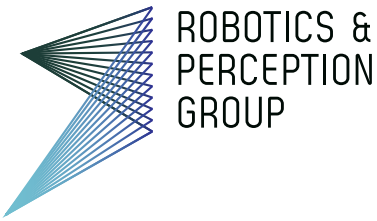
University of Zurich[UZH]
Department of Informatics

University of Zurich[UZH]
ETH zürich
Institute of Neuroinformatics



Rohit Kaushik & Qasim Warraich

Exploiting Semantics and Cycle Association for Domain-adaptive Semantic Segmentation

Semester Thesis

Robotics and Perception Group
University of Zurich

Supervision

Dr. Suman Saha
Menelaos Kanakis
Prof. Luc Van Gool
Prof. Davide Scaramuzza
Aug 2021

Contents

Abstract **ii**

Nomenclature **iii**

1 Introduction **1**
1.1 Related Work 2

2 Approach **3**
2.1 Domain Shift 3
2.2 Pixel-Level Cycle Association 4
2.3 Spatial Aggregation 6
2.4 Network Architecture 6
2.5 Objective 7
2.6 Training Pipeline 7

3 Experiments **9**
3.1 Dataset 9
3.2 Implementation Details 10
3.3 Results and Ablation Study 10

4 Discussion **12**
4.1 Implementational Differences 12
4.2 Analysis 12
4.2.1 Visualisations 12
4.2.2 Statistical Analysis 13
4.3 Limitations 14
4.4 Conclusion 15
4.5 Future Work 15

Acknowledgement **20**

Abstract

Performing domain adaptive semantic segmentation is one of the core vision problems as it facilitates automated visual scene understanding without the need for expensive and labor-intensive annotations. However, due to the inevitable domain shift problem, the model performance significantly degrades when tested on unseen target samples. Research into how best to mitigate the domain shift problem is a highly active research area. The majority of the existing approaches exploit adversarial learning to minimize the distribution discrepancy. Recently contrastive learning has been proposed as an alternative to tackle domain adaptation. In this work, we further investigate this new research direction. More specifically, we study how a contrastive objective, based on pixel-level cycle association, could help learn better a generalisable representation. The idea here is to bring the feature embeddings closer for the visually similar pixels and push the dissimilar pixels further apart. Pixel similarity is measured using a cycle-association technique that outputs pairs of pixels (from source and target images) having the least distance in the embedding space. Our experimental results on the challenging SYNTHIA to Cityscapes benchmark demonstrate that pixel-level contrastive learning is a promising approach that improves semantic segmentation results.

Nomenclature

Acronyms and Abbreviations

PLCA	Pixel Level Cycle Association & [17]
SS	Semantic Segmentation
DA	Domain Adaptation
GAN	Generative Adversarial Network
mIoU	Mean Intersection Over Union

Chapter 1

Introduction

Semantic image segmentation is a dense prediction task that aims to provide pixel level semantic annotations for an image. Semantic image segmentation is a key component in a variety of vision based applications such as autonomous driving, robotics, medical image analysis, etc. In recent years, due to the rapid developments and advancements in the area of deep learning, the task of semantic image segmentation has become one of the most actively researched classification tasks. Despite these advancements in deep learning neural networks, semantic segmentation remains an extremely difficult problem to study primarily due to it's dependence on massive and very densely annotated datasets. Such datasets are extremely laborious and costly to develop and as a result, currently there are only a handful of viable datasets to train semantic segmentation networks with. This barrier has inspired a lot of work towards the goal of utilising unannotated or virtual datasets to aid in the development of semantic segmentation networks.

The goal of using unlabeled and virtual datasets highlights another large research area, Domain Adaptation. Domain adaptation is the process of overcoming *domain shifts* or more explicitly the changes encountered between testing and training environments. Domain shifts can take on a multitude of forms ranging from shifts such as illumination changes to more drastic shifts such as from synthetic to real environments. The promise of using synthetic environments for training is especially interesting as through the exploitation of virtual environment engines a lot of the dense annotation that is normally required for semantic segmentation datasets can be easily obtained.

The majority of work around mitigating the domain shift for semantic segmentation algorithms centers around the use of Generative Adversarial Networks (**GANs**). While a promising approach, GANs come with their own host of sub problems such as tedious hyper parameter tuning and optimisation difficulty. In this work we take a different approach that fuses the ideas of cycle consistency and contrastive learning in order to bridge the domain gap.

