# CLADA: Contrastive Learning for Adversarial Domain Adaptation

Richard Greene and Kevin McGuinness

*School of Electronic Engineering, Dublin City University*

**Abstract**

This paper focuses on the challenging problem of unsupervised domain adaptation of synthetic data for the semantic segmentation task of autonomous driving scenes. It is motivated by the generative adversarial methods that apply image-to-image translation by learning a mapping between the source and target domains. Fully supervised training of deep models for semantic segmentation do not generalize well to unseen target data. By applying domain adaptation, a model can be fit that generalizes to the target domain. Previous work has shown that combining generative adversarial networks with cycle consistency is effective for mapping images between domains, which can then be used to train a domain invariant semantic segmentation model. However, this requires additional networks to implement the cycle-consistency constraint. This paper proposes replacing this with a more efficient contrastive objective for the semantic segmentation task. By reducing the training time and computational resources, more complex end-to-end domain adaptation architectures may be used.

**Keywords:** Deep Learning, Generative Adversarial Network, Domain Adaptation, Contrastive Learning

## 1 Introduction

Deep convolutional neural networks have produced impressive results in many computer vision tasks, such as image classification, segmentation, object detection, and image generation. Semantic segmentation, in particular, has been substantially improved in recent years and has several important applications, including autonomous driving systems. This research focuses on the semantic segmentation of dashcam images captured by a vehicle to assign a semantic label to each pixel, e.g. road, vehicle, building, pedestrian, etc.

Supervised learning is the most common approach to fitting a semantic segmentation model. Using a large labelled dataset, a model can be trained to classify each pixel of the input based on the labels provided. Generating the large datasets required for autonomous driving perception tasks is, however, time consuming and expensive, due to the time and cost associated with manually annotating these datasets with dense pixel-level labels. This supervised approach also assumes both training data and unseen test data are drawn from the same distribution. If this assumption is violated, the model trained on the source data will fail to generalize to unseen test data due to the differences between the two distributions. This is commonly referred to as domain shift.

Domain adaptation is a type of transfer learning that attempts to reduce this domain shift, with the aim of transferring knowledge learnt from labelled data in a sourced domain to another target domain, where labelled data may be unavailable. By leveraging 3D graphics engine technology, commonly used for game development, large amounts of synthetic training images and corresponding labels can be generated in a fraction of the time and cost compared to collecting and hand labelling real world dashcam imagery. A semantic segmentation model can be fit using this synthetic dataset, and domain adaptation techniques applied to reduce the domain shift. Domain adaptation attempts to ensure the models' performance does not drop in the target domain when trained only with the synthetic source domain data. Successful domain adaptation eliminates the need for labelled data in the target domain allowing models to be trained using more cost effective and larger scale synthetic datasets.

This research focuses on the adversarial image-to-image translation approach to unsupervised domain adaptation of synthetic dashcam images, where ground truth labels are only available in the source domain. The goal is to learn a task model that performs well in an unseen target domain. A new contrastive learning based objective

function is proposed as a more efficient alternative to the cycle consistency loss commonly used. We refer to our full network architecture and approach as CLADA, which includes both pixel-, and feature-level domain adaptation to train a target semantic segmentation task model (See Figure 1). Our full approach requires less computational resources, leading to reduced training time.

## 2 Related Work

Several approaches have been taken to solve the domain adaptation challenge and deep learning methods have shown great progress by discovering domain invariant feature representations or by mapping images between the source and target domains [Shrivastava et al., 2017, Bousmalis et al., 2017, Li et al., 2019, Hoffman et al., 2018]. Earlier domain adaptation approaches focused on alignment within the feature space using some distance metric between the first- or second-order statistics of the source and target domains. By aligning the feature space representations of both domains, such that the feature embeddings follow the same distribution, a domain invariant model that generalizes better to the target domain can be learned [Sun and Saenko, 2016, Tzeng et al., 2014, Long et al., 2015]. Domain adversarial objectives have also been applied to feature space alignment, where a domain classifier is trained to distinguish between the source and target representations [Ganin et al., 2016, Tzeng et al., 2017, Tzeng et al., 2015, Ganin and Lempitsky, 2015].

More recently, further improvements have been made by approaching domain adaptation as a pixel-level image-to-image translation problem, leveraging the progress made by generative adversarial networks (GAN) [Goodfellow et al., 2014] in the image synthesis and style transfer domains. Earlier GAN based approaches to image-to-image translation required paired image samples [Isola et al., 2017], which would not be practical for the autonomous driving perception task.

