract relevant information for annotation.
- Annotation Process:
  - The extracted frames were annotated using an interactive interface that allowed for efficient labeling.
  - The process utilized association rule mining to propagate labels across similar objects and scenes, significantly speeding up the annotation process.

2.3. DeepLabV2 Model

Overview: DeepLabV2 is a classic semantic segmentation network that enhances the capability of Deep Convolutional Neural Networks (DCNNs) through the use of atrous convolution and fully connected Conditional Random Fields (CRFs). This model addresses common challenges in semantic segmentation, such as reduced feature resolution, handling multiple object scales, and improving localization accuracy.

Key Features:

- **Atrous Convolution:** Atrous convolution (also known as dilated convolution) allows for controlling the resolution of feature responses within DCNNs. It effectively enlarges the field of view of filters without increasing the number of parameters or computation, maintaining high-resolution feature maps.
- **Atrous Spatial Pyramid Pooling (ASPP):** ASPP captures objects and image context at multiple scales by probing convolutional feature layers with filters at different sampling rates and fields of view. This multi-scale approach is beneficial for segmenting objects of various sizes.
- **Conditional Random Fields (CRFs):** Fully connected CRFs refine the segmentation results by combining the coarse segmentation from DCNNs with fine-grained details, improving the delineation of object boundaries.

Backbone - ResNet-101: DeepLabV2 employs ResNet-101 as its backbone network. ResNet-101, part of the Residual Networks family, is renowned for its deep architecture and ability to mitigate the vanishing gradient problem through the use of residual learning.

In the context of DeepLabV2, ResNet-101's fully connected layers are transformed into convolutional layers, and atrous convolution layers replace standard convolution layers in the later stages of the network. This setup allows the model to maintain high-resolution feature maps crucial for precise semantic segmentation.

2.4. BiSeNet (Bilateral Segmentation Network)

Overview: BiSeNet (Bilateral Segmentation Network) is a real-time semantic segmentation network designed to achieve a balance between high accuracy and fast inference speed. It is particularly well-suited for applications requiring quick responses, such as autonomous driving and video surveillance.

Key Features:

- Spatial Path (SP): The Spatial Path is designed to preserve the spatial resolution of the input image. It consists of three convolution layers with a stride of 2, which helps retain rich spatial details crucial for accurate segmentation.
- Context Path (CP): The Context Path is responsible for capturing sufficient receptive field. It utilizes a lightweight model, such as Xception, along with global average pooling to provide large receptive fields and encode high-level semantic context information.
- Feature Fusion Module (FFM): The Feature Fusion Module combines the features from the Spatial Path and the Context Path. This module effectively balances the low-level spatial information and high-level context information, ensuring that the final segmentation output is both accurate and detailed.
- Attention Refinement Module (ARM): The ARM refines the features of each stage in the Context Path by employing global average pooling to capture global context and computing an attention vector to guide feature learning. This integration enhances the output features without significant computational overhead.

Architecture:

BiSeNet features two distinct pathways: the Spatial Path and the Context Path. The Spatial Path maintains the spatial resolution, while the Context Path focuses on acquiring a comprehensive receptive field. The outputs from these paths are fused through the Feature Fusion Module to produce the final segmentation map.

2.5. Augmentation

Augmentation Set 1: Gaussian Blur

- Description: Gaussian blur is a technique that smoothens the image by applying a Gaussian function. It helps in reducing noise and detail in the image, which can make the model more robust to variations in the input data.

Augmentation Set 2: Horizontal Flip

- Description: Horizontal flipping is a common augmentation technique that involves flipping the image horizontally. It helps the model to learn invariant features that are not dependent on the orientation of the objects in the image.

Combined Augmentation Approach

- Description: This approach combines the two augmentation sets by randomly applying either the Gaussian blur or the horizontal flip to each image. This increases the variability of the training data and further enhances the model's ability to generalize to new, unseen data.

