Real-time Domain Adaptation in Semantic Segmentation

Nicola Biagioli
Politecnico di Torino
s344677@studenti.polito.it

Roffinella Andrea
Politecnico di Torino
s349557@studenti.polito.it

Stinà Giovanni
Politecnico di Torino
s332085@studenti.polito.it

Superno Falco Leonardo
Politecnico di Torino
s338685@studenti.polito.it

Abstract

Semantic segmentation in real-time systems is essential for applications such as autonomous driving. However, models trained on synthetic data often fail to generalize to real-world images due to domain shift, making unsupervised domain adaptation a critical task.

This work focuses on domain adaptation techniques for real-time semantic segmentation. We trained and evaluate a classical model (DeepLabV2) and a real-time model (BiSeNet) on the real-world Cityscapes dataset to establish reference performance. We then trained BiSeNet on the synthetic GTA5 dataset and evaluated its generalization to Cityscapes, highlighting the impact of domain shift.

To address this, we applied data augmentation and adversarial training with a domain discriminator. Additionally, we proposed a multi-level adversarial framework that introduces discriminators at three semantic levels (spatial, context1, and context2) based on intermediate feature maps extracted via convolutional layers.

Finally, we explored alternative segmentation loss functions by combining cross-entropy with focal and dice losses. We also performed loss-weight tuning to balance their contribution during training. The models were evaluated using mean Intersection over Union (mIoU) for segmentation quality, alongside latency, number of floating-point operations per second (FLOPs), and parameter count to assess computational efficiency and real-time feasibility.

1. Introduction

Semantic segmentation is at the center of numerous computer vision applications, where understanding both the content and structure of a scene is essential. From medical diagnostics to autonomous driving, segmentation models enable high-accuracy interpretation of complex visual data. Over the past decade, deep convolutional networks have led to major advances in this task, particularly in settings where accuracy takes precedence over inference time.

However, deploying segmentation models in real-time environments introduces new challenges. In real-time scenarios, such as autonomous driving, computational efficency is the priority. To this end, lightweight architectures such as BiSeNet have been proposed to offer a practical trade-off between speed and segmentation performance.

One issue in training robust segmentation models remains the availability of labeled data. High-quality annotations at pixel level are expensive to produce, especially in real-world conditions. Synthetic datasets, automatically generated from simulated environments, represent a scalable alternative. When models are trained on synthetic data, their accuracy usually drops when applied to real images, this phenomenonis due to domain shift. Domain shift happens when the training data and the test data look different.

In this project, we investigate methods to enhance the transferability of real-time segmentation networks from synthetic to real-world domains. To set a performance reference, we train and evaluate of DeepLabV2 [1] and BiSeNet [7] on the real-world Cityscapes dataset [2]. We then expose the domain shift by training BiSeNet on the synthetic GTA5 [4] dataset and evaluating it on Cityscapes, observing the expected drop in performance.

To mitigate the effect of domain shift, we adopt data augmentation strategies and implement adversarial training using a domain discriminator to promote domain-invariant representations. Following from this, we propose a multi-level adversarial approach adding to the single-level adversarial model a second layer based on an intermediate feature between: spatial, context1, and context2. Although the overall segmentation score slightly decreases compared to the single-discriminator baseline, our analysis reveals improvements in class-specific performance. In particular spatial features improve compact objects with well-defined edges and context features enhance the performance of

larger structural classes.

Finally, we explored alternative loss functions, building upon the single-level adversarial setup, as it provided the strongest baseline performance in our experiments. In addition to standard cross-entropy, we incorporate focal and dice losses and perform fine-tuning of their relative contributions. This setup helps us understand how adjusting the weights of different losses affects training stability and the accuracy of the segmentation results.

The structure of this report is as follows: Section 2 presents an overview of key works in semantic segmentation and domain adaptation. Section 3 outlines the implementation of the baseline models and proposed techniques. Section 4 presents the experimental setup and performance results. Section 5 summarizes the findings and suggests directions for future improvements.

2. Related Works

This section reviews the main contributions that form the basis of our work, focusing on semantic segmentation networks and adversarial approaches for unsupervised domain adaptation.