Building off the work of [17], our study attempts to implement the pixel wise cycle consistency approach and explore its domain shift mitigation potential for the task of semantic segmentation. The idea behind Pixel Level Cycle As-

sociation (**PLCA**) is to bridge the domain gap by bringing together pairs of pixels in source and target images by enforcing a cyclic consistency. The cycle consistency in this case is satisfied by selecting pixels from the source domain, computing the closest pixel in the target with the use of a distance function. This process is then repeated on the computed target pixel to find it's closest match in the source domain. When both source domain pixels are from the same semantic class, this triplet can be considered cyclically consistent.

Our work is clean room style implementation of the original work with the goal of assessing the viability of the idea and validating the claims made by [17]. Through a series of experiments, a visualisation pipeline and statistical analysis we demonstrate that PLCA is a promising approach that improves semantic segmentation results. We additionally highlight some potential concerns and areas for improvement for future work.

1.1 Related Work

Domain adaptation is not a new problem [22, 16, 1, 11]. However, in recent years due to the advancements in deep learning domain adaptation has enjoyed significant new contributions and novel approaches. The majority of contemporary work in this space is based on generative adversarial [13] methods or MMD [32] methods. The adversarial approaches can be broken into two main approaches. Those that perform image to image translation [20, 3, 18, 8, 9], also often termed as style transfer and methods that work on the feature or representation level of images [27, 28, 12].

Adversarial approaches: This category represents the majority of work in domain adaptive segmentation. As previously mentioned adversarial approaches are employed mainly applied in two ways.

1. Implementations that augment image information, also called style transfer [8, 9, 18, 20, 3]. In general these approaches aim to generate altered versions of synthetic images that are closer to the real target domain.

2. Implementations that operate on the feature level or on network predictions [27, 28, 29, 12, 31, 21]. These approaches aim to make feature representation or predictions indistinguishable between two domains. For example [27] uses the output segmentation maps to perform the adaptation on.

Some approaches also combine the two methods [7, 34].

Self training methods: This category represents self training methods [19, 33, 25, 4]. These methods typically iteratively improve segmentation performance. These approaches often employ a curriculum based approach, for example [4] uses a pyramid curriculum to guide the training of the network.

Approaches that fuse additional tasks: Recently, approaches that fuse additional information such as depth or geometric information have also been proposed. [18, 7, 29]

Chapter 2

Approach

In this section we describe our approach for domain adaptive semantic segmentation. We have a labelled source data **S** and an unlabeled target data **T**. The objects present in both target and source are same. The labeled data comes from computer generated environment whereas the unlabeled target data is real world data captured from camera feed of cars in different European cities. We aim to train a network ϕ_θ to classify each target pixel into one of the underlying **M** classes.

As described in the introduction, there is a domain gap in our training and testing dataset. This gap is because of the fact that our training set S comes from a virtual environment whereas we test our network on images from real environment. Since our network should perform with a high accuracy on target **T**, it should produce similar deep features for pixels with same semantic class in both **S** and **T**. Because of the domain shift between **S** and **T**, the features are not the same which affects the semantic segmentation. To deal with this problem, we perform domain adaptive training and follow the approach introduced in Pixel Level Cycle association paper.

2.1 Domain Shift

Domain shift or domain gap is the change in data distribution between the training set and the target dataset. Specific to semantic segmentation, the domain shift is difference in feature or texture of the pixels in source and target images. As we have mentioned, we are interested in training semantic segmentation models that use labeled synthetic data in the source domain and real world data in the target domain. More specifically the domain gap in this setup is between the computer generated synthetic data and real world data.

The domain shifts however, can also arise even if the both source and target images are real world images. This could be because of different weather conditions, different perspectives of the image, illumination and other factors. For a semantic segmentation model to work robustly, the domain gap should be mitigated as much as possible.

Figure 2.1: Example of domain shift between the SYNTHIA and Cityscapes datasets

2.2 Pixel-Level Cycle Association

In the previous section we saw how domain shift between source and target datasets affects the accuracy of semantic segmentation. To reduce the domain gaps, GANs have been predominantly used by researchers in previous work. In this work we look at a novel approach for domain adaptation based on a pixel level cyclic association. We now describe the idea behind pixel level cycle association.

Several works such as [35], [30] have used cycle consistency in different applications and have found it to be effective. CycleGAN [35] popularised the use of cycle consistency for domain adaptation. CycleGan uses cycle consistency to perform style transfer in order to diminish the gap at an image-level. In the field of semantic segmentation, researchers have frequently used GANs to mitigate the image level domain gap. The idea behind this is to build a discriminator that can classify between real image and virtual images. However, this approach typically operates on a globel image level does not consider the rich pixel wise relationship which has the potential to improve the semantic segmentation.