2.6. Domain Adaptation Techniques

Fourier Domain Adaptation (FDA): Fourier Domain Adaptation (FDA) is a technique in unsupervised domain adaptation to minimize the domain shift in such a way that changing the low frequency spectrum of the source images domain to that of the target images domain. This technique does not involve the use of training for alignment of the data domain, just Fourier Transform and its inverse. This causes the low-level statistics between the source and target domains to be matched, which enhances the efficiency of the models in the target domain's semantic segmentation. The main adjustable for FDA is the size of spectral neighborhood to be swapped, denoted by β, which controls the amount of domain alignment.
Domain Adaptation via Cross-domain Mixed Sampling (DACS): Domain Adaptation via Cross-domain Mixed Sampling (DACS) is an improvement over DA in which samples from the source and target domains are merged to form a training set. This method creates the mixed-domain images by having the contents of the source images and the styles of the target images, which promotes cross-domain features learning. Actually, DACS achieves the integration of the positive aspects of both domains and enhances the model's generalization performance toward the target domain.

2.7. Evaluation Metrics

- **Mean Intersection over Union (mIoU):** The primary evaluation metric used to assess model performance. mIoU measures the accuracy of the predicted segmentation maps against the ground truth annotations.
- **FLOPs (Floating Point Operations):** Measures the computational complexity of the model.
- **Params (Number of Parameters):** Indicates the model size.
- **Latency:** Measures the inference speed of the model

it is worth to mention that FLOPs, Params and Latency were calculated only in step 2.

2.7.1 Training Details

- Learning Rate:
  - For step 2: $1 * 10^4$
  - For the rest: $1 * 10^2$
- Number of Epochs: 50
- Batch Size: 4
- Optimizer: SGD with momentum of 0.9 and weight decay of $1 * 10^4$
- Loss Function: Cross-entropy loss with an ignore index for unlabeled pixels (255).

3. Experiment

3.1. Classic Semantic Segmentation Network

Objective: The objective of this experiment is to evaluate the performance of a classic semantic segmentation network, DeepLabV2, on the Cityscapes dataset.
Experimental Setup: For this experiment, we used the following setup:

- Dataset: Cityscapes
- Training and Test resolution: 1024x512
- Backbone: ResNet-101 (pre-trained on ImageNet)

The performance of DeepLabV2 on the Cityscapes dataset after 50 epochs is summarized in Table 1. The table includes mIoU, latency, FLOPs, and the number of parameters. Also, the loss-epoch plot and mIoU-epoch plot for training and validation is illustrated in figure 1.

3.2. Real-time Semantic Segmentation Network

Objective: The objective of this experiment is to evaluate the performance of a real-time semantic segmentation network, BiSeNet, on the Cityscapes dataset.
Experimental Setup: For this experiment, we used the following setup:

- Dataset: Cityscapes
- Training resolution and Test: 1024x512
- Backbone: ResNet-18 (pre-trained on ImageNet)

Table 2 demonstrates the performance of a real-time semantic segmentation of BiSeNet model on the Cityscapes dataset

3.3. Evaluating the Domain Shift Problem

Objective: The objective of this experiment is to evaluate the domain shift problem in semantic segmentation by training a real-time segmentation network (BiSeNet) on synthetic images from the GTA5 dataset and evaluating its performance on real images from the Cityscapes dataset.
Experimental Setup: For this experiment, we used the following setup:

- Dataset: GTA5 (source domain), Cityscapes validation split (target domain)
- Training resolution (GTA5): 1280x720
- Test resolution (Cityscapes): 1024x512
- Backbone: ResNet-18 (pre-trained on ImageNet)

3.4. Data Augmentations to Reduce the Domain Shift

Objective: The objective of this experiment is to improve the generalization capability of the segmentation network trained on the synthetic domain by using data augmentations during training. Data augmentations help to virtually expand the dataset size and modify the visual appearance of source (synthetic) images to make them more similar to target (real) ones.
Experimental Setup: For this experiment, we repeated the previous setup with the addition of data augmentations at training time. The augmentations used are:

- Augmentation 1: Horizontal Flip
- Augmentation 2: Gaussian Blur
- Combined Augmentation: Combination of Augmentation 1 and Augmentation 2

The probability of performing augmentation was set to 0.5.

