Real-time Domain Adaptation in Semantic Segmentation

Tommaso Mazzarini
Politecnico di Torino
tommaso.mazzarini@studenti.polito.it

Leonardo Merelli
Politecnico di Torino
leonardo.merelli@studenti.polito.it

Giovanni Stinà
Politecnico di Torino
giovanni.stina@studenti.polito.it

Abstract

In a context where real-time neural networks are becoming increasingly fundamental for critical applications such as autonomous driving and surveillance video, domain adaptation is a crucial challenge. This work illustrates the application of domain adaptation techniques on real-time neural networks for semantic segmentation. The main goal is to enhance the networks' accuracy when transitioning from synthetic to real-world domains. Initially, we trained a classical segmentation network (DeepLabV2) and a real-time segmentation network (BiSeNet) on the Cityscapes dataset to compare their results and establish an upper bound for performance. Then, we explored the domain shift problem by training the BiSeNet model on the synthetic GTA5 dataset and evaluating its accuracy on the Cityscapes dataset, assessing the negative impact on performance. To mitigate this impact, we applied data augmentation techniques, followed by an adversarial learning method for domain adaptation to further increase the model's effectiveness. The results obtained were evaluated using metrics such as mean Intersection over Union (mIoU), FLoating point Operations Per second (FLOPs), latency, and the number of parameters of the models. The proposed techniques demonstrated significant improvements, bringing real-time networks closer to the performance of classical networks despite the domain change.

1. Introduction

Semantic segmentation is a fundamental task in computer vision that involves assigning category labels to each pixel in an image. This task is crucial for applications ranging from surveillance video to medical image segmentation, as it provides detailed spatial and contextual information. Recent advancements in deep learning, particularly on convolutional neural networks (CNNs), have significantly improved the accuracy of classical semantic segmentation models.

However, as the demand for real-time applications increases, particularly in areas like autonomous driving, the necessity for real-time semantic segmentation networks becomes essential. These networks need to balance accuracy and computational efficiency, operating on limited hardware while keeping latency low. This project aims to explore real-time semantic segmentation, with a specific focus on the BiSeNet architecture.

A significant challenge in developing robust semantic segmentation models is the limited availability and high cost of labeled real-world data. For example, building a comprehensive dataset for self-driving requires extensive manual annotation of street scenes across various weather and lighting conditions, which is time-consuming and expensive. Synthetic datasets, such as those derived from video games, present a promising alternative. However, models trained on synthetic data often underperform in real-world scenarios due to discrepancies in texture, lighting, and other visual attributes between synthetic and real environments, a phenomenon known as domain shift. This project tackles the challenge of domain shift by investigating domain adaptation techniques, with a particular emphasis on adversarial methods. The goal is to bridge the gap between synthetic and real-world domains, aiming to improve the transferability of models trained on synthetic data to real-world domain.

Specifically, this survey will investigate the performance of classical and real-time semantic segmentation networks, namely DeepLabV2 [1] and BiSeNet [13] respectively, on the Cityscapes dataset [2], a benchmark for urban scene understanding. We will then explore the impact of domain shift by training BiSeNet on the synthetic GTA5 [9] dataset and evaluating its performance on Cityscapes. To mitigate the domain shift, we will implement and evaluate various domain adaptation approaches, including data augmenta-

tion and an adversarial discriminative method.

The report is structured in four sections. Related Works reviews some classical semantic segmentation networks, real-time networks, and adversarial domain adaptation techniques. Methods provides a detailed description of DeepLabV2 and BiSeNet networks, as well as the chosen domain adaptation approaches. Experimental Results present datasets, implementation setups, metrics used, and detailed results analysis. Conclusion summarizes the findings and provides final remarks.

Through this study, our aim is to explore domain adaptation applied to real-time semantic segmentation networks, focusing on improving their performance and applicability in the field of computer vision and its practical scenarios.

2. Related Works

This section explores important developments in semantic segmentation and domain adaptation techniques, discussing classical and real-time networks for semantic segmentation, as well as different approaches to overcome the challenges associated with domain shift.

2.1. Classical Semantic Segmentation Networks

Among the various classic networks of semantic segmentation, three notable models stand out.

