Synthetic Dataset Acquisition for a Specific Target Domain

Joshua Niemeijer
German Aerospace Center (DLR)
Braunschweig, Germany
Joshua.Niemeijer@dlr.de

Sudhanshu Mittal
University of Freiburg
Freiburg, Germany
mittal@cs.uni-freiburg.de

Thomas Brox
University of Freiburg
Freiburg, Germany
brox@cs.uni-freiburg.de

Abstract

Intelligent sampling from simulation becomes crucial due to storage and hardware constraints. This research focuses on developing an intelligent acquisition strategy for synthetic data and evaluates multiple approaches to address the limitations of existing domain adaptation methods. Selecting suitable synthetic data for real-world model training presents challenges, as accurately representing the real world remains elusive. We tackle the task of adapting from synthetic to real-world data through unsupervised domain adaptation, a challenging setting for perception systems. The performance of our acquisition function is measured by its facilitation of this adaptation.

We showcase different strategies, to assign value to synthetic images. Acquisition functions either operate based on synthetic data alone or take the given real world target domain into account, to assign a value to synthetic images. Leveraging assumptions from semi-supervised learning, we identify challenging real-world images and find their counterparts in the synthetic world. Evaluation is conducted using the GTA-5 dataset as the representative synthetic world and the Cityscapes and ACDC dataset as the target domain. State-of-the-art unsupervised domain adaptation approaches are employed to assess the effectiveness of our acquisition function.

By advancing the utilization of synthetic data in training perception systems, this research contributes to improved real-world performance. Our findings demonstrate the potential of intelligent acquisition strategies for enhancing the adaptation from synthetic to real-world domains.

1. Introduction

Synthetic data represents a promising building block for training perception systems. Especially for tasks like semantic segmentation, which require about 90 minutes per frame to annotate [2], simulation is a crucial alternative for automatically creating labeled data. Hence, the creation of simulation engines has been the focus of the computer vision community. However, the number of possible scenarios and configurations such simulators can produce is still uncountably large. Due to the constraints in storage space and training hardware, sampling from the simulation in an intelligent way is crucial. We do not have enough memory to generate all necessary cases and even if we had, training on such a dataset becomes infeasible, since convergence is slow since, due to redundancy, important data points occur rarely during training. Some parts of synthetic data could even introduce a bias that negatively affects real-world performance.

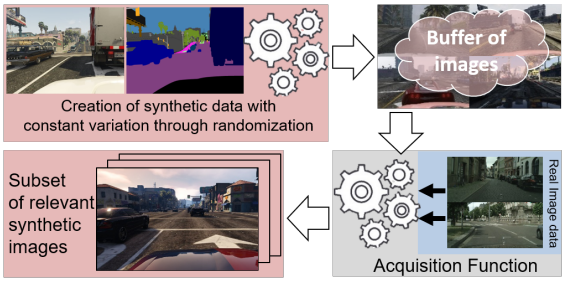

Figure 1. A system for creating a synthetic dataset. The simulation creates images with constant large variation. The images are written into a buffer. The acquisition function selects the synthetic images relevant for the real world target domain. As a result a smaller subset of relevant synthetic images is obtained.

In this work, we introduce the research direction of finding an intelligent acquisition strategy for synthetic data and showcase the performance of different approaches. Such acquisition functions face several challenges. Finding an objective to select synthetic data suitable to train models for the real world is not obvious. While realism and variation are two crucial objectives, identifying an objective that accurately represents the real world can be elusive. The two most commonly used Synthetic datasets, the GTA5 dataset [18] and the SYNTHIA dataset [19], e.g., selected their samples based on random or uniform sampling. While these strategies provide a wide range of scenes and variations, they may not consider important factors such as the distribu-

tion of challenging or informative data points. Additionally, random or uniform sampling might not effectively capture the specific characteristics or challenges of the real-world target domain. Therefore, there is a need for intelligent acquisition strategies.

In this work, we address the limitations of such existing acquisitions. The objective is to score the value of a simulated frame by estimating its usefulness in training perception models to perform well on the real-world target domain. More specifically, we focus on the challenging setting of adapting from the created synthetic dataset world to the real-world use case via unsupervised domain adaptation. By integrating unsupervised domain adaptation, we aim to leverage the power of adaptation techniques and minimize the need for simulation of aspects that can be effectively covered by domain adaptation. The performance of our acquisition function is hence defined by how well it facilitates the adaption. We seek to develop an intelligent acquisition strategy that selects synthetic data points based on their potential to enhance adaptation.