Shrivastava et al. [Shrivastava et al., 2017] proposed SimGAN to translate unpaired images from synthetic source images to a target domain with the introduction of an additional self-regularizing function. This approach is successful in domains where there is a limited domain shift in pixel space. The addition of this L1 reconstruction loss for the generator during training preserves the annotations of the source data by penalizing large changes to the global structure during translation. Preserving this structural content is essential in pixel-level domain mapping, otherwise the source annotations would not accurately represent the new translated data used for supervised learning of the semantic segmentation model.

Zhu et al. [Zhu et al., 2017] introduced CycleGAN, which proposed a learned mapping applied to an input $x$ in both directions should be cycle consistent. That is, mapping a sample $x$ from the source domain $X$ to a target domain $Y$ using a learned mapping function $G_{s \to t}$, and then mapping back to the source domain using a learned mapping function $G_{t \to s}$, the result should be consistent with the original input $x$. CycleGAN uses the L1 distance to measure the reconstruction error, which they call cycle consistency loss. The forward and backward consistency constraint is used to train the generator model along with the original discriminator GAN loss. Zhu et al. [Zhu et al., 2017] reported better results for several image translation experiments when compared to SimGAN. CycleGAN, however, requires additional generator and discriminator models to implement cycle consistency, which results in higher computational requirements and time during training.

Hoffman et al. [Hoffman et al., 2018] proposed CyCADA, a combined approach to domain adaptation for the semantic segmentation task by applying both feature space domain invariant feature learning and pixel space domain mapping. CyCADA separates domain adaptation into two sequential steps. First, performing image translation from the source to target domain with a CycleGAN, and then further decreasing the domain gap by adding a domain adversary to the features of the semantic segmentation model. The advantage of pixel space adaptation is that it is more human interpretable, which allows visualizing the progress of the model as it is trained. This approach allows for interpretability at the pixel level, while also regularizing the feature level.

Li et al. [Li et al., 2019] introduced a bidirectional learning framework that uses both a CycleGAN-based image translation network and a segmentation adaptation network, similar to CyCADA. However, an end-to-end bi-directional training process was used, requiring more resources to train the closed-loop end-to-end architecture. Park et al. [Park et al., 2020] recently proposed a new image translation approach, introducing an alternative to the cycle consistency loss that does not require the additional generator and discriminator models for the two-way

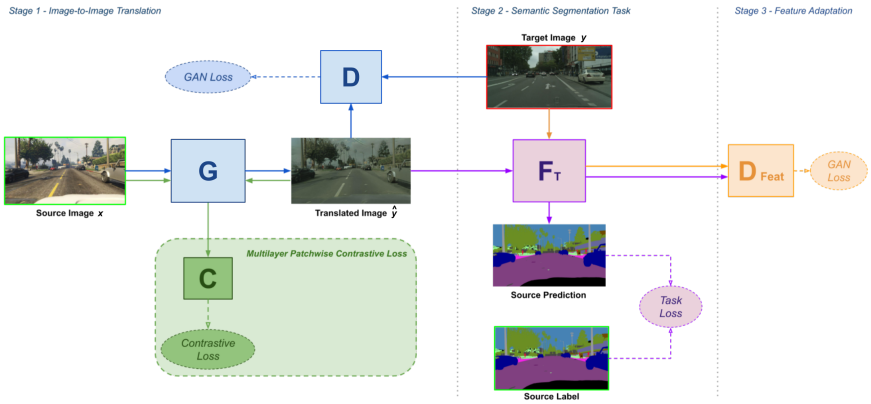



Figure 1: **Proposed CLADA architecture overview:** pixel-level GAN losses are in (blue), the PatchNCE loss in (green), the semantic segmentation task loss in (purple), and the additional feature-level GAN loss in (orange).

cycle consistency calculation. A multilayer patchwise contrastive loss (PatchNCE) is used to learn a one-way unpaired image translation that maintains the content of the input image while allowing the appearance to be adapted. The authors suggest that this alternative method is faster and requires less computational resources than a CycleGAN, which relies on additional auxiliary networks. Park et al. [Park et al., 2020] claim that their full method, including an additional identity loss, is 40% faster and 31% more memory efficient than CycleGAN at training time. The results shown in [Park et al., 2020] strongly suggest that PatchNCE could provide a more efficient alternative to cycle consistency in domain adaptation.