Fully Convolutional Network (FCN). The FCN [4, 6], introduced by Long et al. (2016), was the first deep neural network designed exclusively for pixel-wise semantic segmentation, resulting in a notable increase in segmentation accuracy with respect to classical approaches. By employing skip connections, FCN combined deep and shallow feature maps innovatively to enhance segmentation accuracy. This approach highlights the advantage of utilizing both shallow and deep layer features to improve semantic segmentation accuracy significantly.

U-Net. Developed by Ronneberger et al. (2015), U-Net [4, 10] introduced a symmetrical encoder-decoder architecture with dense skip connections, fully leveraging features from each layer by connecting the encoder's features to their corresponding layers in the decoder. This network revolutionized medical image analysis by preserving fine details throughout the segmentation process. Its design facilitated robust feature propagation between encoder and decoder layers, improving segmentation precision.

DeepLabV2. DeepLabV2 [1, 4], a model within the DeepLab family specialized in semantic segmentation, incorporated dilated convolutions and Conditional Random Fields (CRFs) to preserve feature resolution, manage objects of different scales, and improve spatial accuracy. Detailed by Chen et al. (2018), these architectural improvements enabled DeepLabV2 to effectively capture both global and local contexts in images, significantly enhancing the accuracy of semantic segmentation.

These classical semantic segmentation models represent significant advancements in semantic segmentation. They have introduced innovative methodologies aimed at preserving intricate spatial details and improving contextual understanding. These advancements collectively enhance the accuracy and performance of semantic segmentation techniques across a wide range of applications.

2.2. Real-time Semantic Segmentation Networks

Real-time semantic segmentation networks aim to balance accuracy and efficiency, while traditional semantic segmentation networks, despite their high accuracy, struggle with inference time due to high computational costs, limiting their real-time application. This section discusses two of the most important real-time semantic segmentation networks: ENet and BiSeNet.

Efficient Neural Network (ENet). Introduced by Paszke et al. (2016), ENet [4, 8] is a pioneering model designed for real-time semantic segmentation. Based on factorized convolution and an early downsampling strategy, ENet significantly reduces computational complexity while maintaining competitive accuracy. Its innovative architecture enables rapid image processing, making it suitable for real-time applications on devices with limited computational resources. ENet has demonstrated robust performance across various datasets of road scenes and indoor environments, proving its efficacy. This balance of speed and accuracy makes ENet particularly valuable for many real-time applications.

Bilateral Segmentation Network (BiSeNet). Proposed by Yu et al. (2018), BiSeNet [4, 13] represents a significant advancement in real-time semantic segmentation. This network introduces a novel architecture that divides the segmentation task into two paths: a context path for capturing semantic information and a spatial path for preserving fine spatial details. BiSeNet exploits a feature fusion module to integrate information from these paths, achieving a good balance between accuracy and computational efficiency.

These networks illustrate ongoing efforts to improve the efficiency of semantic segmentation models without sacrificing accuracy.

2.3. Adversarial Discriminative Methods

Adversarial discriminative techniques are a powerful class of methods in domain adaptation that use adversarial objectives to induce domain confusion and promote the learning of domain-invariant features. These methods bridge the gap between the source and target domains by making the extracted features from both domains indistinguishable to a domain discriminator. This section delves into some key adversarial discriminative model.

Domain-Adversarial Neural Network (DANN). Introduced by Ganin et al., DANN [3, 14] is one of the pioneer-

ing models in adversarial discriminative techniques. This model optimizes the H-divergence between the source and target domains using a gradient reversal layer. This layer adversarially trains the feature extractor to maximize the domain discriminator's loss, making the source and target features indistinguishable. This helps the model learn domain-invariant features.

Adversarial Discriminative Domain Adaptation (ADDA). Proposed by Tzeng et al., ADDA [12, 14] separates the optimization process into two independent objectives for the generator and the discriminator using an inverted label GAN loss. This method stabilizes the training process and avoids issues such as vanishing gradients, effectively aligning the feature distributions of the source and target domains.

Conditional Domain Adversarial Network (CDAN). Presented by Long et al., CDAN [7, 14] focuses on aligning conditional distributions across domains. This model uses a conditional discriminator that captures the cross-covariance between feature representations and classifier predictions. This approach improves the discriminability of the model and aligns joint distributions.

Transferable Adversarial Training (TAT). TAT [5, 14], that was introduced by Liu et al., extends feature-level alignment by generating transferable examples that help the classifier to learn a more robust decision boundary. This method adapts feature representations and increases the model's robustness in handling domain shifts.