Figure 2.2: Illustration of Pixel Cycle Association from [17]

In this paper, we propose to use cycle consistency to build the pixel-level correspondence. Specifically, given the source feature map $F^s \in \mathbb{R}^{C \times H^s \times W^s}$ and the target feature map $F^t \in \mathbb{R}^{C \times H^t \times W^t}$ for an arbitrary source pixel $i \in \{0, 1, ..., H^s \times W^s - 1\}$ in F^s , we compute its pixel-wise similarity with all of the target pixels in F^t . The pixel-wise similarity between the features is measured using cosine similarity given by the equation below:

$$D(F_i^s, F_j^t) =< \frac{F_i^s}{\left\|\left(F_i^s\right)\right\|} \cdot \frac{F_j^t}{\left\|\left(F_j^t\right)\right\|} > \tag{2.1}$$

where $F_i^s \in \mathbb{R}^C$ and $F_j^t \in \mathbb{R}^C$ are the features of source pixel i and target pixel j. The $\cdot$ represents the dot product operation.

For each of the source pixel i, we select the closest target pixel j^* using the equation:

$$j^* = \underset{j' \in 0,1,\ldots,H^t \times W^t - 1}{\operatorname{argmax}} D(F_{i^*}^s, F_{j'}^t) \tag{2.2}$$

In a similar way, we get the closest pixel i^* to the computer target pixel j^* from the source domain using the equation:

$$i^* = \underset{i' \in 0,1,\ldots,H^s \times W^s - 1}{\operatorname{argmax}} D(F_{j^*}^t, F_{i'}^s) \tag{2.3}$$

We additionally enforce a semantic condition for cycle consistency to hold. The conditions for cycle consistency is only satisfied when the source pixel i and final source pixel i* come from the same semantic class.

We mitigate the domain gap by bringing the feature representation of associated pixels (i, j^*) and $(j^*, i*)$ closer. The pixels are brought closer in representation by performing contrastive learning rather than directly maximizing their feature similarity since this might introduce bias. Learning contrastively, the similarity of the association pair is made higher than any other possible pair. During the training, we minimize the contrastive association loss which is given by the formula:

$$L^{fass} = -\frac{1}{|\hat{I}^s|} \sum_{i \in \hat{I}^s} \log \{ \frac{\exp \{D(F_i^s, F_{j^*}^t)\}}{\sum_{j'} \exp \{D(F_i^s, F_{j'}^t)\}} \frac{\exp \{D(F_{j*}^t, F_{i^*}^s)\}}{\sum_{i'} \exp \{D(F_{j^*}^t, F_{i'}^s)\}} \} \quad (2.4)$$

2.3 Spatial Aggregation

In the previous section we described how the cycle association between the source and target pixels is established. In practice only a small portion of target pixels would form cycle associations given that the source and target come from two different domains. Also since these source and target images are sampled randomly, the semantic class overlap between the images could be different and hence it also restricts the number of associations.

To regularise the training and the back propagation of error through all pixels, we spatially aggregate the features for each target pixel before performing the cycle association. In the spatial aggregation step we represent each target pixel as a weighted sum of features of all the other target pixels.

$$\hat{F}_j^t = (1 - \alpha) \times F_j^t + \alpha \sum_{j\prime} w_{j\prime} F_{j\prime}^t \quad (2.5)$$

where the weight is given by:

$$w_{j\prime} = \frac{\exp D(F_j^t, F_{j\prime}^t)}{\sum_{j\prime} \exp D(F_j^t, F_{j\prime}^t)} \quad (2.6)$$

In our method, we use $\alpha = 0.5$ as described in the paper [17]. The important thing to note here is that gradients are diffused unevenly and it is determined by the similarity of features.

2.4 Network Architecture

DeepLab[5] is the state of the art architecture for semantic segmentation. Most work on semantic segmentation uses the DeepLab v2 network as a backbone. We also use the DeepLab v2 network as a backbone albeit with a modification. Instead of using the ResNet-101[15] backbone we use ResNet-18. The reason for using ResNet-18 is that it is a smaller architecture and hence more GPU memory efficient. Respecting our GPU limitations, we wanted to train our network with

a batch size greater than one. The reason for this is that batch size is an important parameter in contrastive learning. Given the aforementioned GPU limitations ResNet-18 was the only viable option that allowed for the usage of a batch size greater than one.