Motivated by [Park et al., 2020] and [Hoffman et al., 2018] this research evaluates the impact of replacing cycle consistency with a PatchNCE loss for unsupervised domain adaptation of synthetic autonomous driving dashcam images, which are then used to perform supervised learning for semantic segmentation. It shows that competitive results can be achieved for the semantic segmentation task using a simplified model architecture and less resources, producing faster training times. To our knowledge, no study has evaluated this approach on unsupervised domain adaptation for semantic segmentation.

## 3 Approach

Provided with source data $X$ and ground truth labels, and target data $Y$, with no labels, the aim is to learn a task model $F_t$ that when trained on the source data can correctly predict the semantic labels for the target data. Figure 1 shows the proposed architecture.

Similar to the staged-based approach taken by [Hoffman et al., 2018], we begin by fitting an image translation model $G$ (identical to $G_{s\to t}$ in [Hoffman et al., 2018]) that will apply pixel level domain adaptation to reduce the domain gap between the source and target data. This model $G$ is trained using a generative adversarial approach where it learns to map source images to the target domain, thus fooling an adversarial discriminator $D$ based on the GAN loss:

$$\mathscr{L}_{\text{GAN}}(G,D,X,Y) = \mathbb{E}_{y\sim Y} \log D(y) + \mathbb{E}_{x\sim X} \log(1 - D(G(x))). \tag{1}$$

To preserve the content structure of the source samples $x_s$, many previous approaches have used the cycle consistency loss. Here we propose to replace this with the contrastive learning based PatchNCE loss proposed by [Park et al., 2020]. This is based on a type of contrastive loss function, the InfoNCE loss [Oord et al., 2018], which aims to learn an encoder that associates corresponding patches with each other. The aim is to match corresponding patches between the input and output images. For example, a patch in the input image showing a traffic light should be associated with the corresponding patch showing a traffic light in the translated image. [Park et al., 2020] propose selecting multiple positive and negative pairs of patches from several layers within the feature stack of

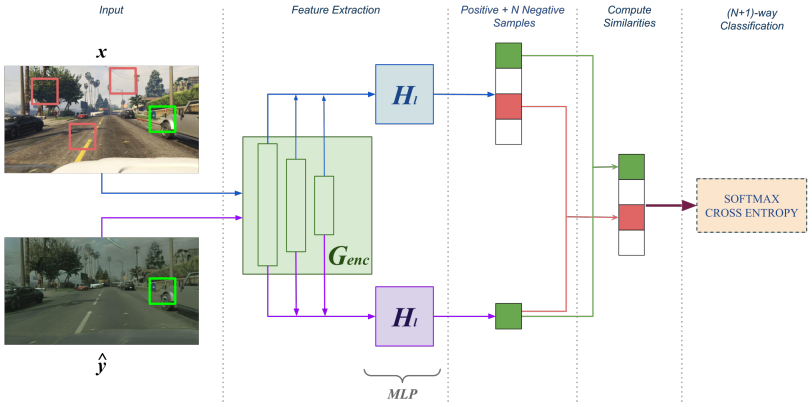



Figure 2: Multilayer patchwise contrastive loss.

the encoder $G_{\text{enc}}$ based on the normalized temperature scaled cross-entropy (NT-Xent) loss:

$$\ell(v, v^+, v^-) = -\log\left[\frac{\exp(v \cdot v^+/\tau)}{\exp(v \cdot v^+/\tau) + \sum_{n=1}^{N} \exp(v \cdot v_n^-/\tau)}\right], \tag{2}$$

where $v$, $v^+$ and $v^-$ are patches taken from layers and spatial locations within the feature stack of the image translation generator and $\tau$ is the temperature. By feeding the feature maps into a small multi-layer perceptron $H_l$ and selecting 1 positive and $N$ negative samples from a number of spatial locations, an $(N+1)$ way classification problem is setup. The contrastive loss (PatchNCE),

$$\mathscr{L}_{\text{PatchNCE}}(G_{\text{enc}}, H, X) = \mathbb{E}_{x \sim X}\left[\sum_{l=1}^{L}\sum_{s=1}^{S_l} \ell(\hat{z}_l^s, z_l^s, z_l^{S/s})\right], \tag{3}$$

can then be calculated and fed back to the generator during the training cycle. $L$ represents the layers within the generator $G_{\text{enc}}$ passed to $H_l$, $S_l$ represents the spatial locations within each layer, and $\hat{z}_l^s$, $z_l^s$ and $z_l^{S/s}$ represent the query, positive and negative patches respectively. This bypasses the need for a predefined similarity function. Figure 2 shows an overview of this multilayer patchwise contrastive loss architecture.