2.1. Semantic Segmentation Architectures

Among the key architectures for semantic segmentation, DeepLabV2 introduced a combination of dilated convolutions and Atrous Spatial Pyramid Pooling (ASPP) to effectively capture multi-scale context without sacrificing spatial resolution. Additionally, the use of fully connected Conditional Random Fields (CRFs) as post-processing allowed for to get clearer edges between objects [1]. These innovations made DeepLabV2 a strong baseline for high-accuracy segmentation, though with high computational demands.

To address real-time constraints, BiSeNet was proposed with a dual-branch design: a Spatial Path that preserves fine-grained details and a Context Path that captures semantic information through rapid downsampling and global average pooling [7]. These are fused via a Feature Fusion Module and refined with Attention Refinement Modules, achieving a reasonable trade-off between speed and accuracy. Given its efficiency, BiSeNet is especially suited for real-time tasks such as autonomous driving.

2.2. Synthetic vs Real-world Data

Training deep segmentation models requires large-scale pixel-level annotated data, which is both costly and time-consuming to collect. Real-world datasets like Cityscapes provide high-quality labels but require up to 90 minutes per image for manual annotation [2]. To mitigate this issue, synthetic datasets have been proposed as a scalable alternative.

One example is GTA5, a dataset generated from the photorealistic video game Grand Theft Auto V without engine access [4]. Through a method of detouring, rendering metadata is captured and used to propagate semantic labels across frames efficiently. This enabled the authors to label nearly 25,000 images in under 50 hours, faster than manual annotation on real data.

However, synthetic images often differ significantly from real-world counterparts in terms of texture, lighting, and noise. This discrepancy causes models trained on synthetic data to underperform when deployed in real-world scenarios.

2.3. Unsupervised Domain Adaptation and Adversarial Learning

A key challenge in using synthetic datasets is the domain shift which refers to the performance drop observed when models trained on synthetic data are applied to real-world images. To bridge this gap, numerous unsupervised domain adaptation (UDA) techniques have been proposed. Among them, adversarial learning methods have shown promising results in semantic segmentation.

Tsai et al. [6] proposed a novel adversarial approach that operates not in the feature space, but in the output space of the segmentation network. Their insight is that, although images differ across domains, the structural layout of segmentation maps is more domain-invariant.

Their architecture includes a discriminator that distinguishes between predicted segmentation outputs from source and target domains, encouraging the segmenter to generate domain-invariant predictions by "fooling" the discriminator. This technique yields improved generalization without requiring labels in the target domain.

From a theoretical perspective, this method belongs to the broader class of adversarial discriminative approaches, which minimize domain discrepancy through a min-max game between generator and discriminator.

Furthermore, to address limitations of single-level adversarial adaptation, they introduce a multi-level adversarial learning strategy, where discriminators are attached to intermediate layers of the segmentation model (e.g., conv4 and conv5 in DeepLabV2). This allows adaptation to take place at multiple semantic depths, improving alignment between source and target domains and enhancing robustness. Their experiments on GTA5 $\rightarrow$ Cityscapes and SYNTHIA $\rightarrow$ Cityscapes confirm that this approach outperforms previous UDA methods in terms of mean IoU.

This adversarial framework has inspired our own multi-discriminator strategy applied to the internal feature maps of BiSeNet (spatial, context1, context2), extending the concept of multi-level alignment to real-time segmentation networks.

2.4. Alternative Loss Functions

In addition to architectural and domain adaptation advances, recent work has focused on improving the loss functions used during training, especially in scenarios characterized by severe class imbalance or noisy boundaries. Lin et al. [3] introduced Focal Loss to mitigate the dominance of well-classified examples during training. By dynamically scaling the cross-entropy loss, Focal Loss emphasizes harder, misclassified examples.

Complementarily, Dice-based losses have been widely adopted in medical imaging and are gaining traction in general segmentation due to their ability to directly optimize overlap-based metrics. Sudre et al. [5] proposed the Generalized Dice Loss, which re-weights class contributions based on inverse frequency, making it effective under extreme imbalance.