We showcase the performance of several strategies. On the one hand, we try to sample based on objectives that are independent of the real-world target domain. We utilize class uniform sampling and active learning acquisition functions like entropy and coreset to assign a value to the synthetic images. On the other hand, we take the assumption that underlies many semi-supervised learning methods: "Unsupervised learning performs best if correspondences between the labeled and unlabeled sets exist". We hence identify challenging images in the real world and then try to find their counterparts in the synthetic world. For repeatability, we choose the GTA-5 dataset to represent the synthetic world we can sample from. The Cityscapes and ACDC datasets represent our target domain. This setting simulates a scenario that is demonstrated in figure 1, where we assume that a simulation engine (e.g., CARLA [4]) produces simulation data constantly and writes it to a buffer. We aim to acquire the optimal subset of synthetic images from this buffer.

The paper is structured as follows: In section 2, we give a brief insight into the relevant state of the art. Then, in section 3, we describe our methods, which are applied to the problem in question. In section 4, we present our setup and experimental obtained results. Finally, in section 5, we summarize the paper and present our future work.

2. Related Work

2.1. Simulation engines and acquisition strategies

There are several existing synthetic simulation engines and acquisition strategies that have been used to create synthetic datasets. These engines play a crucial role in generating diverse and realistic synthetic data for various applications. Here, we discuss some notable examples:

GTA5 [18] dataset was created by extracting frames from the game Grand Theft Auto V (GTA5) and annotating them with pixel-accurate semantic labels. Frames were recorded during gameplay using the tool RenderDoc. Synthetic frames were sampled by taking every 40th frame to capture diverse scenes within the game world. The dataset's main purpose is to provide a large-scale, pixel-level annotated dataset for training semantic segmentation models. The dataset leverages the content of GTA5 to supplement real-world images and enhance the accuracy of semantic segmentation models.

SYNTHIA [19] dataset provides diverse synthetic frames with semantic segmentation annotations. The dataset was generated using a virtual city created with the Unity development platform. The synthetic frames were sampled from multiple camera viewpoints, incorporating dynamic objects, illumination conditions, and textures to improve visual variability. By combining SYNTHIA with real-world data, the dataset aims to improve the performance of the segmentation on real world data, just as the GTA5 dataset.

SHIFT [26] dataset addresses domain shifts in autonomous driving. It captures various real-world conditions using CARLA [4] as the simulation engine. The dataset provides annotations for semantic segmentation and other perception tasks, enabling research in adaptation strategies and robustness assessment, specifically in autonomous driving scenarios. The dataset consists of 2,500,000 annotated frames with fixed and continuously shifting conditions, allowing the development and evaluation of semantic segmentation models under diverse domain shifts.

Acquisition strategies: Synthia, GTA5, and SHIFT datasets were sampled in a random or uniform way, e.g. GTA5 dataset chose every 40th frame. All of them tried to capture a wide variety of weather, viewpoint, and lighting conditions. However, they did not did not choose frames w.r.t. epistemic uncertainties of existing network architectures and did not select frames with regard to the real-world target domain. There are some approaches that present solutions for acquiring corner case data in the synthetic world.

For instance, the approach by Kowol et al. [11] focuses on enriching training data for AI algorithms in automated driving by generating safety-critical driving situations called "corner cases". These were simulated and recorded using CARLA [4]. A test rig is employed, where one operator observes the original virtual image while another operator monitors the output of a real-time semantic segmentation network. The second operator acts as a safety

driver, intervening when the network's assessment is considered unsafe. The objective is to generate corner cases that challenge the network's perception and enhance its performance in safety-critical scenarios. CARLA serves as the simulation engine to create realistic driving scenarios and gather data for training and testing the AI algorithms.

The approach of Möller et al. [15] proposes a novel approach to generate out-of-distribution (OOD) datasets.The generated OOD samples are based on in-distribution datasets using methods like Gaussian Hyperspheric Offset and Soft Brownian Offset. The purpose is to enhance OOD detection and validate the generalization performance of neural networks. The use case includes applications such as object detection, trajectory prediction, and automated driving. However, they do not use simulation engines but models like variational autoencoders.