The models presented are just a part of the broader spectrum of adversarial discriminative techniques used in domain adaptation. These methodologies represent a versatile and effective approach to addressing domain shift in adaptation tasks. By leveraging adversarial training to induce domain confusion and promote the learning of domain-invariant representations, these techniques play a crucial role in reducing the domain shift impact in many applications.

3. Methods

This section outlines the methods employed in our project, focusing on two advanced neural network architectures, DeepLabV2 and BiSeNet. We also address the domain shift issue by discussing an adversarial domain adaptation method.

3.1. DeepLabV2: Classical Semantic Segmentation Network

DeepLabV2 [1] is a significant advancement in the field of semantic segmentation, building on the strengths of Deeplab by incorporating atrous (dilated) convolutions and an improved multi-scale contextual aggregation method. Atrous convolutions are pivotal in DeepLabV2, they solve the problem of spatial reduction without increasing the number of parameters or the computational complexity, thereby preserving the spatial resolution of the feature maps. This is achieved by inserting zeros between the filter weights, effectively enlarging the field of view of the filters. The Atrous Spatial Pyramid Pooling (ASPP) applies atrous convolutions with different dilation rates in parallel, which helps in capturing objects and context at multiple scales. Additionally, DeepLabV2 employs the fully connected Conditional Random Field (CRF) as a post-processing operation to refine the segmentation maps. The CRF models the spatial dependencies between pixels, improving the alignment of segmentation boundaries with the actual edges of objects in the images. This combination of atrous convolutions, ASPP, and CRF enables DeepLabV2 to achieve great performance on various challenging benchmarks, demonstrating its efficacy in accurately segmenting objects within images while maintaining a reasonable computational complexity.

3.2. BiSeNet: Real-Time Semantic Segmentation Network

In semantic segmentation, employing a real-time network is critical for many applications where achieving both high accuracy and low latency is crucial. Traditional semantic segmentation models often struggle in balancing these requirements due to their computational complexity, which makes them unsuitable for real-time applications. To address these challenges, we exploited BiSeNet (Bilateral Segmentation Network) [13], a model designed to have an excellent compromise between accuracy and efficiency.

BiSeNet's architecture is based on two important paths: the Spatial Path (SP) and the Context Path (CP). The Spatial Path preserves the spatial resolution of the feature maps by employing three convolutional layers with stride 2 that capture fine-grained spatial details. This ensures that the network preserves the detailed spatial information of the input image. Additionally, within the Spatial Path, the Attention Refinement Module (ARM) refines features by generating attention maps that highlight important areas and suppress irrelevant information, improving the quality of the segmentation output by focusing on significant regions of the image.

The Context Path, on the other hand, provides high-level semantic context. It utilizes a lightweight model to capture context at multiple scales through a process that includes global average pooling. This path effectively expands the receptive field of the network, allowing it to aggregate contextual information across the entire image, which is vital for understanding the broader scene and the relationships between different objects within it.

To integrate the features extracted from both paths, BiSeNet employs a Feature Fusion Module (FFM), which merges the detailed spatial features from the Spatial Path

with the high-level context features from the Context Path. The FFM incorporates an attention mechanism to emphasize significant features, enhancing the overall segmentation performance selectively.

3.3. Domain Shift Problem and Solutions

Domain shift is a critical challenge in semantic segmentation, occurring due to differences in data distributions between the source domain (e.g., synthetic images) and the target domain (e.g., real-world images). Synthetic datasets, such as GTA5, are often used for training because of their large volume and easily obtainable pixel-level annotations. However, models trained on these datasets often experience significant performance drops when tested on real-world datasets like Cityscapes. This drop in performance arises because the model's features do not generalize well across domains that differ in visual characteristics, textures, lighting conditions, and scene layouts.

AdAddressing domain shift is essential for developing robust semantic segmentation models that can operate effectively in diverse real-world settings or for aligning synthetic data with real-world data. In our approach, we used techniques such as data augmentation and adversarial learning in order to improve the generalization capabilities of our model with more reliable performance in different domains. The following sections delve into the specifics of the data augmentation and adversarial learning techniques used in our project.

3.3.1 Augmentations

One primary approach to mitigating domain shift is the use of data augmentation techniques. These techniques enhance the generalization capability of models trained on synthetic datasets by introducing variations that simulate real-world conditions.