3. Methods

This section presents the methodologies adopted in our project, we explore two advanced neural network architectures: DeepLabV2 and BiSeNet. To address the domain shift challenge, we explore a data augmentation strategy alongside an adversarial domain adaptation approach. Finally, we aim to enhance performance by training the model using a combination of differently weighted loss functions.

3.1. DeepLabV2: Classical Semantic Segmentation Network

The main tasks covered by DeepLabV2 are to limit the reduction in feature resolution, manage objects at multiple scales, and solve the problem of the reduction of localization accuracy due to the property of invariance. When in the architecture of a model we have multiple combinations of max-pooling and down-sampling, the spatial resolution is significantly reduced. The solution implemented in DeepLabV2 is to use the Atrous convolution that enlarges the receptive field, without increasing the computational cost. Indeed, putting some zeros in the convolutional kernel, we can create a filter with a larger field that can see a larger context. Based on the number of holes that we add to the kernel, we can decide the dimension of the field of view. The Atrous Spatial Pyramid Polling (ASPP) is used to manage objects at different scales and is also computationally efficient. ASPP uses multiple parallel Atrous convolutional layers with different sampling rates, in this way it can obtain smaller and larger views of the same image. Another important topic is invariance that can be defined as the ability to recognize things with different positions and orientations, so it is an important property in the image classification field. But, on the other hand, it reduces the localization accuracy, so in DeepLabV2 the Conditional Random Field (CRF) is introduced to solve this problem. CRF connects a pixel to all the other pixels and clarifies the edges of things working at a pixel-level.

3.2. BiSeNet: Real-Time Semantic Segmentation Network

In the context of real-time semantic segmentation, ensuring both high accuracy and low latency is critical. Conventional segmentation networks often struggle to meet these dual demands due to their heavy computational requirements and the trade-offs between spatial detail and semantic context.

To address this challenge, we adopt the Bilateral Segmentation Network (BiSeNet), a model specifically designed to strike an optimal balance between accuracy and computational efficiency. BiSeNet achieves this by decomposing the learning objective into two parallel, specialized pathways:

Spatial Path (SP): The Spatial Path captures detailed spatial features using a shallow stack of three convolutional layers with stride 2. While it progressively reduces resolution to 1/8 of the input, it maintains enough detail to preserve object boundaries and fine structures.

Context Path (CP): Complementing the SP, the Context Path is optimized to capture rich semantic context and global information. It utilizes a lightweight backbone network (e.g., ResNet18) **SE è L'ESEMPIO DEL PAPER CITIAMOLO, ALTRIENTI ELIMINIAMO L'ESEMPIO** that rapidly downsamples the input, thereby enlarging the effective receptive field. To further improve the model's understanding of the overall scene, a global average pooling layer is added at the end of the backbone network. This layer collects information from the entire image, helping the network to recognize larger patterns and relationships between different objects within the scene. In addition to the global pooling, the Context Path integrates one or more Attention Refinement Modules (ARM) after intermediate stages of the backbone. Each ARM uses global average pooling to compute a channel-wise attention vector, which is then used to reweight the feature maps. This mechanism helps the network focus on the most informative features, refining the representation before the final fusion with the Spatial Path.

Feature Fusion Module (FFM): Since SP and CP produce features at different abstraction levels (low-level spatial detail and high-level semantic context), a dedicated fusion mechanism is essential. The FFM addresses this by first concatenating the outputs of both paths, normalizing the combined feature map, and then applying a lightweight attention-based reweighting mechanism. This enables the network to adaptively prioritize informative features during fusion, thereby enhancing overall segmentation quality.

This dual-pathway design, along with attention-driven fusion, enables BiSeNet to achieve real-time performance

(e.g., 105 FPS on Cityscapes [2]) without sacrificing segmentation quality, making it ideal for applications requiring both speed and precision.