- Dataset: GTA5 (source domain), Cityscapes validation split (target domain)
- Training resolution (GTA5): 1280x720
- Test resolution (Cityscapes): 1024x512
- Backbone: ResNet-18 (pre-trained on ImageNet)

3.5. Image-to-Image Adaptation Approach

Objective: The objective of this experiment is to evaluate the performance of the segmentation network using image-to-image adaptation techniques, specifically FDA (Fourier Domain Adaptation) and DACS (Domain Adaptation via Cross-domain Mixed Sampling), to mitigate the domain shift problem.
Experimental Setup: For this experiment, we used the following setup:

- Source Synthetic Labeled Dataset: GTA5
- Target Real-World Unlabeled Dataset: Cityscapes
- Image-to-Image Adaptations: FDA and DACS
- Best Setting from previous step: The best performing data augmentation setting from previous step was used (Horizontal Flip, Gaussian Blur).

4. Results

4.1. Classic Semantic Segmentation Network

Table 1 demonstrates the performance of DeepLabV2 on the Cityscapes dataset.

Classic Cityscapes	mIoU (%)	Latency	FLOPs	Params
DeepLabV2	53.57%	72.94 ms	0.375 T	43.901M

Table 1. Performance of DeepLabV2 on the Cityscapes

we observe that the mean Intersection over Union (mIoU) achieved by DeepLabV2 after 50 epochs of training on the Cityscapes dataset was 53.57%. Also, the latency for inference was 72.94 milliseconds, and the model required 0.375 FLOPs and had 43.901M parameters. figure 1 illustrates loss and mIoU value for each epoch.

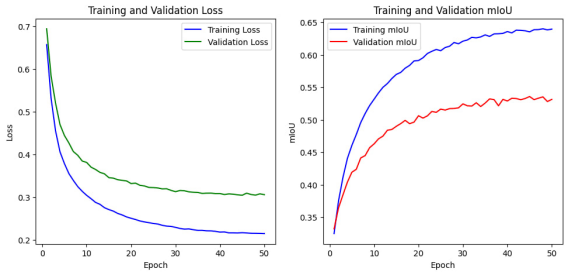

Figure 1. Train and Test Loss and mIoU

As observed in the training and validation curves for loss and mIoU in the figure 1, it is observed that the training process is proper as both training and validation loss is decreasing consistently epoch wise. Furthermore, the mIoU increases successively also for the training set and the validation set indicating that the model is learning successfully and on the validation set also. The marginal differences between the training and validation mIoU curves at some later iteration might suggest a minor overfitting; however, it could be alleviated with more elaborate forms of regularization or data augmentation.

4.2. Real-time Semantic Segmentation Network

Table 2 summarizes the performance of BiSeNet on the Cityscapes. The table includes mIoU, latency, FLOPs, and the number of parameters.

Real-time Cityscapes	mIoU (%)	Latency	FLOPs	Params
BiSeNet	34.62 %	5.99 ms	25.78 G	12.582 M

Table 2. Performance of BiSeNet on the Cityscapes

in this step, we observe that the mean Intersection over Union (mIoU) achieved by BiSeNet after 50 epochs of training on the Cityscapes dataset was 34.62%. Also, the latency for inference was 5.99 milliseconds. Finally, the model required 25.78 G FLOPs and had 12.582 M parameters. figure 2 demonstrates loss vs.epoch and mIoU vs.epoch plots.