All of the presented acquisition strategies have in common that they do not explicitly take into account the actual real world target domain for evaluating a frame. There are, however, some "sampling" strategies that have been developed for other tasks that do.

Strategies for transfer learning: Sun et al. [25] propose a transfer learning method for semantic segmentation that is relevant to our work. This method uses labels from both real and synthetic data. It focuses on adaptively selecting similar synthetic regions to bridge the gap between insufficient real data and abundant synthetic data. Hierarchical weighting networks are introduced to score similarity at different levels (pixel, region, and image) and guide the joint learning process. They showcase improvements when training a segmentation model together on the real and the synthetic world.

Kim et al. [10] focus on semi-supervised domain adaptation, where a small fraction of labeled data in the real world target domain is given. They propose the Source Domain Subset Sampling (SDSS) method, which samples data from the synthetic source domain. It involves pixel-level sampling, where a model, pretrained on a small target domain subset, biases the source domain data towards the target domain at the pixel level by removing pixels that do not match the ground truth labels. Additionally, image-level sampling selects a subset of the source domain data based on class balance scoring, prioritizing images with balanced class distributions and accurate samples. By combining these sampling techniques, SDSS extracts a subset of relevant synthetic source domain samples to improve the adaptation process in scenarios with limited labeled target domain data.

2.2. Unsupervised domain adaptation (UDA)

Many synthetic datasets are created to be used to enhance models on real world data. UDA from the synthetic to the real world even allows training without any manual annotation. Since we present strategies to build synthetic datasets for adapting to the specific real world, understating the current principles of UDA are important.

UDA in semantic segmentation is a current field of research that aims to develop methods for adapting DNNs from an annotated source domain [21] to an unlabeled target domain without human labeling effort. As shown in the survey [21], this field has received great attention over recent years. This is especially true for the synthetic to real world domain change. Generally speaking, UDA methods aim to align the distributions in the input (e.g., Hoffman et al.[7], Yunsheng et al.[12], Termöhlen et al.[27] and Yang et al.[30]), the feature (e.g., Niemeijer et al.[17, 16], Hoffman et al.[8] and Marsden et al.[13]), or the output space (e.g., Vu et al.[29], Tsai et al.[28] and Zheng et al.[31]) of a neural network. The recent state of the art is dominated by transformer based architectures that utilize adaptation techniques in the feature and output space as, for instance, presented in [9]. In our experiments, we make use of such architectures.

3. Methods

In this section, we develop an acquisition function to intelligently select important images from the buffer filled constantly by the simulation engine with synthetic images of high variety. With a specified budget B of synthetic images to be acquired, the function aims to improve the performance of the semantic segmentation network on the target domain in a scenario where methods of unsupervised domain adaptation (UDA)–see section 2–are utilized during training. All acquisition functions share the similar generic structure shown in algorithm 1. Each sample from the currently buffered synthetic data receives a score. The score is to reflect the value of the corresponding synthetic image for improving the segmentation performance on the real world target domain. In the end, we select the B images with the largest score from the synthetic world, irrespective of the approach used.

We propose two types of acquisition functions, which can be distinguished by the computation of the scores.

- ***Correspondence-based*** *acquisition functions* leverage synthetic and real-world data to select synthetic images representing meaningful correspondences to the real world target domain. These acquisition functions select synthetic samples that align well with clusters or patterns in the real-world domain.

- ***Simulation-only*** *acquisition functions* focus solely on information and statistics from the synthetic world. The detailed computation of scores in each approach is provided in subsections 3.1 and 3.2.