2.5 Objective

As with any other classification task, for the source data, our loss function is pixel-wise cross entropy, given by:

$$L^{ce} = \frac{-1}{|I^s|} \sum_{i \in I^s} \log P_i^s(y_i^s) \tag{2.7}$$

This loss is only calculated on the source images, since we do not use the labels from target images and hence $|I^s|$ denotes the source image.

Given that the number of pixels for each semantic class would vary a lot and some semantic class like building, road, sky would have more number of pixels than other classes, we apply Lovasz-Softmax [2] loss to regularize the effect of imbalanced data.

The Lovasz-Softmax loss is given as:

$$L^{lov} = \frac{1}{|C|} \sum_{c \in C} \Delta_{J_c}(m(c)) \tag{2.8}$$

where C is the number of semantic classes and J is the Jaccard index. The final loss is the association loss and since we only use associations at the final feature layer so term it as L^{fass}.

Overall our final objective for the adaptation process is:

$$L^{full} = L^{ce} + \beta_1 L^{lov} + \beta_2 L^{assoc} \tag{2.9}$$

2.6 Training Pipeline

In this section we describe our training pipeline and the steps required to perform the contrastive learning.

We use the pretrained ResNet-18 [26] backbone and train it on SYNTHIA[24] source data and Cityscapes[10] target data. We only use the labels from source dataset SYNTHIA and for the pair of source and target data we also perform association and calculate association loss. But before performing the association, several preliminary steps needs to be followed:

1) Apply spatial aggregation on the target image, as described in the previous section this is done to represent each pixel as a weighted sum of every other pixel so that the gradient from association loss can cover every pixel.

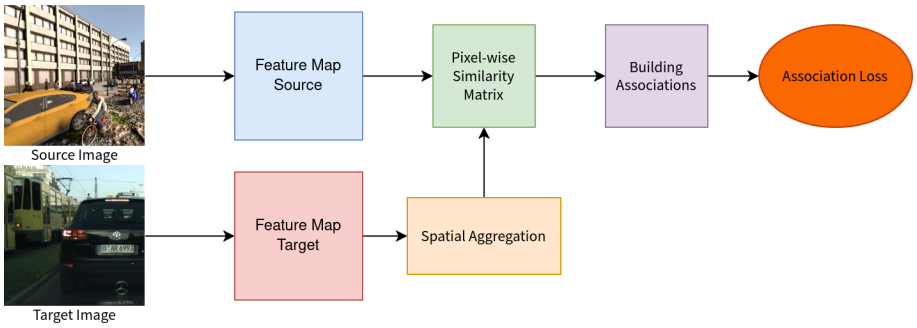


Figure 2.3: Training Pipeline for Contrastive Learning

2) Calculate the distance matrices using cosine similarity metric, to do this efficiently the feature map F_s & F_t are first flattened and a matrix multiplication operation is performed to calculate pixel-wise distance.

3) Contrast normalization on the the Distance matrix, that is for every pixel the distance vector is standardized with mean and variance.

4) Finally performing the association by computing the valid associations as described in the section of pixel-wise cycle association, Note the association are performed at the last feature map and the dimension of the last feature map is 92×92.

Chapter 3

Experiments

To study the proposed approach, we perform three different experiments. The result of these experiments are discussed in detail in the ablation section. Our aim was to study the performance of contrastive loss and lovasz loss as they are applied sequentially. We use mean intersection over union (**mIoU**) as the accuracy metric which a common metric for the task of semantic segmentation.

Below we briefly describe the three different experiments:

- **Baseline**: The first step was setting our baseline for comparison. We follow the usual approach and use the DeepLab v2 as the backbone. Since we wanted to train the models with a larger batch size (4), we modified the DeepLab architecture to use ResNet-18 as opposed to ResNet-101. ResNet-18 is a much smaller network and it allows us to run our experiments with higher batch size.
- **Baseline with Lovasz loss**: As an additional loss to the cross entropy loss used in the baseline, we also perform an experiment with Lovasz loss. The Lovasz loss is used to regularize the imbalance in dataset. Certain semantic classes like, road, sky etc. will take up most of the pixels in image and hence the target predictions can be biased. Lovasz loss helps deal with this imbalance.
- **Domain adaptive model**: In our final experiment, we introduce the contrastive pixel wise association loss. Our aim is to study improvement in semantic segmentation accuracy relative to baseline and Lovasz experiment.