The training and validation curves for loss and mIoU indicate that the training process converges well, with both

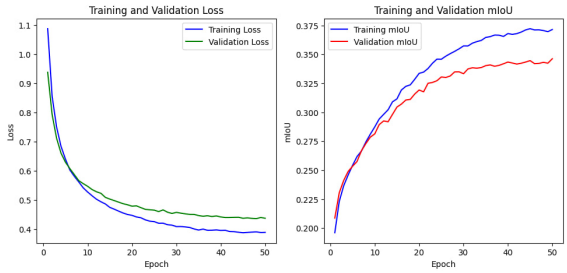

Figure 2. Train and Test Loss and mIoU

training and validation loss decreasing steadily over the epochs. The mIoU improves consistently for both training and validation, suggesting that the model is learning effectively and generalizing well to the validation set. The slight gap between the training and validation mIoU curves towards the end of training might indicate a minor overfitting, which could be addressed with additional regularization or data augmentation techniques.

4.3. Evaluating the Domain Shift Problem

The mean Intersection over Union (mIoU) achieved by BiSeNet after 50 epochs of training on the GTA5 dataset and evaluating on the Cityscapes dataset was 21.79%. The domain shift problem occurs because the synthetic images from the GTA5 dataset differ significantly from the real images in the Cityscapes dataset in terms of texture, lighting, and other visual features.

figure 3 demonstrates loss and mIoU value for each epoch.

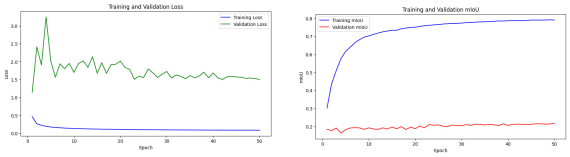

Figure 3. Train and Test Loss and mIoU

The results demonstrate that the domain shift problem has a significant impact on the performance of the segmentation network. The mIoU drop from 34.62% to 21.79% when moving from a source domain (GTA5) to a target domain (Cityscapes) underscores the need for domain adaptation techniques to bridge this gap. The detailed class-wise performance in Table 3 can provide insights into which classes are most affected by the domain shift.

4.4. Data Augmentations to Reduce the Domain Shift

The performance of BiSeNet trained on the GTA5 dataset with different augmentations and evaluated on the Cityscapes dataset is summarized in Table 4. The table includes mIoU for each class and the overall mIoU.

The mean Intersection over Union (mIoU) achieved by BiSeNet after 50 epochs of training on the GTA5 dataset with Augmentation 1 (Gaussian Blur) and evaluating on the Cityscapes dataset was 23.60%. The mean Intersection over Union (mIoU) achieved by BiSeNet after 50 epochs of training on the GTA5 dataset with Augmentation 2 (Horizontal Flip) and evaluating on the Cityscapes dataset was 21.89%. The mean Intersection over Union (mIoU) achieved by BiSeNet after 50 epochs of training on the GTA5 dataset with combined Augmentation 1 and Augmentation 2 and evaluating on the Cityscapes dataset was 23.56%.