Park et al. [Park et al., 2020] introduces two variants of their architecture, which they refer to as CUT and FastCUT. Their CUT model includes an additional identity loss to impose a content structure constraint and selects $N = 256$ patches from $L = 5$ layers of the encoder, where FastCUT omits the identity loss and selects only $N = 16$ patches from within each layer. These proposed $L$ and $N$ values were also chosen for our CUT and FastCUT model variants in our experiments. FastCUT also applies a weight, $\lambda$, to the PatchNCE loss to constrain the content structure in the absence of the identity loss. The resulting loss function is

$$\mathscr{L}_{\text{GAN}}(G, D, X, Y) + \lambda_X \mathscr{L}_{\text{PatchNCE}}(G_{\text{enc}}, H, X) + \lambda_Y \mathscr{L}_{\text{PatchNCE}}(G_{\text{enc}}, H, Y). \tag{4}$$

For the CUT model, $\lambda_X$ and $\lambda_Y$ are set to 1.0 to jointly train with the identity loss. [Park et al., 2020] proposes using $\lambda_Y = 0.0$ for FastCUT to omit the identity loss and $\lambda_X = 10.0$ to compensate for its absence. We found $\lambda_X = 10.0$ too high in our experiments: qualitative results showed the model failed to translate the Cityscapes style, resulting in images more similar to the source GTA5 data. Reducing $\lambda_X$ to 5.0 improved the image translation results. We refer to this model variant as FastCUTL5 in the remainder of the paper.

Once an image-to-image translation model for producing translated images that are similar to images in the target domain has been fit, the learned model $G$ is used to generate a new translated dataset. This new translated data, along with the corresponding source labels, is used as training data for the next stage, where a fully supervised learning approach is taken to train a target semantic segmentation model $F_t$.

Lastly, additional domain adaptation is applied within the feature embedding space of the task model $F_t$ using a domain adversarial approach. By introducing a feature level GAN loss, we fit a discriminator $D_{\text{feat}}$ that can

(a) Source (b) CycleGAN (c) CUT (d) FastCUT L5

Figure 3: **Image translation:** GTA5 source input and translation models output examples. We observed all models successfully transferred the Cityscapes saturation levels and textures, such as the smoother road surface.

distinguish between the feature embeddings of inputs from the translated source and target datasets, when passed through the task model $F_t$, and feed that back to the task model during a round of fine tuning.

## 4 Experiments

| Model | FID | Time |
|---|---|---|
| Source | 68.6 | – |
| CycleGAN | 28.6 | 426ms |
| CUT | 29.4 | 321ms |
| FastCUT L5 | 33.2 | 170ms |

Table 1: Fréchet Inception Distance (FID) and training time for a single iteration.

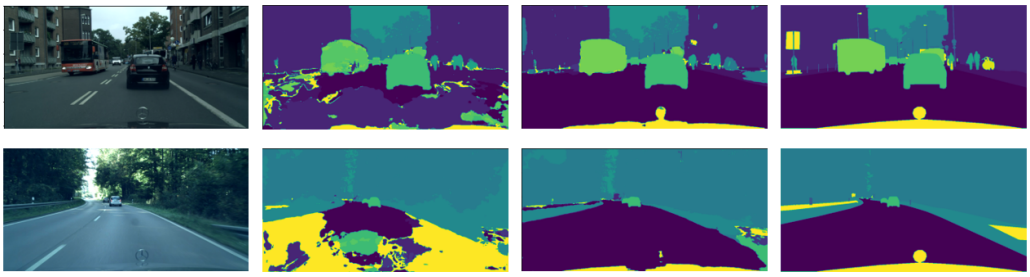
We evaluate our approach on the challenging unsupervised adaptation of the GTA5 [Richter et al., 2016] to Cityscapes [Cordts et al., 2016] datasets for the semantic segmentation task. Given that the ground truth labels are only available for the source GTA5 dataset, a task model is fit and its performance is evaluated in the target Cityscapes domain, where labelled data is available in a validation set. GTA5 is made up of 24,966 synthetic images extracted from the GTA5 computer game, with corresponding semantic labels at $1914 \times 1052$ resolution. Cityscapes provides real world dashcam images captured in Germany, and is split into train, validation, train extra and test sets at a resolution of $2048 \times 1024$. The train split has 2,975 images with dense ground truth labels, the validation split has 500 images with dense ground truth labels, and the train extra has 19,998 images without labels. Given that no labels are available for the test split, we use the validation split to evaluate our models performance.