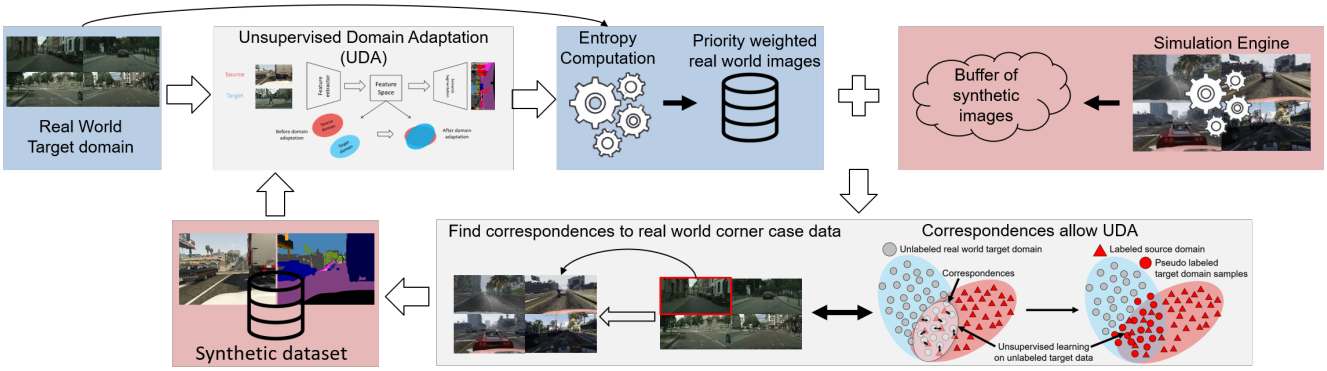

Figure 2. We begin with an initial synthetic training set. Unsupervised Domain Adaptation (UDA) is applied to adapt the model to the real-world target domain. Priority scores are computed for real-world images using the trained model, emphasizing those where the segmentation performs poorly. Synthetic images with the highest correspondence to the prioritized real-world images are selected. The chosen synthetic images are incorporated into the training dataset, enhancing its diversity and improving performance.

Algorithm 1: Acquisition function

```
Data: synthetic_data, real_world_data, budget
Result: selected_data
correspondence_matrix =
  compute_correspondence(synthetic_data,
  real_world_data);
sorted_real_world_data =
  sort_by_priority(real_world_data);
selected_data = [];
while len(selected_data) <budget do
    best_candidate = None;
    best_score = 0;
    foreach synthetic_sample in synthetic_data do
        if synthetic_sample not in selected_data then
            score =
              compute_score(synthetic_sample,
              sorted_real_world_data,
              correspondence_matrix);
            if score > best_score then
                best_score = score;
                best_candidate = synthetic_sample;
            end
        end
    end
    selected_data.append(best_candidate);
end
```

3.1. Simulation-only acquisition functions

Class uniform sampling: This acquisition function aims to achieve a balanced synthetic set w.r.t. the class statistics occurring in the segmentation labels. To keep track of how often a certain class appears in our already chosen data, we build a histogram that is incremented for a certain class if it is present in the segmentation label of the chosen image. During the class uniform sampling, we start with a random selection of an image. After updating the histogram, we compute the class with the lowest value in the histogram and choose an image with a label map that contains that class. The 'compute score' step in algorithm 1 is maximized if the most seldom class in the histogram is present in the synthetic image. If many images fulfill that requirement, a random selection is applied amongst them.

Entropy [23]: Entropy acquisition is based on ideas presented in active learning literature. We compute the semantic segmentation for a given simulated image and the corresponding pixel-wise entropy over the a-posteriori class distributions, representing the value of 'compute score' in algorithm 1. This represents the epistemic uncertainty of the model. We chose the images with the largest epistemic uncertainty until B is full.

CoreSet [22]: The Coreset approach selects a batch of samples that cover the whole data distribution of the simulated world. It formulates this batch selection as a robust k-center selection problem and hence, tackles the redundancy in the simulated world. The value of 'compute score' in algorithm 1 represents how well the distribution of all synthetic images is represented by the acquired batch when this image was added.

3.2. Correspondence-based acquisition functions

The *clustering assumption* of semi-supervised learning (SSL) yields, that if two points belong to the same cluster, their outputs are likely to be close and can be connected by a short curve [1, 14]. In other words, if the synthetic samples are aligned with clusters of the unlabeled real-world samples, it follows the cluster assumption of semi-supervised

learning. In such a case, UDA methods, which are a variant of semi-supervised learning (the target domain is the unlabeled data), should yield good performance. Therefore, to maximize semi-supervised learning performance, newly selected synthetic samples must cover the unlabeled real-world clusters, which are not covered by synthetic samples yet.

Figure 2 shows how we want to take advantage of the cluster assumption. The image represents the general feedback loop for selection. We start with a small initial synthetic set and train a DNN for segmentation while using UDA to the real-world target domain. The UDA method aligns the distributions of the source and target domain in the feature space of the DNN.