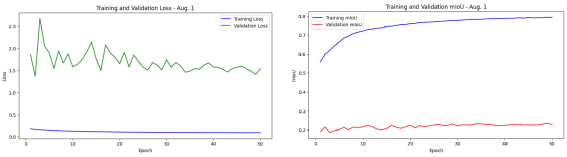

Figure 4. Train and Test Loss and mIoU Augmentation 1

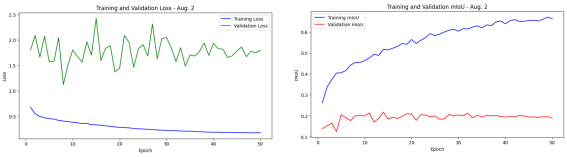

Figure 5. Train and Test Loss and mIoU Augmentation 2

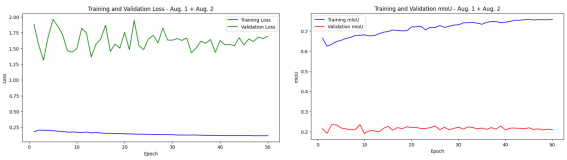

Figure 6. Train and Test Loss and mIoU Augmentation 1 & 2

Thus, the results show that the considered data augmentations can to a certain extent improve the situation with the domain shift. More precisely, it can be noted that horizontal flipping (Augmentation 1) gave the highest increase in mIoU among all the augmentations. The results showed that when adding the two augmentations together, it was not much different from applying only horizontal flipping, implying that some augmentations are not synergistic or could actually hinder each other. Specifically, the detailed class-wise performance in Table 4 can give ideas about which classes are most affected with the domain shift and what augmentations benefit most.

4.5. Image-to-Image Adaptation Approach

The mean Intersection over Union (mIoU) achieved by BiSeNet after 50 epochs of training on the GTA5 dataset with FDA adaptation and evaluating on the Cityscapes dataset was 29.21%. FDA improved the mIoU significantly compared to the baseline results without adaptation (Table []), highlighting the effectiveness of these image-to-image adaptation technique.

The result suggest that using image-to-image adaptation techniques means the degree of the domain shift is low and FDA' mIoU score is improved. In improving the mIoU of the source and target domains, adaptation of the FDA was established to bring about substantial enhancement. The information in terms of classes has been presented in Table 5 in detail where it could be observed that which classes are positively impacted by which of these adaptation techniques the most.

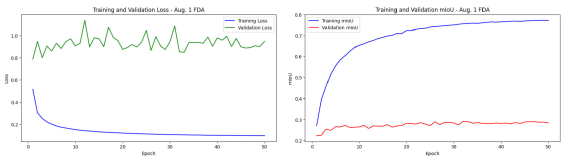

Figure 7. Train and Test Loss and mIoU FDA

5. Conclusion

In this study, we addressed the significant challenge of domain shift in semantic segmentation by evaluating and enhancing the performance of two key models: It is composed of two architectures, namely DeepLabV2 and BiSeNet. Using Cityscapes dataset, the performance of these models was established baselines with DeepLabV2 having 53 mIoU. reduction in SegNet reaching up to 57% and BiSeNet achieving 34. 62%. However, if trained on the synthetic GTA5, BiSeNet's performance drastically decreased to 21 on Cityscapes. It results show that the model obtained 79% mIoU, pointing out that domain shift clearly affects the results.

To combat this, we used data augmentation consisting of Gaussian blurs and horizontal flips to the images, which enhanced results by a small amount. Moreover, we discussed other techniques in the field of DA such as FDA that improved the mIoU to 29. 21%. This implies that data augmentation indeed improves the model performance on unseen data, but domain adaptation is a better solution to the problem of difference between augmented and original data. It is with these results in mind that more work in domain adaptation must be done to precisely improve and enhance semantic segmentation in different conditions.

References

[1] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Gläser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1341–1360, 2021. 1

[2] B. Kayalibay, G. Jensen, and P. van der Smagt, "Cnn-based segmentation of medical imaging data," *arXiv preprint arXiv:1701.03056*, 2017. 1

[3] J. V. Hurtado and A. Valada, "Semantic scene segmentation for robotics," in *Deep learning for robot perception and cognition*, pp. 279–311, Elsevier, 2022. 1

[4] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 102–118, Springer, 2016. 1, 2

[5] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016. 1, 2

[6] 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, 2017. 1, 2

[7] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 325–341, 2018. 1, 2

[8] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019. 2

[9] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017. 2

[10] 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," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7472–7481, 2018. 2

[11] Y. Yang and S. Soatto, "Fda: Fourier domain adaptation for semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4085–4095, 2020. 2

[12] W. Tranheden, V. Olsson, J. Pinto, and L. Svensson, "Dacs: Domain adaptation via cross-domain mixed sampling," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1379–1389, 2021. 2