Similar to CyCADA, a staged based approach is taken in which the image translation models are trained first. This allows us to interpret the progress of the pixel-level domain adaptation stage, and qualitatively evaluate the impact of replacing cycle consistency loss (CCL) with PatchNCE before proceeding to subsequent stages that include training the task model $F_t$ and further adaptation in feature space.

We trained a CycleGAN using the network architecture and training procedure from CyCADA. The images were resized to a width of 1024, maintaining the aspect ratio, from which random $400 \times 400$ crops were taken as input. The model was trained with a batch size of 1 for 400k iterations with a learning rate of $2 \times 10^{-4}$. After 200k iterations, the learning rate was linearly decayed to 0. The same procedure was used to train additional CUT and FastCUT models, where CCL was replaced with PatchNCE loss. As discussed $\lambda_X = 10$ for the PatchNCE loss was found to be too high for the GTA5-Cityscapes task; reducing it to $\lambda_X = 5$ achieved better results.

Following initial domain adaptation at the pixel-level, all models produced good quality images when translated to the Cityscapes appearance (Fig 3 illustrates an example). In particular, we noted that our CUT based models learned to adapt similar characteristics of the target Cityscapes domain to those adapted by CycleGAN, such as the image saturation levels, contrast and texture. All models, for example, learned that the road surface is much smoother in the target Cityscapes domain. We also noted that both the CycleGAN and full CUT model attempt to transfer the hood ornament, but FastCUT was not as prone to this. This is likely due to the lack of the additional identity loss, which may cause such features to be transferred.

Once all image translation models were trained and producing good qualitative image translation results, the



(a) Test Image (b) Source Only (c) CLADA (d) Ground Truth

Figure 4: **Semantic segmentation:** a test image (a) along with the corresponding source only model (b) predictions; our CLADA model (c) predictions and the ground truth masks (d).

| | road | sidewalk | building | wall | fence | pole | traffic light | traffic sign | vegetation | terrain | sky | person | rider | car | truck | bus | train | motorbike | bicycle | **mIoU** | **fwIoU** | **Pixel acc.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Source Only | 31.3 | 14.0 | 54.2 | 10.9 | 9.8 | 21.8 | 21.4 | 4.9 | 76.5 | 19.5 | 66.2 | 41.9 | 2.2 | 53.5 | 13.8 | 5.6 | 0.0 | 2.8 | 0.0 | 23.7 | 44.7 | 56.4 |
| CycleGAN | 78.9 | 30.4 | 75.6 | 20.5 | 1.3 | **29.5** | 24.3 | 0.0 | 80.3 | 32.2 | 69.2 | 48.0 | 0.0 | 78.9 | 24.6 | 31.0 | 0.0 | 0.0 | 0.0 | 32.9 | 71.0 | 79.2 |
| CUT | 77.4 | 27.2 | 75.1 | 18.4 | 17.9 | 27.9 | 0.0 | 0.0 | **80.4** | 29.8 | 72.9 | 47.6 | 0.0 | 79.6 | 28.3 | 28.6 | 0.0 | 0.0 | 0.0 | 32.2 | 70.4 | 78.4 |
| FastCUT L5 | 78.6 | **32.1** | **76.8** | 24.1 | **19.6** | 24.9 | 11.4 | **13.1** | 79.4 | 31.2 | **73.2** | 47.0 | 5.7 | 80.0 | 22.6 | 27.3 | 0.0 | 0.3 | 0.0 | 33.4 | 71.4 | 79.0 |
| CLADA (FastCUT L5 + FeatAda) | **79.9** | 31.0 | **76.8** | **24.5** | 18.7 | 28.0 | **24.4** | 12.7 | 79.6 | **31.6** | 72.2 | **51.0** | **11.7** | **81.5** | **29.9** | **33.9** | **3.4** | **7.6** | 0.0 | **36.8** | **72.2** | **80.3** |
| Oracle | 92.1 | 68.0 | 84.6 | 41.2 | 41.9 | 44.2 | 32.7 | 51.4 | 87.9 | 48.2 | 87.5 | 67.6 | 41.3 | 89.7 | 50.8 | 59.3 | 42.5 | 1.2 | 61.8 | 57.6 | 84.8 | 89.5 |