3.1 Dataset

We perform the domain adaptive semantic segmentation training on the source (SYNTHIA) virtual domain and target (Cityscapes) real domain. These datasets have been widely used by various other previous work and hence for the sake of comparability we use these datasets. The Cityscapes consists of street scenes images captured from camera feed of the car around various cities in Europe. The dimension of the images are 2048×1024. The dataset in split

into train and test set and the test set contains 500 images from Cityscapes. The labels for Cityscapes images are not used in training and the images are used to train our adaptation model.

3.2 Implementation Details

Most of the work related to semantic segmentation, uses the DeepLab V2 architecture with ResNet-101 as the backbone, but given the limitation of GPU memory and the batch size being an important parameter in the contrastive learning, we modify our pipeline to implement DeepLab V2 with ResNet-18 as the backbone. This allows us to use a higher batch size and perform the training much more effectively. The downside is the mIoU is lower and is not directly comparable to other works since they typically use ResNet-101. For this reason, we wish to study only the relative improvements from the introduced contrastive loss.

In the training, we train our model based on stochastic gradient descent (SGD) with momentum of 0.9 and weight decay of 5×10^{-4}. We employ the poly learning rate schedule with the initial learning rate at 2.5×10^{-4} . We train for 30K iterations with batch size of four. For all of the tasks, we resize images from both domains with the length of shorter edge randomly selected from [760, 988], while keeping the aspect ratio. The 730×730 images are randomly cropped for the training. The resolution for building the associations is 92×92. At the test stage, we first resize the image to 1460×730 as input and then upsample the output to 2048×1024 for evaluation. In all of our experiments, we set β_1, β_2 to 0.75, 0.1 respectively.

3.3 Results and Ablation Study

Our experiments show some promising and interesting results. We observe that introducing Lovasz loss, the mIoU improves from **18.7** to **19.32**. The Lovasz loss also improves the class wise accuracy of some under represented classes such as light and bicycle.

		SYNTHIA ⟶ Cityscapes															
Experiment	**mIoU**	Road	Sidewalk	Building	Wall	Fence	Pole	Light	Sign	Vegetation	Sky	Person	Rider	Car	Bus	Motorcycle	Bicycle
Baseline	**18.7**	16.48	14.6	38.89	0.35	0.0	21.12	0.19	3.18	55.46	59.34	41.32	1.64	43.65	2.2	0.0	0.76
Lovasz-loss	**19.32**	13.41	15.34	40.52	1.72	0.04	23.61	1.28	5.81	50.48	62.64	42.67	3.61	41.5	2.4	0.14	4.0
Lfass	**20.32**	7.47	13.28	50.74	3.77	0.15	17.18	1.4	11.9	55.14	64.04	35.89	7.4	43.16	6.49	1.9	5.27

Table 3.1: Our Results (PLCA with ResNet-18 backbone) with batch size 4

Our final experiment with the inclusion of the pixel-wise contrastive loss also leads to an improvement in mIoU of **1.0**. The thing to note here, is that the architecture used in our experiments is different and we use a ResNet-18 backbone which is smaller network and hence the mIoU from our experiments cannot be directly compared to that of [17]. For this reason, we study the relative increase in mIoU as the new losses are introduced in the three experiments. In the following table, we show the relative improvements in accuracy:

	L^{ce}	L^{lov}	L^{fass}
Ours	18.7	+3.31%	+5.17%
PLCA[17]	35.4	+2.82%	+12.3%

Table 3.2: Relative increases in MIoU

From the above table, we see our experiment with contrastive loss gives a relative improvement of **5.17%** as compared to **12.3%**. We wanted to study the reason for this and analyze some potential problems or areas that could be further developed. One reason could be that our setup for experiments is different than [17]. We use a smaller backbone and during training we use a batch size of four as opposed to eight. Batch size has been found to be an important parameter in contrastive learning. In the next section, we provide a more in depth study and discussion with our visualisation pipeline and statistical study to better understand the potential problems with this approach.

Chapter 4

Discussion

In the last section, we saw the results from the three experiments we performed. We also looked at the relative improvements when individual losses are systematically introduced. From the experiments, we observe a lower relative improvement with the contrastive pixel-wise loss as compared to the original paper [17]. While experiment setup and other hyper-parameters like batch size can affect the accuracy, we also believe there are other potential problems like the random pairing of the source and target images and the pixel cycle associations that we form during the loss calculation. We perform some statistical and visual study to understand these in further detail.