Selecting the real world images: Given the *clustering assumption* we now need to find the best correspondences between target and source domain. Therefore, we first analyze three ways of selecting real-world images that serve as targets for correspondence from the synthetic world. In algorithm 1 this is the 'sort_by_priority' function:

- **Global (G):** Choose the synthetic images with the largest correspondence with any real world image.
- **Arbitrary (Arb):** Choose the synthetic image with the largest correspondence with a random subset of real world images.
- **Corner Case (CC):** Choose images with largest correspondence with a set of real world corner case images. To determine the corner case score, we utilize the entropy of the prediction on the real world images.

We assume a given similarity or distance matrix that describes the correspondence of each real-world image with each synthetic image: $C_{ji} = \text{correspondence}(s_j, r_i)$, Where $r_i \in \text{real\_world\_data}$ and $s_j \in \text{synthetic\_data}$. How we compute such a correspondence matrix is shown later in the section. If the correspondence is a distance, we compute the set of synthetic images that minimize the sum of all correspondence values with the real world target images, and vice-versa we maximize the sum of the correspondences if the correspondence represents similarity:

$$\text{compute\_score}(s_j, r_i, C_{ji}) = \min_{s_j} C_{i,j} \tag{1}$$

$$\text{compute\_score}(s_j, r_i, C_{ji}) = \max_{s_j} C_{i,j} \tag{2}$$

The value computed in 'compute score' in algorithm 1, hence is the minimum distance or the maximum similarity. This represents a greedy approximation to minimizing or maximizing the sum of correspondences.

Finding a good measure for correspondence To compute the correspondence matrix, we either need a measure of similarity or dissimilarity (see equations (1) and (2)) that captures the relevant features of the real and synthetic images for semi-supervised learning. For this, we take our current trained model and compute the feature space representation of the real and synthetic images. A synthetic image "corresponds" to a real-world image if the embedding is similar or if the feature space distance is small.

We use cosine similarity as a measure of similarity. A global average pooling over the spatial dimensions reduces the feature space to one vector. As a measure of distance, we use the Hausdorff [5] distance or the Euclidean distance between the Gram matrices [6] of the feature space representations of real world or synthetic images. An essential aspect for both is the encoder that is being used. In our case, we use the feature space of the trained segmentation network in most cases. For some experiments, we also employed a VGG-19 [24] ImageNet [3] encoder.

Deciding on final methods: Finally, we develop and apply different methods from the building blocks we described:

- "CC min/max sim": Indicates the minimization or maximization of the cosine similarity with real world corner case images
- "max sim I-NET": Uses the the VGG-19 ImageNet encoder to compute feature space representation/similarities.
- "Arb max sim": Instead of maximizing the similarity with corner case images, a random corner case score is assigned to each real world image.
- "G max sim": The synthetic images that represent the pairs in the correspondence matrix with the largest similarity
- "Gram Matrix"/"Hausdorf": The distance of Gram matrices or the symmetric Hausdorf distance is used
- "uni": This prefix implies that the class uniform sampling is introduced as a requirement

In some cases, we add class uniformity, as described above, as an additional optimization criterion. In the end, the selection of synthetic images is added to the synthetic training set, and the model in retrained, starting the cycle again.

4. Experiments

This section briefly describes our experimental setup and compares the proposed methods introduced in section 3.

4.1. Setup

During our experiments, we choose the GTA5 dataset to represent our synthetic world. It consists of 24966 synthetic frames (source domain). In each acquisition step, we sample a budget $B = 250$ frames, which represent roughly 1% of the data, five times, which accumulates up to 5% of the whole dataset. This setup would correspond to a system that constantly creates synthetic data with many variations and saves the resulting images into a buffer. Our acquisition functions would select from such a buffer and add the resulting frames to the final set.

We employ the state-of-the-art UDA approach "DAFormer" to adapt to the real world. The model is trained for 20,000 iterations with a batch size of two images. The real-world target domain is represented by the Cityscapes [2] and ACDC [20] dataset. The Cityscapes dataset represents a very "clean" distribution with limited variations w.r.t. illumination, weather, and season conditions. The ACDC dataset represents a distribution with large variability w.r.t. weather (e.g., fog, rain, snow) and illumination (day, night) conditions. Experiments on multiple target domains are important since different acquisition functions might work in different scenarios.