3.3. Domain Shift Problem and Solutions

Semantic segmentation of real-world images typically requires large-scale datasets with dense pixel-level annotations, which are costly and time consuming to obtain. To mitigate this limitation, synthetic dataset such as GTA5 provide automatic, accurate and large-scale semantic annotations. So they can be used to train deep segmentation models on a large amount of data. However, synthetic data introduce a new challenge: Domain shift. This is due to the discrepancy between the source domain and the target domain in visual appearance, scene layouts and data distributions. So we can see a drop in performance when models trained on synthetic data are evaluated on real-world images. To address this issue, we explored different techniques to help the model to generalize better, reduce the impact of the domain shift and increase the transferability of learned features between the source and the target domain. In particular we implemented the data augmentation and adversarial learning approach with some extensions as the multi-level adversarial learning and additional losses used in combination with the cross-entropy loss as the focal the dice loss.

3.3.1 Augmentations

One effective strategy to reduce domain shift between synthetic and real-world data is through data augmentation. These methods improve a model's ability to generalize by introducing controlled variations that reflect the diversity encountered in real environments. In our project, we applied a set of augmentations including Resize, Horizontal Flip, Rotation, Color Jitter, Gaussian Blur, and Random Resized Crop.

- **Resize** standardizes image dimensions, ensuring consistency in input size for efficient training.
- **Horizontal Flip** generates mirrored versions of the original image, helping the model handle objects appearing in different orientations.
- **Rotation** introduces controlled angular variations, enabling the model to better manage different viewing angles and orientations commonly seen in real-world scenes.
- **Color Jitter** introduces random modifications to brightness, contrast, and saturation, simulating variable lighting and color conditions found in real scenes.
- **Gaussian Blur** softens image details by reducing high-frequency noise, encouraging the model to focus on broader structures rather than fine-grained textures.
- **Random Resized Crop** randomly crops and resizes portions of the image, enabling the model to learn from different spatial contexts and improving scale invariance.

These enhancements were deliberately chosen to modify the appearance of the image while preserving semantic integrity, which is essential for realistic training. For example, applying Horizontal Flip maintains the natural structure of a scene, whereas transformations like Vertical Flip were excluded because they create implausible images, potentially confusing the model during training for real-world deployment.

3.3.2 Adversarial Discriminative Model

In our project, we addressed the problem of domain adaptation in semantic segmentation using an adversarial discriminative strategy. This method operates in the output space, with the main idea being to align the predicted label distributions of source and target domains, following the intuition introduced in [6]. The framework consists of two core components: the generator, that is a segmentation network, and the discriminator. The goal of the generator is to make segmentation predictions of source and target images close to each other, instead the role of the discriminator is to distinguish whether the output from generator corresponds to a source or target image. The semantic segmentation outputs contains rich spatial and contextual information, so also images of different sources that look very different in appearance will be structurally similar, so this strategy can be very powerful to reduce the domain shift. Training proceeds as follows: The discriminator is trained with a standard classification loss $\mathcal{L}_d$, encouraging it to correctly distinguish between source and target outputs. The generator is supervised with a segmentation loss $\mathcal{L}_{seg}$, computed on the labeled source data to ensure accurate predictions and an adversarial loss $\mathcal{L}_{adv}$, which guides the training of the generator using target data as reported in 1.

$$\mathcal{L}(I_s, I_t) = \mathcal{L}_{\text{seg}}(I_s) + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}}(I_t), \tag{1}$$

$\mathcal{L}_{\text{adv}}$ defines a measure of divergence between the probability distributions of the segmentation outputs for source and target images. We give $\mathcal{L}_{\text{adv}}$ a discriminator output of a target image and the label the discriminator would assign to source images. The generator tries to fool the discriminator, maximizing the probability that it classifies target outputs as if they were from the source domain.

3.3.3 Multi-Level Adversarial Model

To achieve better domain adaptation, we extend the single-level adversarial framework by aligning feature distributions at multiple levels. As in the single-level implementation, we use the BiSeNet segmentation head to extract

feature maps. This multi-level extension targets the adaptation of both high-level and low-level features, especially those far from the output layer, which may exhibit different distributions.

The total loss function used in the multi-level adversarial approach is defined as:

$$\mathcal{L}(I_s, I_t) = \sum_i \lambda^i_{\text{seg}} \mathcal{L}^i_{\text{seg}}(I_s) + \sum_i \lambda^i_{\text{adv}} \mathcal{L}^i_{\text{adv}}(I_t), \quad (2)$$

where the index i refers to the level within the generator (i.e., segmentation network).