In our project, we implemented several data transformations, including Resize, Horizontal Flip, Color Jitter, Gaussian Blur, and Random Resized Crop.

Resize changes the dimensions of images to a specified size.

Horizontal Flip creates mirrored versions of images, increasing the model's robustness to variations in orientation.

Color Jitter randomly alters color parameters such as brightness, contrast, and saturation, mimicking different lighting conditions.

Gaussian Blur applies a smoothing effect to reduce noise and minor texture variations, helping the model focus on significant features.

Random Resized Crop performs random cropping and resizing, allowing the model to see different parts of images during training and promoting a better understanding of scenes at various scales.

These augmentations were selected because they alter the look of images while maintaining realism, which is critical for effective training. For instance, using Horizontal Flip makes logical sense, as it mirrors the scene without disrupting its realism. Conversely, we avoided improvements such as vertical flipping, as flipping an image vertically would produce unrealistic scenes with the sky below and the ground above, which would not be beneficial for training models intended for real-world applications.

3.3.2 Adversarial Discriminative Model

In our project, we employed an adversarial discriminative approach to tackle the challenge of domain adaptation in semantic segmentation. Our approach is based on adversarial learning in the output space, where the intuition is to make the predicted label distributions close to each other across source and target domains as [11]. This method involves a segmentation network, called generator, and a discriminator. The generator's objective is to generate similar distributions in the output space features for either source or target images capable of fooling the discriminator. Conversely, the discriminator's task is to distinguish whether the output from generator corresponds to a source or target image.

This approach is effective because the output space contains rich information, both spatially and locally. Even if images from two domains are very different in appearance, their segmentation outputs share a significant amount of similarities.

Specifically, in this method the generator is trained using segmentation loss L_{seg} on the labeled source data, while the discriminator utilizes the discriminator loss L_d to learn how to distinguish whether an image comes from the source or target domain.

The key point of this method lies in minimizing another loss, known as adversarial loss L_{adv}, to encourage the generator to produce similar segmentation distributions in both, target and source, domains. In particular the generator tries to fool the discriminator into thinking that the segmented output of a target image belongs to the source domain. During training, errors are back-propagated to the feature level from the output labels, thus adapting the features to align the segmentation distributions between the two domains.

By leveraging this adversarial approach, we were able to mitigate the negative effects of domain shift. This method effectively aligned the synthetic and real-world data distributions, ensuring more reliable and accurate segmentation outcomes in practical applications.

4. Experimental Results

In this section, we present the details and results of our implementations. After an analysis of the datasets we used, we show the evaluations of the models trained

on real-world images from the Cityscapes dataset. Subsequently, we illustrate experiments on synthetic-to-real unsupervised domain adaptation. The code and model are available at https://github.com/MazzariniTommaso/Real_time_Domain_Adaptation_in_Semantic_Segmentation

4.1. Datasets

Cityscapes. Cityscapes [2] is a large-scale urban street scene dataset, which contains high quality pixel-level annotations of 5000 images collected in street scenes from 50 different cities. All these images have a resolution of 2048×1024, in which each pixel is annotated to pre-defined 19 semantic labels belonging to 7 super categories: ground, construction, object, nature, sky, human, and vehicle (the void label is not considered).

We utilized a subset of the dataset, comprising 1572 images. Instead of using the original segmented images for training, we employed grayscale masks, where each pixel has been associated with a specific class.

GTA5. GTA5 [9] (Grand Theft Auto) consists of 24,966 images synthesized from the video game, based on the city of Los Angeles, with an original image size of 1914×1052. It has 19 classes that are compatible with the Cityscapes dataset. For our implementation, we used a subset of 2,500 images from the GTA5 dataset rather than the full set.

To optimize the training process and improve computational efficiency, we initially converted all segmented images in the dataset from color to grayscale masks. This preprocessing step allows us to directly employ the original images and their associated masks during training, without the need to perform conversion each time.

4.2. Implementation Protocol

Training Details. We employ DeepLabV2 and BiSeNet to understand what is the highest possible performance with our resources, for classic and real-time segmentation respectively. Regardless of the model and dataset, we used mini-batches of 8 images. Two different optimizers were employed: SGD with a momentum of 0.9, weight decay of 5e-4, and varying learning rates depending on the model and dataset; and Adam with a learning rate of 2.5e-4 and betas set to (0.9, 0.99). To further enhance training results, we applied the polynomial learning rate scheduler, which updates the learning rate at each iteration according to the formula from [1, 13] with a power of 0.9. Our loss function is the cross-entropy loss, where all positions and labels are equally weighted in the overall loss function, except for unlabeled pixels which are ignored.