As introduced in section 3, we employ different acquisition functions for the synthetic world. The Entropy, Coreset and, Class uniform acquisitions only depend on the synthetic world, while the selection of synthetic data based on real world data takes the target domain into account. For the latter, we evaluate the variation presented in section 3. Each variation function is evaluated by the maximum Mean Intersection over Union (MIoU) on the Cityscapes and ACDC dataset and the area under the MIoU curve (AUC [14]) that results from the acquisition steps.

4.2. Quantitative Evaluation

In our quantitative evaluation, we assessed the performance of different acquisition functions for the synthetic world as introduced in section 3. We analyze the given acquisition functions on the two domain changes (GTA5 to Cityscapes and GTA5 to ACDC).

GTA5 to Cityscapes Shift: Table 1 presents these values to evaluate the segmentation trained on the GTA5 dataset and adapted to the Cityscapes dataset. Most of the performance can be achieved with a small amount of synthetic data. We achieve a $\max mIoU = 66.17\%$ with 5% of synthetic data only, which is 95.6% of the performance achieved with 100% of the data ($\max mIoU = 67.8\%$). This indicates that the synthetic world contains a lot of redundancy. Except for Entropy, Coreset, and Gram Matrix, all acquisition functions performed better than random sampling. Class uniform sampling proved crucial in the GTA5 to Cityscapes domain change, as having samples from all classes was particularly important for unsupervised domain adaptation (UDA). Notably, the class uniform sampling does not utilize information from the real world target domain but still achieves the best results. However, the Entropy and Coreset acquisition functions that utilize uncertainty or distribution features from the synthetic domain perform worse than functions utilizing the information from the real world domain. W.r.t. the latter class of acquisition functions the Hausdorff and Gram matrix distance as a measure of correspondence performed worse than utilizing the cosine similarity. Interestingly, similarity minimization works better than maximization on GTA5 to Cityscapes domain change if corner case images are chosen for computing correspondences, and no class uniformity is enforced. Introducing the class uniformity requirement does help all acquisition strategies to be on par with the class uniform sampling, which is the best strategy for this domain change. Additionally, to class uniformity introducing a priority for real world corner case images during the computation of

	Acq Function Metric →	**max**	**AUC**
-	Random	64.53	2.36
S	Entropy on GTA5	62.01	2.38
S	Coreset on GTA5	62.93	2.39
S+R	Hausdorf	62.26	2.37
S+R	Gram Matrix	61.94	2.33
S+R	CC min sim	65.78	2.55
S+R	CC max sim	63.57	2.47
S	uni sampling	66.17	2.56
S+R	uni CC max sim	65.87	2.55
S+R	uni CC max sim I-NET	65.47	2.54
S+R	uni CC min sim	65.55	2.53
S+R	uni G max sim	65.55	2.55
S+R	uni Arb max sim	65.35	2.54
–	100% data	67.8	–

Table 1. Domain change from GTA5 to Cityscapes datasets. Maximum mean Intersection over Union (MIoU) in % and the area under the curve after sampling from 1%-5% of the data.

	Acq Function Metric →	**max**	**AUC**
-	Random	46.91	1.82
S+R	CC min sim	46.68	1.8
S+R	CC max sim	47.83	1.84
S	uni sampling	47.5	1.88
S+R	uni CC max sim	49.3	1.98
S+R	uni G max sim	48.89	1.91
–	100% data	46.37	–

Table 2. the GTA5 to ACDC domain change. Maximum mean Intersection over Union (MIoU) in % and the area under the curve after sampling from 1%-5% of the data.

correspondences does not improve the results. Finally, utilizing the VGG-19 ImageNet encoder does not results in a more meaningful correspondence score as the results do not improve (e.g. 'uni CC max sim I-NET' compared to uni 'uni CC max sim').

GTA5 to ACDC Shift: Table 2 presents the GTA5 to ACDC use case results. For this domain shift using 100% of the data seems to decrease performance compared to all acquisition functions, including the random sampling, even though only 1%-5% of the data is sampled. That even indicates that some images may introduce a bias into the model that makes unsupervised learning on the real world more difficult. On the GTA5 to ACDC domain change, the uniform sampling did not give the best performance but still offered an improvement in combination with similarity maximization. Finally, we can observe that, other than in the GTA5 to Cityscapes shift, assigning priority for real world corner case images during correspondence computation ('uni CC max sim') is introducing a benefit in the GTA5 to ACDC domain change (better than 'uni CC max sim').