4.1 Implementational Differences

As, at the time of writing, the original code of [17] was not available for review our implementation cannot be compared with that of [17] in a one to one fashion. We implemented the approach from scratch in PyTorch[23] and our code is available publicly. Following the exact setting and experiments of PLCA is a computationally expensive task. Computing pixel wise similarity indices takes up a lot of space. We perform cycle association at the feature and the feature map is 92×92, which means that have distance for every pair of pixel from source and target would require a matrix of 8464×8464. Besides the space intensive correlation matrices, original paper uses a batch size of eight. Several works [6, 14] show how batch size is an important parameter in semantic segmentation and contrastive learning. Given the limitation of GPU memory, we were only able to run our experiments with a batch size of four and with a modified version of DeepLab v2 which uses ResNet-18 for the backbone.

4.2 Analysis

4.2.1 Visualisations

We developed a visualisation pipeline to interpret the cycle associations we obtain from the trained model. Obtaining good cycle associations is essential, since

we want to bring together the representations closer together. The visualisation pipeline allows us to:

- Visually check associations and look for anomalies.
- See the class based association performance.
- Validate our implementation.

Below you can see a visualisation for a single pixel from car and its corresponding association.

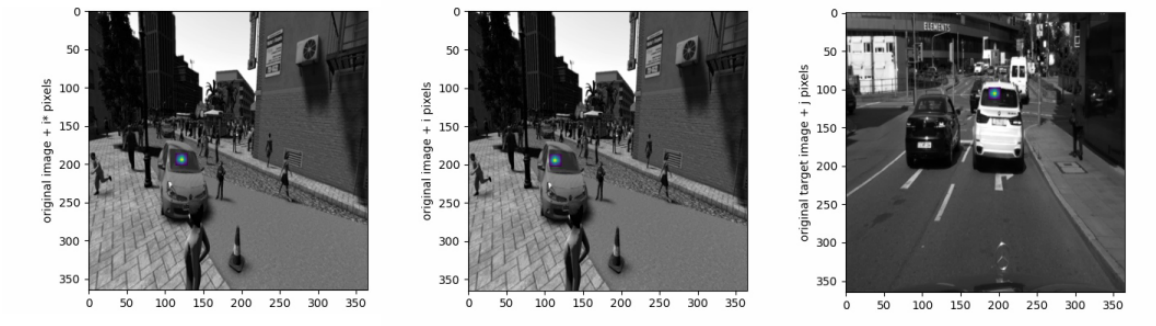


Figure 4.1: Visualisation output of a single cyclically consistent triplet of pixels.

For each pixel (i) in the source image its closest pixel, that is most visually similar pixel, in the target (j) is computed using a distance function. Cosine similarity is the distance metric we employ. The process is then applied to that target pixel to compute another closest source pixel (i*). When i and i* are from the same semantic class, the triplet can be considered cyclically consistent.

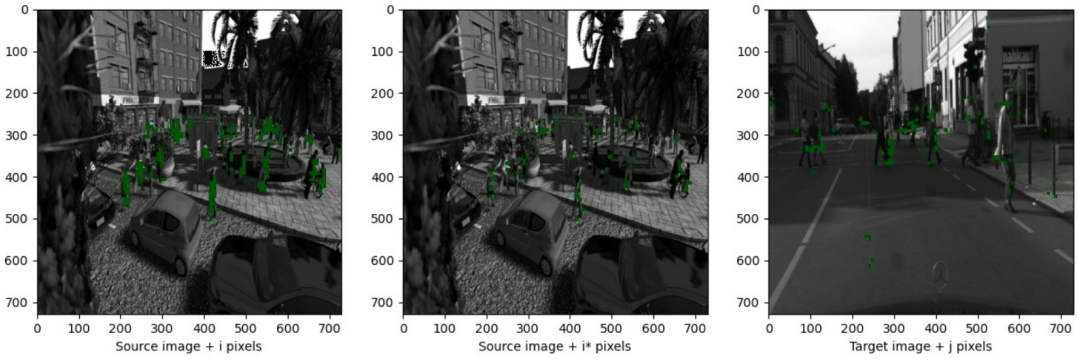


Figure 4.2: Visualisation output of a single class (person)

.