In addition to the final output, we investigated three different levels for adversarial alignment:

The intermediate feature map of **context1**, extracted from an early block in the context path, provides limited semantic abstraction but retains some local spatial details.

The second level, **context2**, corresponds to a deeper context feature with a larger receptive field, which captures more abstract and semantic information about the scene.

Finally, **spatial** refers to the output of the Spatial Path, a shallow CNN branch designed to preserve spatial resolution and fine details, which is particularly beneficial for segmenting small or thin structures.

3.3.4 Loss extension: focal loss and dice loss

While standard cross-entropy loss is commonly used for image segmentation, it is known to struggle in scenarios with significant class imbalance, particularly when foreground regions (e.g., bikes, riders, or small objects) occupy only a small fraction of the image. To address this issue, we extend the baseline loss with two complementary formulations: **dice loss** and **focal loss**.

Dice Loss is derived from the Dice similarity coefficient, which quantifies the overlap between predicted and ground truth masks. For each class c, it is defined as:

$$\text{Dice}_c = \frac{2\sum_{i=1}^{N} p_{i,c} r_{i,c} + \epsilon}{\sum_{i=1}^{N} p_{i,c} + \sum_{i=1}^{N} r_{i,c} + \epsilon} \quad (3)$$

where:

- $p_{i,c} \in [0, 1]$ is the predicted probability that pixel i belongs to class c,
- $r_{i,c} \in \{0, 1\}$ is the one-hot encoded ground truth label for pixel i and class c,
- N is the total number of pixels,
- ϵ is a small constant (e.g., 10^{-6}) added for numerical stability.

Multiclass Dice Loss. In multi-class segmentation settings with C classes, the Dice Loss is computed by averaging the per-class Dice Coefficients and subtracting the result from 1:

$$\mathcal{L}^{\text{multi}}_{\text{Dice}} = 1 - \frac{1}{C} \sum_{c=1}^{C} \text{Dice}_c \quad (4)$$

This formulation ensures that each class contributes equally to the total loss, regardless of its pixel frequency in the dataset. It is particularly beneficial in class-imbalanced scenarios, where dominant classes might otherwise overshadow rare ones.

Focal Loss. Focal Loss addresses class imbalance by down-weighting easy examples and focusing learning on hard, misclassified ones. It modifies the cross-entropy loss with a modulating factor:

$$\mathcal{L}_{\text{Focal}} = -(1 - p_t)^{\gamma} \log(p_t) \quad (5)$$

where:

- p_t, probability of the true pixel label
- $\gamma \geq 0$ is the focusing parameter that controls the down-weighting of easy examples.

In our implementation, we set $\gamma = 2$. This loss significantly reduces the relative loss for well-classified background pixels, which are typically over-represented in segmentation datasets.

Combined Loss. We optionally combine the different loss terms as weighted sums:

$$\mathcal{L}^{(1)}_{\text{Total}} = \lambda_{\text{CE}} \mathcal{L}_{\text{CE}} + \lambda^{\text{multi}}_{\text{Dice}} \mathcal{L}^{\text{multi}}_{\text{Dice}} \quad (6)$$

$$\mathcal{L}^{(2)}_{\text{Total}} = \lambda_{\text{CE}} \mathcal{L}_{\text{CE}} + \lambda_{\text{Focal}} \mathcal{L}_{\text{Focal}} \quad (7)$$

where $\lambda_{\text{CE}}, \lambda^{\text{multi}}_{\text{Dice}}, \lambda_{\text{Focal}}$ are scalar weights controlling the contribution of each component.

We first evaluate the combination of Cross-Entropy and multi-class Dice Loss to exploit both pixel-wise accuracy and region-level overlap. We then experiment with Cross-Entropy and Focal Loss to address class imbalance and hard-to-classify regions in the segmentation task.

4. Experimental Results

This section reports the experimental setup and results of our study. We begin by describing the datasets used for training and evaluation. We then analyze the performance of BiSeNet and DeepLabV2 in supervised and unsupervised settings, comparing different adaptation strategies and loss functions.