Comparison: For both domain shifts, very little data achieves similar or better performance on the respective real world target domain. These results highlight the need for more effective acquisition strategies, saving memory and training time, and avoiding a bias in the data that introduces a worse performance despite having more data. The latter problem, which is present in the GTA5 to ACDC shift, probably is due to the fact that the dataset distributions do not overlap as well as the GTA5 and Cityscapes datasets (see figure 5). We conjecture that either 'wrong' information is introduced (image information represents a different label) or that the imbalance of little corresponding relevant and lots of non-corresponding irrelevant data causes the model to assign less weight to the former. We further conjecture that this is part of the reason why similarity maximization combined with class uniformity has the advantage in this domain shift, given that it picks only corresponding information between the domains.

4.3. Qualitative evaluation

Figure 5 displays the TSNE representations of the synthetic and the real world images in the feature space of VGG-19 encoders and the trained segmentation net. The TSNE plots are generated for both the GTA5 to Cityscapes domain change and the GTA5 to ACDC domain change. We can observe that the number of correspondences, i.e., yellow and purple data points close to each other, is sparse in all domain changes and for encoders. This effect even increases in the case of the GTA5 to ACDC domain change. This can be interpreted in different ways. For one, it could be that our encoders do not capture the features important

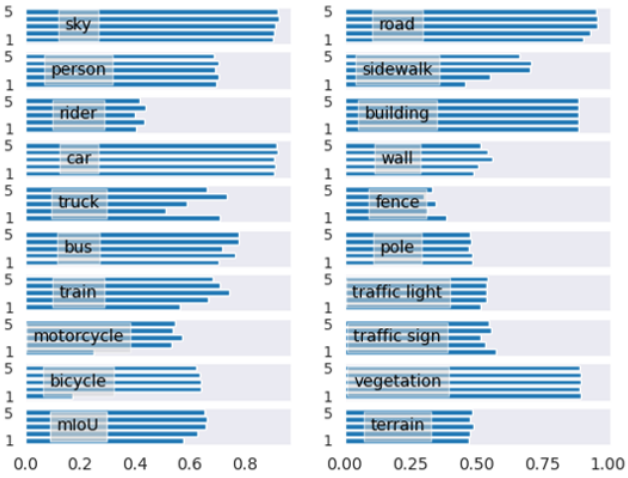

Figure 3. Class-wise IOU over the acquisition steps. The bottom bar is the first iteration the top bar the last. The Cityscapes dataset is the target domain. Class uniform sampling used.

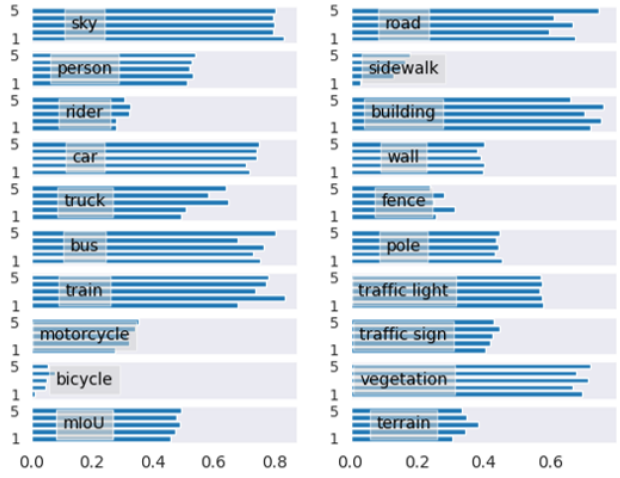

Figure 4. Class wise IOU over the acquisition steps. The bottom bar is the first iteration the top bar the last ACDC is the target domain: Class uniform sampling + similarity maximization used.

for similarity well. On the other hand, it could be a hint that the GTA5 dataset is not ideal for capturing the Cityscapes or ACDC domain. There appears to be more overlap between the Cityscapes and the GTA-5 than between the GTA5 and the ACDC datasets, which could further explain the lower absolute MIoU values obtained in the latter domain change. Given that most of the samples do not overlap between the two domains, this could also be a reason why sampling 100% of the GTA5 data introduces a bias and works worse than sampling 1-5%.