4.2.2 Statistical Analysis

In the last section, we introduced the visualisation pipeline. The visualisations help us with analyzing if the pixel level associations we observe are meaningful. In our examples, we see that while we get good associations, we also have quite

a lot of wrong associations which can drive our model training to perform worse. An example of such a bad association is, sidewalks being cycle associated with roads in the target. We also see that there our pairs of source and target image, where the class overlap between the pairs are minimal and hence getting good cycle association is not possible. We believe this to be a problem arising from random pairing of source and target images. To study this more analytically we calculate class wise out of class examples, that is, how many of the instances exist are such that the semantic class appears in target but not in source. This is done for 30k iterations or random pairings and the results are described in the table below.

	Out of class examples															
	Road	Sidewalk	Building	Wall	Fence	Pole	Light	Sign	Vegetation	Sky	Person	Rider	Car	Bus	Motorcycle	Bicycle
Instances	420	1662	406	8432	13442	268	8959	1583	821	2719	6371	19599	1359	20756	20686	13355

Table 4.1: Statistics per class

From the table above we can see, that there are lot of examples where the target contains a class that is not present in source image. This outlines a problem with the random pairing in the dataloader. We believe a better dataloader which uses some heuristics to pair images in a better way can further help with generating better cycle associations and hence could potentially improve the semantic segmentation performance. An interesting thing to note is the class building which has least number of out of class examples also show the greatest gain in mIoU when contrastive loss is applied as seen here in table 3.1

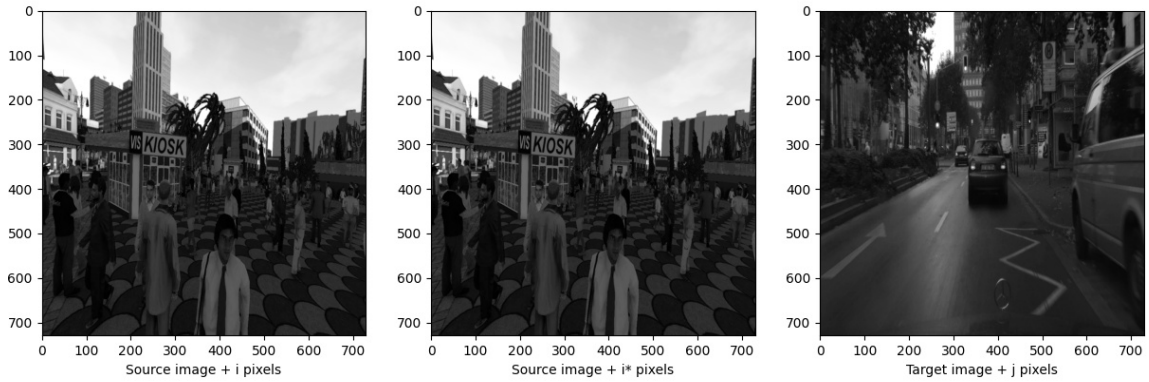


Figure 4.3: Example of an image pairing with low class overlap

4.3 Limitations

Our replication study was subject to a few limitations that prevented us from performing an one to one replication of the original work by [17]. It appears that implementing Pixel-wise Cycle Association is an expensive computation

and required us to make a few compromises. Due to a limitation on available GPUs on our cluster we had to settle for a smaller batch size of four as opposed to the batch size of eight as described in the original work. Furthermore, due to time constraints on the cluster we had to utilise a leaner back end, namely ResNet-18 as opposed to ResNet-101 as described in the [17]. We initially performed experiments on a ResNet-101 back end but were limited to a batch size of 1, which explains our eventual shift to the smaller ResNet-18.

In addition to hardware limitations, during the time of our project we had no access to the codebase of [17], despite having attempted to contact the original authors. This also prevented us from doing a 100% replication of the [17] work as we had to self implement the entire pipeline. This was performed using the information shared by [17] in their research paper.

4.4 Conclusion

In this work, we explored a novel idea of exploiting pixel-wise similarities to mitigate the domain gap. Our experiments suggest that contrastive learning with cycle association can help improve semantic segmentation performance by mitigating the gap in a novel way. Although interesting, there are some challenges in performing PLCA. We perform a statistical study and provide some visualisation to show some problems with the random pairing of source and target images and how it can affect the cycle associations we get between pixels. We hope this study and its contributions will help any further research in the direction of domain shift mitigation using pixel-wise cycle association.

4.5 Future Work

Semantic segmentation and domain adaptation are extremely active areas of research and the room for future contributions is very broad. With respect to the idea of pixel-wise cycle association there is also space for other researchers to pick up where [17] and this work left off. First of all, it would be interesting to see a team without hardware limitations attempt a one to one replication of [17] with the use of the same backbone and batch size. Furthermore it would be interesting to look into the idea of "cherry picking". That is, to select image pairs from the source and target domain that consist of a large class overlap to help establish more cyclically consistent pairs of pixels to better train the network.