4.1. Datasets

Cityscapes [2] is a well-established benchmark for semantic segmentation of urban street scenes, consisting of 5,000 finely annotated images collected across 50 European cities. All images are high-resolution (2048×1024) and annotated with 19 semantic classes. While the dataset is highly realistic and diverse, the fine-level annotation process is extremely labor-intensive, reportedly taking around 90 minutes per image.

We conducted our experiments using a curated subset of 1,572 training images along with their corresponding grayscale label masks, where each pixel directly encodes a class index.

GTA5 [4] is a synthetic dataset created from the photorealistic open-world video game Grand Theft Auto V. It contains 24,966 high-resolution (1914×1052) frames annotated with dense pixel-level labels compatible with the Cityscapes format. Unlike other synthetic datasets that require access to rendering engines or manual annotations, GTA5 was built by intercepting communication between the game and the graphics hardware without access to the game source code.

In our experiments, we selected a subset of 2,500 images from the full GTA5 dataset. We converted the RGB label masks to grayscale label masks as a preprocessing step, rather than on the fly, in order to reduce the computational load.

4.2. Implementation Protocol

Training Details. We trained both DeepLabV2 and BiSeNet to establish supervised baselines on Cityscapes. DeepLabV2 was used as a high-accuracy reference model, while BiSeNet served as our real-time architecture. For all experiments, we used mini-batches of 6 images and trained with Adam optimizer using as initial learning rate 2.5e-4, then the learning rates were scheduled with the poly policy [**?**, **?**], decreasing as $(1 - \frac{iter}{max\_iter})^{0.9}$. The base segmentation loss was standard cross-entropy, applied only to labeled pixels.

Data Augmentations. To enhance generalization and address domain shift, we employed data augmentation via the `Albumentations` library. Input images were resized to 1280×720 for GTA5 and 1024×512 for Cityscapes. Augmentations included resize, horizontal flip, rotation, color jitter, gaussian blur, random resized crop. These transformations helped diversify the synthetic data domain and simulate some of the variability observed in real-world images.

Adversarial Domain Adaptation. For domain adaptation experiments, we implemented an adversarial framework inspired by [**?**], where a segmentation network (generator), implemented using BiSeNet, is trained jointly with a domain discriminator. The discriminator is applied to the predicted segmentation maps and trained to distinguish between source and target domain outputs. We used `BCEWithLogitsLoss` as the adversarial loss for the discriminator, and applied a coefficient $\lambda_{adv} = 0.001$ to balance the adversarial and segmentation losses in the generator's optimization.

Multi-level Adversarial Extension. We then extended the adversarial adaptation framework to operate at multiple semantic levels within BiSeNet. Specifically, we extracted three types of intermediate features from the model: *spatial*, *context1*, and *context2*, corresponding to different stages of the network. Each feature map was passed through a `Conv2d` layer and fed to a separate discriminator. The goal was to enforce domain invariance at multiple abstraction levels.

Alternative Loss Functions. In a separate set of experiments, using the standard single-discriminator setting, we investigated whether alternative loss functions could improve segmentation quality and training robustness. In particular, we augmented the standard cross-entropy loss with two widely used additions. First, we experimented with the focal loss, then, we explored the use of dice loss. To better understand the trade-offs involved, we implemented a fine-tuning procedure over the relative weights of each loss term.

4.3. Metrics

To comprehensively evaluate both the effectiveness and efficiency of our semantic segmentation models, we employ a combination of accuracy and computational performance metrics.

Intersection over Union (IoU) serves to measure the quality of segmentation by quantifying the overlap between the predicted mask and the ground truth. It is defined as the ratio between the intersection and the union of the two masks, formally expressed as:

$$IoU = \frac{target \cap prediction}{target \cup prediction} \tag{8}$$

For each semantic class, we compute its IoU score, and then report the average across all classes, known as mean Intersection over Union (mIoU).We used this as a primary indicator of segmentation accuracy.

Computational complexity is assessed using the `fvcore` library, which provides standardized tools for analyzing model architecture. Specifically, we measure:

FLOPs, or floating-point operations, which reflect the total number of arithmetic operations needed to process a single input image.

Number of parameters, which counts the total trainable weights in the model. This metric gives insight into the model's representational capacity and storage footprint.