Data Augmentations. We employed data augmentation using the *Albumentations* library. Regardless of our goal, for both GTA5 and Cityscapes, we resized images to (720, 1280) and (512, 1024) respectively to reduce input dimensions, thereby occupying less memory and accelerating the training process. To address domain shift, we experimented with various combinations of augmentations including HorizontalFlip, GaussianBlur, ColorJitter, and RandomResizedCrop.

Adversarial Learning Our approach to domain adaptation leverages adversarial learning, employing a discriminator-generator architecture. We implemented the discriminator as described in [11], while using BiSeNet as our generator. And like the referenced work, we opted to train the segmentation network and discriminator simultaneously.

For the segmentation loss, we utilized cross-entropy loss, while for the discriminator and adversarial loss, we employed BCEWithLogitsLoss, which incorporates a sigmoid layer along with binary cross-entropy loss.

A critical aspect of optimizing the generator network is maintaining a balance between segmentation and adversarial loss. Following the recommendation in [11], we applied a coefficient of 0.001 (λ_{adv}) to the adversarial loss function to achieve this balance.

The optimization process utilized different learning rates for each network component. For the discriminator, we used the Adam optimizer with a learning rate of 1e-3. The generator, also optimized with Adam, was assigned a lower learning rate of 2.5e-4.

4.3. Metrics

Intersection over Union (IoU). The Intersection over Union (IoU) is a metric used to quantify the percentage of overlap between the target mask and the predicted output mask. Specifically, IoU measures the number of pixels common to both the ground truth and prediction masks, divided by the total number of pixels present in either mask. This is mathematically represented as:

$$IoU = \frac{target \cap prediction}{target \cup prediction} \tag{1}$$

We calculated the IoU for each class label and computed the mean across all labels to determine the mean Intersection over Union (mIoU). This metric is crucial for evaluating the accuracy of our segmentation models.

FLOPs and **Number of Parameters**. To quantify the computational complexity of the models, we employ the *Fvcore* library. This library provides essential functionalities common to various computer vision frameworks. Fvcore includes a class specifically designed for calculating FLoating-Point Operations Per second (FLOPs) and counting the number of parameters utilized by a model to process one image.

- **FLOPs**: measure a model's performance based on the number of floating-point arithmetic calculations it performs. This metric is important for understanding the computational demands of a model.

- **Number of Parameters**: this metric counts the total number of trainable parameters within a model, offering insight into the model's capacity and complexity.

By leveraging Fvcore, we can accurately assess and compare the computational requirements of different models, providing a comprehensive understanding of their efficiency.

Latency. Latency refers to the time taken by a neural network to produce a prediction for a single input sample. Minimizing latency is crucial for achieving high inference speed in real-time applications. To measure the latency of a neural network in PyTorch, we use the time module to track the duration of a forward pass through the network. This measurement provides critical insights into the responsiveness and suitability of the model for time-sensitive tasks.

By systematically evaluating these metrics: mIoU for segmentation accuracy, FLOPs and parameter count for computational complexity and latency for real-time performance, we can assess the efficacy and efficiency of neural network models.

4.4. Results

Quantitative Results. As we said earlier, we trained DeepLabv2 on the Cityscapes dataset and achieved the highest performance with a mIoU of 50.09% using the Adam optimizer. In contrast, using the same optimizer, BiSeNet achieved a mIoU of 48.04% on the same dataset, as shown in Table 1.This difference in performance, while small, is attributed to the fact that BiSeNet is a smaller and less complex network capable of real-time prediction, unlike DeepLabV2.

The next step is to demonstrate the effectiveness of domain adaptation techniques in enhancing semantic segmentation performance across various scenarios. The impact of domain shift is evident when BiSeNet, trained on GTA5 and tested on Cityscapes without adaptation, suffers a drop in mIoU to 20.61%, underscoring the significant challenge posed by domain shift.

To address this issue, we initially tested various augmentation strategies. Among these, the most effective strategy (Aug 4 that includes Resize, HorizontalFlip, GaussianBlur, ColorJitter, RandomResizedCrop) achieved an mIoU of 27.43%, as shown in Table 2. To further enhance the BiSeNet's accuracy, we combined data augmentations with adversarial learning, achieving an mIoU of 30.44%, a notable increase as shown in Table 3.