Table 2: GTA5-Cityscapes semantic segmentation task model evaluation results showing IoU for individual classes and mean IoU, frequency weighted IoU and pixel accuracy.

full GTA5 dataset was used to generate new translated datasets using each trained image translation model. A qualitative and quantitative evaluation was then performed. Each of these new translated datasets were compared to the Cityscapes data and the Fréchet Inception Distance (FID) [Heusel et al., 2017] calculated. Table 1 shows the results, and illustrates that, based on the statistical comparison between the translated datasets and the target Cityscapes data, the adapted images are more similar to Cityscapes than the original GTA5 images. Training time was measured and shown to be reduced when using the contrastive approach, in particular for the FastCUT variant that omits the identity loss. We found FastCUT is 47% faster than CUT and 60% faster than CycleGAN during training, where CycleGAN takes 426ms per iteration and FastCUT only takes 170ms (see Table 1).

The new translated datasets are then used in the next stage where we trained our task semantic segmentation models. The goal is to train a task model $F_t$ that performs well when evaluated on the Cityscapes validation split. Each task model is evaluated using three metrics: mean intersection-over-union (mIoU), frequency weighted intersection-over-union (fwIoU) and pixel accuracy (pixel acc.):

$$\text{mIoU} = \frac{1}{N} \cdot \frac{\sum_i n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}} \qquad \text{fwIoU} = \frac{1}{\sum_k t_k} \cdot \frac{\sum_i n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}} \qquad \text{pixel acc.} = \frac{\sum_i n_{ii}}{\sum_i t_i}, \tag{5}$$

where $N$ is the total number of classes, $n_{ij}$ is the number of pixels of class $i$ predicted as class $j$ and $t_i = \sum_j n_{ij}$ is the total number of pixels of class $i$.

For the semantic segmentation task, we use an EfficientNetB3 [Tan and Le, 2019] model, pretrained on ImageNet [Deng et al., 2009], as an encoder within a U-Net [Ronneberger et al., 2015] architecture. Each task model was trained with a batch size of 8 for 120k iterations with an initial learning of $2 \times 10^{-4}$, which was stepped down to $10^{-5}$ for the final 40k iterations. The same training procedure was used for all model variants, with the encoder frozen for the first 40k iterations. The *Source Only* and *Oracle* models were used to set lower and upper bounds on the achievable accuracies. The *Source Only* model was trained using the original GTA5 data and labels and the *Oracle* model was trained using the Cityscapes data and dense ground truth labels provided in the train split. We then trained our semantic segmentation task models using each of the translated datasets created using our CycleGAN, CUT, and FastCUTL5 image translation models. Finally, each of these trained task models was evaluated on the Cityscapes validation split using our task evaluation metrics (see Table 2).

| | road | sidewalk | building | wall | fence | pole | traffic light | traffic sign | vegetation | terrain | sky | person | rider | car | truck | bus | train | motorbike | bicycle | **mIoU** | **fwIoU** | **Pixel acc.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Source Only | 60.8 | 54.0 | 30.5 | 30.3 | 32.1 | 22.4 | 11.3 | 46.5 | 11.4 | 28.7 | 21.3 | 25.7 | 39.1 | 36.2 | 37.0 | 53.7 | 42.5 | -1.6 | 61.8 | 33.9 | 40.2 | 33.1 |
| Ours (CLADA) | 12.2 | 37.0 | 7.8 | 16.7 | 23.2 | 16.2 | 8.3 | 38.7 | 8.3 | 16.6 | 15.3 | 16.6 | 29.6 | 8.2 | 20.9 | 25.4 | 39.1 | -6.4 | 61.8 | 20.8 | 12.6 | 9.2 |
| **% Performance Gain** | 79.9 | 31.5 | 74.4 | 44.8 | 27.8 | 27.8 | 26.7 | 16.7 | 27.4 | 42.2 | 28.1 | 35.4 | 24.3 | 77.4 | 43.6 | 52.7 | 8.0 | - | 0.0 | 38.7 | 68.6 | 72.2 |

Table 3: Performance gap for the *source only* model and our full CLADA model, when compared to the oracle performance. The % performance gain is also shown for our full CLADA model versus the source only model. *Note: for the motorcycle class, our model outperforms the oracle*

The final end task results achieved using PathNCE during the image translation stage are comparable to those achieved with a CCL based CycleGAN. All models perform well on common classes. The performance of all models is poor for the *train* and *bicycle* classes due to these classes being under represented in the dataset. Semantic mask predictions (Fig 4) show qualitative results that correlate with the quantitative results. At this stage, FastCUTL5 has slightly better results, and has closed the performance gap for mIoU by approx 29%, which is competitive to CycleGAN and suggests that the PatchNCE contrastive objective is a viable replacement for CCL if faster training times is required, which may also lead to reduced training costs.