Latency is defined as the average inference time required to process a single image. It is computed in PyTorch using the `time` module, ensuring GPU synchronization to obtain precise measurements. This metric is fundamental for real-time applications, where prediction speed is a critical constraint.

4.4. Results

We first trained DeepLabV2 and BiSeNet on Cityscapes in a fully supervised setting to establish a baseline. DeepLabV2 reached 51.69% mIoU, while BiSeNet achieved 46.80% with lower latency and complexity (Table 1).

Model	mIoU (%)	Latency (s)	FLOPs	Params (M)
DeepLabV2	51.69	0.036	375G	43.90
BiSeNet	46.80	0.010	25.78G	12.58

Table 1. Accuracy and parameter analysis of DeepLabV2 and BiSeNet trained and evaluated on Cityscapes.

Training BiSeNet on GTA5 and testing on Cityscapes without domain adaptation (DA) resulted in a mIoU drop to 17.41%, highlighting the domain shift. Data augmentation improved performance only marginally, with the best configuration (Aug3) reaching 17.82% mIoU (Table 2).

Applying adversarial training on the output space (single-level) raised mIoU to 28.11% (Table 3). We then explored multi-level adversarial learning by attaching discriminators to BiSeNet's internal branches (spatial, context1, context2). The spatial branch achieved the best result (26.89%), enhancing small-object segmentation, while context branches favored larger structures.

Lastly, we tested focal and dice losses within the single-level framework. Adjusting loss weights ($\lambda_{\text{CE}}, \lambda_{\text{Focal}}, \lambda_{\text{Dice}}$) as in Table 4, we obtained a peak mIoU of 26.73% with dice loss, with gains on rare classes like rider and bus (Table 5).

Parameter	Values
λ_{CE}	$\{0.1, 0.5, 1.0\}$
λ_{Dice}	$\{0.1, 0.5, 1.0, 2.0\}$
λ_{Focal}	$\{0.1, 0.5, 1.0, 2.0\}$

Table 4. Loss weighting parameters explored during training.

5. Conclusion

This work investigated domain adaptation for real-time semantic segmentation from synthetic (GTA5) to real-world (Cityscapes) data. Starting from a baseline mIoU of 17.41% without adaptation, we improved performance to 28.11% using adversarial training combined with data augmentation.

We extended the standard adversarial approach to a multi-level setting by attaching discriminators to BiSeNet's internal branches. While overall mIoU slightly decreased, results showed that spatial features help with small objects and context features with larger structures.

We also evaluated alternative loss functions within the single-level setup. Incorporating Dice Loss led to the best performance (26.73% mIoU), especially benefiting underrepresented classes.

Although a gap remains with fully supervised training on real data, our results highlight the effectiveness of adversarial adaptation and tailored loss functions in reducing domain shift. Future work may explore better loss balancing or self-supervised signals to further close the gap.

Experiments	mIoU	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle
Without DA	17.41	12.33	8.18	53.89	3.70	2.65	13.93	6.78	6.26	75.34	2.33	58.97	24.68	0.20	55.48	4.37	1.52	0.00	0.21	0.00
Aug 1	16.24	11.91	10.72	52.84	2.70	1.68	11.58	9.76	3.82	71.18	6.09	55.12	27.14	0.28	42.10	1.21	0.01	0.00	0.00	0.34
Aug 2	16.08	5.83	3.10	54.23	4.16	3.29	11.93	8.02	7.18	71.68	2.02	48.47	23.53	0.30	53.80	5.65	2.05	0.00	0.14	0.19
Aug 3	**17.82**	17.66	1.96	64.33	4.70	5.05	12.68	7.64	6.24	69.90	2.17	55.95	27.17	0.97	52.18	6.08	0.01	0.00	0.28	3.68
Aug 4	17.68	12.07	3.14	62.65	1.14	8.26	18.17	12.58	10.32	69.35	2.03	44.34	30.48	0.86	55.01	4.49	0.88	0.01	0.19	0.00

Table 2. Performance (in %) of data augmentation techniques designed to reduce the domain gap between the synthetic GTA5 dataset and the real-world Cityscapes benchmark.