Qualitative Results Figure 1 provides visual examples of the segmentation results, showing for each target image from the Cityscapes dataset the predictions of the Deeplab and BiSeNet models trained on it. The last two columns show the difference between the predictions of BiSeNet trained on GTA5 before and after our adaptation to Cityscapes. The adapted BiSeNet model shows noticeable improvements over the non-adapted version across various urban scenes. Segmentation boundaries are more defined and accurate, particularly for classes that occupy a large space in the image, like road, building, and sky. It also shows an enhanced ability to capture smaller objects which were often missed or misclassified in the non-adapted version. Overall, the adapted model demonstrates a better grasp of the overall scene layout, with more coherent segmentation across different semantic regions.

These visual results corroborate our quantitative findings, illustrating the effectiveness of our domain adaptation approach in bridging the gap between synthetic and real-world domains for semantic segmentation tasks.

Model	Optimizer	mIoU (%)	Latency (s)	FLOPs	Params (M)
DeepLabV2	SGD	43.11	0.039	0.375T	43.901
	Adam	**50.09**	0.037	0.375T	43.901
BiSeNet	SGD	32.81	0.013	25.78G	12.582
	Adam	**48.04**	0.013	25.78G	12.582

Table 1. Accuracy and parameter analysis of models trained and validated on Cityscapes with two different types of optimizers.

5. Conclusion

This study demonstrated that domain adaptation techniques can substantially enhance real-time semantic segmentation performance when transitioning from synthetic to real-world environments. Specifically, our findings indicate that employing data augmentation in conjunction with adversarial learning effectively mitigates the domain shift issue, resulting in a notable increase in mean Intersection over Union (mIoU) from 20.61% to 30.44%. Despite the significant improvement in our adapted model, there remains a performance gap compared to models trained directly on real-world data. This highlights the necessity for further research, which could involve exploring various hyperparameter configurations, optimizers, loss functions, or alternative domain adaptation methodologies to bridge this gap and achieve more robust performance in real-world applications.

Experiments	mIoU	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle
No Aug	20.61	33.90	11.14	61.35	7.34	5.31	15.29	9.30	6.19	75.13	3.18	63.08	30.09	0.21	63.14	4.63	1.14	0.01	1.11	0.03
Aug 1	26.36	67.42	26.68	69.06	12.94	6.55	19.87	13.57	12.00	77.17	16.70	73.61	36.25	0.50	54.46	6.90	5.43	0.06	1.45	0.18
Aug 2	20.96	33.27	5.81	69.44	5.42	12.72	15.93	11.67	6.27	71.47	3.33	57.74	32.44	2.05	62.22	2.32	0.87	0.00	5.06	0.20
Aug 3	26.67	55.87	32.02	74.90	17.23	14.49	24.44	14.72	9.00	79.29	20.23	76.72	39.88	3.28	29.76	8.85	1.95	0.02	3.19	0.84
Aug 4	**27.43**	67.21	25.57	72.32	16.63	13.20	19.03	18.39	11.94	79.28	20.19	74.09	36.04	3.82	44.91	7.89	3.42	0.36	3.57	3.32

Table 2. Performance comparison (in %) of different data augmentations applied to reduce domain shift between GTA5 and Cityscapes.

	mIoU	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle
Before DA	20.61	33.90	11.14	61.35	7.34	5.31	15.29	9.30	6.19	75.13	3.18	63.08	30.09	0.21	63.14	4.63	1.14	0.01	1.11	0.03
Ours	**30.44**	86.37	21.72	79.69	22.04	15.07	22.04	18.39	11.10	77.32	21.72	69.98	32.51	3.06	73.55	12.53	5.97	0.43	3.98	0.95

Table 3. Performance comparison (in %) to show the accuracy improvement using domain adaptation techniques on BiSeNet trained on GTA5 and tested on Cityscapes.