To further close the performance gap with the upper bound, as per [Hoffman et al., 2018], we performed further domain adaptation in feature space using a domain adversarial approach where we fine-tuned the FastCUTL5 model using a domain discriminator to classify the feature embeddings of the intermediate layer of the task model when source and target data are used as input. The final model, which we call CLADA, was evaluated using the same procedure and metrics and the results show the mIoU gap is closed by a further 10% (see Table 2). Overall, CLADA recovers approx, 39% mIoU lost to domain shift for the target segmentation task. In some cases, for well represented classes such as road, building, and car, it recovered >70% of the IoU performance lost (Table 3).

## 5 Conclusion

Our experiments show that by using a contrastive learning based objective function, PatchNCE, similar results to cycle consistency loss can be achieved for the challenging GTA5-Cityscapes semantic segmentation task, with faster training times and using a simplified model architecture. For the full approach, CLADA, the image translation stage is 60% faster during training and recovers approximately 39% of the mIoU performance lost to domain shift for the target semantic segmentation. By reducing the resources, costs, and time required to train generative adversarial domain adaptation models, our findings support further research into more complex end-to-end image translation approaches to domain adaptation of synthetic data for semantic segmentation.

## References


[Bousmalis et al., 2017] Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., and Krishnan, D. (2017). Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Computer Vision and Pattern Recognition (CVPR)*.

[Cordts et al., 2016] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Deng et al., 2009] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition (CVPR)*.

[Ganin and Lempitsky, 2015] Ganin, Y. and Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning (ICML)*, volume 37.

[Ganin et al., 2016] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016). Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research*, 17(1):2096–2030.

[Goodfellow et al., 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. In *Advances in Neural Information Processing Systems (NIPS)*, volume 27.

[Heusel et al., 2017] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *International Conference on Neural Information Processing Systems (NIPS)*.

[Hoffman et al., 2018] Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A., and Darrell, T. (2018). CyCADA: Cycle-Consistent Adversarial Domain Adaptation. In *International Conference on Machine Learning (ICML)*.

[Isola et al., 2017] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-Image Translation with Conditional Adversarial Networks. *Computer Vision and Pattern Recognition (CVPR)*.

[Li et al., 2019] Li, Y., Yuan, L., and Vasconcelos, N. (2019). Bidirectional learning for domain adaptation of semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR)*.

[Long et al., 2015] Long, M., Cao, Y., Wang, J., and Jordan, M. (2015). Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning (ICML)*.

[Oord et al., 2018] Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

[Park et al., 2020] Park, T., Efros, A. A., Zhang, R., and Zhu, J.-Y. (2020). Contrastive Learning for Conditional Image Synthesis. In *European Conference on Computer Vision (ECCV)*.

[Richter et al., 2016] Richter, S. R., Vineet, V., Roth, S., and Koltun, V. (2016). Playing for data: Ground truth from computer games. In *European Conference on Computer Vision (ECCV)*, pages 102–118.

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In *LNCS*, volume 9351, pages 234–241.

[Shrivastava et al., 2017] Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., and Webb, R. (2017). Learning from simulated and unsupervised images through adversarial training. In *Computer Vision and Pattern Recognition (CVPR)*.

[Sun and Saenko, 2016] Sun, B. and Saenko, K. (2016). Deep CORAL: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision (ECCV)*.

[Tan and Le, 2019] Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning (ICML)*, pages 6105–6114.

[Tzeng et al., 2015] Tzeng, E., Hoffman, J., Darrell, T., and Saenko, K. (2015). Simultaneous deep transfer across domains and tasks. In *International Conference on Computer Vision (ICCV)*.

[Tzeng et al., 2017] Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. (2017). Adversarial discriminative domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7167–7176.

[Tzeng et al., 2014] Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., and Darrell, T. (2014). Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*.

[Zhu et al., 2017] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *International Conference on Computer Vision (ICCV)*.