Experiments	mIoU	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle
Single Level Adv	**28.11**	84.33	14.58	77.72	20.12	7.15	20.24	15.65	6.04	77.60	13.97	76.67	28.07	0.82	77.48	10.77	2.46	0.00	0.38	0.00
Multi Level Adv Context1	24.55	69.41	14.80	75.45	18.58	10.72	20.98	11.77	6.09	65.23	9.80	55.43	25.93	0.45	71.17	8.58	1.65	0.00	0.46	0.00
Multi Level Adv Context2	26.64	87.78	14.39	77.13	17.68	9.92	11.02	7.37	4.40	77.50	18.34	73.51	21.08	0.39	72.92	9.94	1.75	0.03	0.48	0.61
Multi Level Adv Spatial	26.89	86.35	15.44	77.67	19.45	8.88	22.87	11.70	6.63	69.46	17.92	58.58	27.26	0.84	74.02	11.18	1.50	0.05	0.10	0.04

Table 3. Performance comparison (in %) of single-level and multi-level adversarial learning methods applied to mitigate domain shift in synthetic-to-real semantic segmentation (GTA5 to Cityscapes).

Experiments	mIoU	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle
Single Level Adv	**28.11**	84.33	14.58	77.72	20.12	7.15	20.24	15.65	6.04	77.60	13.97	76.67	28.07	0.82	77.48	10.77	2.46	0.00	0.38	0.00
Focal Loss	26.00	67.46	17.59	75.88	13.33	8.16	21.80	12.18	11.30	76.91	7.24	72.67	28.29	0.73	69.65	7.32	1.12	0.00	2.19	0.14
Dice Loss	26.73	83.11	9.94	73.35	7.76	7.63	19.54	15.28	6.02	77.11	13.25	73.72	30.49	4.26	70.82	7.13	3.33	0.00	2.34	2.84

Table 5. Quantitative results (in %) of alternative loss functions (focal and dice)integrated into the single-level adversarial training framework for domain adaptation from GTA5 to Cityscapes.

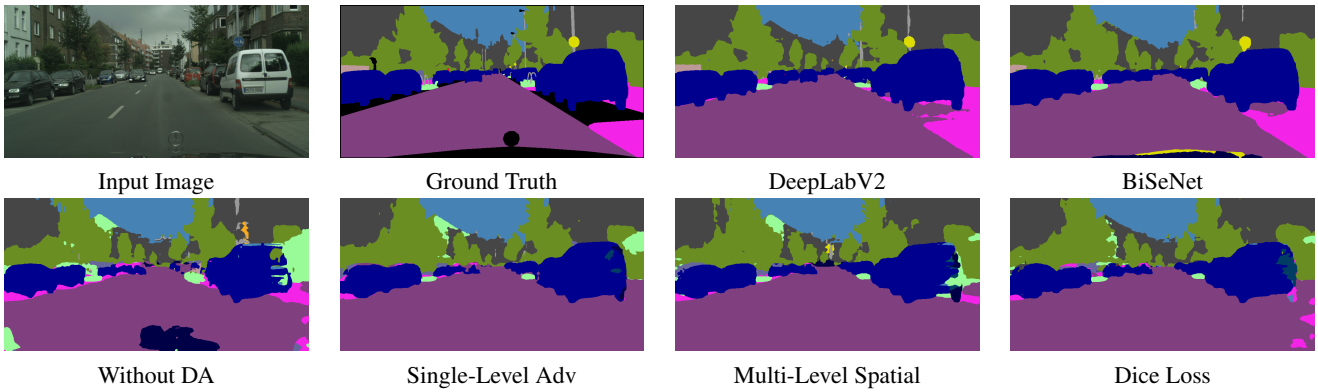

Figure 1. Qualitative comparison of segmentation results on a sample from the Cityscapes dataset. The top row shows the input image, ground truth, and predictions from DeepLabV2 and BiSeNet. The bottom row illustrates the impact of various domain adaptation techniques.

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

[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, 2, 4, 6

[3] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017. 3

[4] 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, 2, 6

[5] Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sébastien Ourselin, and M. Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. *CoRR*, abs/1707.03237, 2017. 3

[6] 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. 2, 4

[7] 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