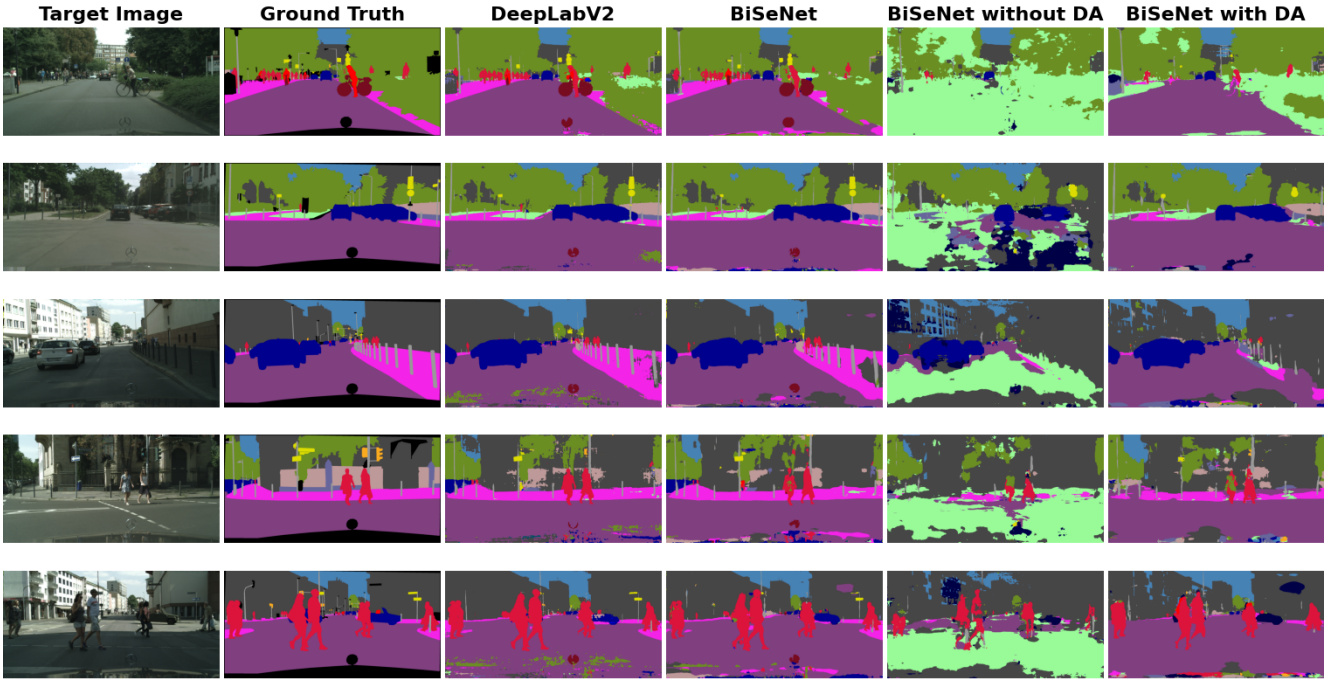

Figure 1. The figure shows from left to right the original image, ground truth, DeepLabV2 and BiSeNet results trained on Cityscapes, and BiSeNet results trained on GTA5 with and without domain adaptation.

References

[1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *CoRR*, abs/1606.00915, 2016. 1, 2, 3, 5

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. *CoRR*, abs/1604.01685, 2016. 1, 5

[3] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016. 2

[4] Shijie Hao, Yuan Zhou, and Yanrong Guo. A brief survey on semantic segmentation with deep learning. *Neurocomputing*, 406:302–321, 2020. 2

[5] Hong Liu, Mingsheng Long, Jianmin Wang, and Michael Jordan. Transferable adversarial training: A general approach to adapting deep classifiers. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4013–4022. PMLR, 09–15 Jun 2019. 3

[6] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *CoRR*, abs/1411.4038, 2014. 2

[7] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31, 2018. 3

[8] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *CoRR*, abs/1606.02147, 2016. 2

[9] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *European Conference on Computer Vision (ECCV)*, volume 9906 of *LNCS*, pages 102–118. Springer International Publishing, 2016. 1, 5

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. 2

[11] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. *CoRR*, abs/1802.10349, 2018. 4, 5

[12] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. *CoRR*, abs/1702.05464, 2017. 3

[13] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation, 2018. 1, 2, 3, 5

[14] Sicheng Zhao, Xiangyu Yue, Shanghang Zhang, Bo Li, Han Zhao, Bichen Wu, Ravi Krishna, Joseph E. Gonzalez, Alberto L. Sangiovanni-Vincentelli, Sanjit A. Seshia, and Kurt Keutzer. A review of single-source deep unsupervised visual domain adaptation. *CoRR*, abs/2009.00155, 2020. 2, 3