Another interesting future approach would be to look into fusing additional tasks such as depth or geometric information about the scene to assist in bridging the domain gap. There are already works such as [7, 29, 18] exploring the potential of fusing additional information in addition to semantic class labels to help bridge the domain gap.

Bibliography

[1] Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137, 2007.

[2] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4413–4421, 2018.

[3] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3722–3731, 2017.

[4] Wei-Lun Chang, Hui-Po Wang, Wen-Hsiao Peng, and Wei-Chen Chiu. All about structure: Adapting structural information across domains for boosting semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1909, 2019.

[5] 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. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[7] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1841–1850, 2019.

[8] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. Crdoco: Pixel-level domain transfer with cross-domain consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1791–1800, 2019.

[9] Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6830–6840, 2019.

[10] 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. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[12] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] Hal Daumé III. Frustratingly easy domain adaptation. *CoRR*, abs/0907.1815, 2009.

[17] Guoliang Kang, Yunchao Wei, Yi Yang, Yueting Zhuang, and Alexander G Hauptmann. Pixel-level cycle association: A new perspective for domain adaptive semantic segmentation. *arXiv preprint arXiv:2011.00147*, 2020.

[18] Kuan-Hui Lee, German Ros, Jie Li, and Adrien Gaidon. SPIGAN: privileged adversarial learning from simulation. *CoRR*, abs/1810.03756, 2018.

[19] Qing Lian, Fengmao Lv, Lixin Duan, and Boqing Gong. Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6758–6767, 2019.

[20] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4500–4509, 2018.

[21] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through

self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[22] Sinno Jialin Pan, Ivor W. Tsang, James T. Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011.

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[24] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016.

[25] Tong Shen, Dong Gong, Wei Zhang, Chunhua Shen, and Tao Mei. Regularizing proxies with multi-adversarial training for unsupervised domain-adaptive semantic segmentation. *arXiv preprint arXiv:1907.12282*, 2019.

[26] PyTorch Team. Deep residual networks pre-trained on imagenet. `https://pytorch.org/hub/pytorch_vision_resnet/`.

[27] 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. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7472–7481, 2018.

[28] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

[29] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Mathieu Cord, and Patrick Pérez. Dada: Depth-aware domain adaptation in semantic segmentation. In *ICCV*, 2019.

[30] Xiaolong Wang, Allan Jabri, and Alexei A. Efros. Learning correspondence from the cycle-consistency of time, 2019.

[31] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wenmei Hwu, Thomas S. Huang, and Honghui Shi. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12632–12641, 2020.

[32] Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition, pages 2272–2281, 2017.

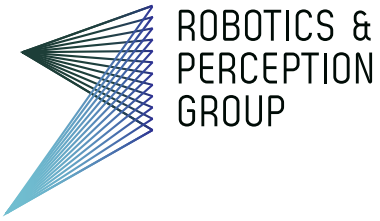
[33] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 2020–2030, 2017.

[34] Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, and Tao Mei. Fully convolutional adaptation networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6810–6818, 2018.

[35] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Acknowledgement

We would like to thank the Computer Vision Laboratory at the ETH for facilitating our project. A special thanks goes to our ETH Computer Vision Laboratory supervisors, Dr. Suman Saha and Menelaos Kanakis. We also extend our gratitude to Professor Davide Scaramuzza for his role as a supervisor for our project.



Title of work:

Exploiting Semantics and Cycle Association for Domain-adaptive Semantic Segmentation

Thesis type and date:

Semester Thesis, Aug 2021

Supervision:

Dr. Suman Saha
Menelaos Kanakis
Prof. Luc Van Gool
Prof. Davide Scaramuzza

Students:

Name:	Rohit Kaushik
E-mail:	rohit.kaushik@uzh.ch
Legi-Nr.:	97-906-739
Name:	Qasim Warraich
E-mail:	qasim.warraich@uzh.ch
Legi-Nr.:	18-787-796

Statement regarding plagiarism:

By signing this statement, we affirm that we have read the information notice on plagiarism, independently produced this paper, and adhered to the general practice of source citation in this subject-area.

Information notice on plagiarism:

`http://www.lehre.uzh.ch/plagiate/20110314_LK_Plagiarism.pdf`

Zurich, 3. 5. 2022: _____