When comparing the ImageNet VGG-19 encoder with the Segmentation network encoder based on their feature representations, the overlap seems to be slightly larger when

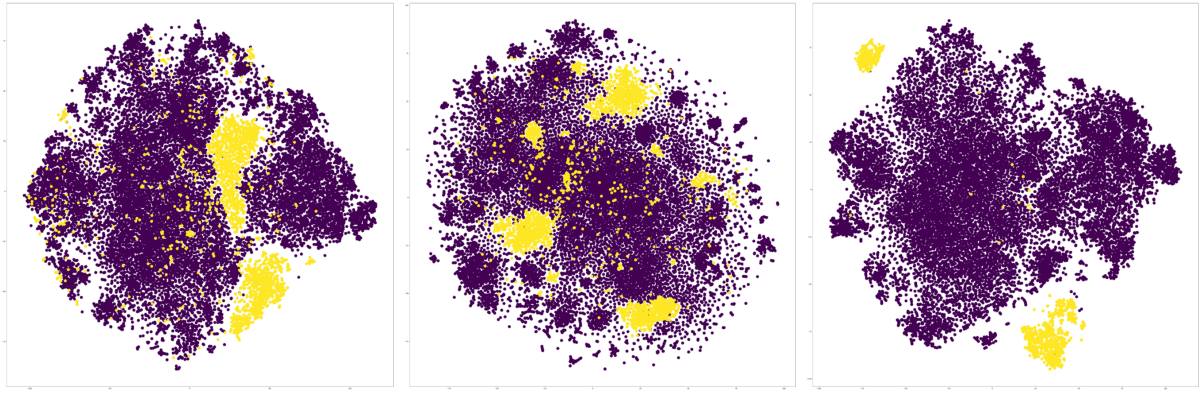

Figure 5. TSNE plots of feature representations: GTA5 to Cityscapes domain change based on Segmentation network (left); GTA5 to Cityscapes domain change based on VGG-19 Image-Net encoder (middle); GTA5 to ACDC domain change based on Segmentation network (right); Feature represenstations of GTA5 are purple and both Cityscapes and ACDC representations are yellow

using the VGG19 encoder. Generally speaking, using the feature representations of the segmentation network encoder makes sense since the displayed distribution is the same that the classification is performed on. Hence similarity in this feature space indicates the similarity of two images w.r.t. the segmentation task.

4.4. Limitations and discussion

The difference in the performance of the studied acquisition functions depending on the domain change has to be analyzed more deeply. Analyzing factors contributing to the varying performance could give us a direction to improve the current functions. We conjecture that enhancing the similarity function, possibly by utilizing a different feature space embedding, could be crucial for better results. Moreover, investigating how particular images or subsets create biases that lead to sub-optimal performance on the target domain is essential to eliminate such issues. As figures 3 and 4 show, even using the current best acquisition functions in the respective domain shift lead to a fluctuation in the class wise IoU values, showing that not all information added over the iterations is having a positive influence. Given the different behaviors of acquisition functions in different domain shifts, the real-world applicability in improving actual perception systems still requires more validation across various real-world scenarios.

5. Conclusion

In this work, we addressed and, to our knowledge, introduced the research question of how to construct a synthetic dataset that facilitates unsupervised adaptation to a specific target domain. Our experiments demonstrated that by sampling a small portion of the synthetic data (1-5%), a performance quite close to or better than achieved with the entire dataset can be obtained. Moreover, we discovered that training on the complete synthetic dataset can introduce a bias and degrade performance when the synthetic domain poorly represents the real-world domain. Through an extensive study of different acquisition functions, we identified their benefits in enhancing adaptation. The findings presented in this paper can be leveraged to create a synthetic dataset tailored to specific real-world target domains.

We propose utilizing the CARLA simulation engine to generate a stream of data and iteratively select the most relevant samples, as demonstrated in this work. The insights gained from this research contribute to advancing the field of synthetic data utilization and have implications for improving perception systems in real-world applications.

A deeper investigation of the acquisition functions' underlying mechanisms is warranted in future work to gain a more comprehensive understanding of their effectiveness across diverse real-world scenarios. Further refinement and optimization of the similarity function, potentially by exploring alternative feature space embeddings, could enhance the acquisition function's performance. Additionally, conducting large-scale validation experiments on various real-world camera systems would validate the applicability and robustness of the proposed approach, thereby paving the way for its wider adoption in practical settings.

Acknowledgement

The research leading to these results is funded by the German Federal Ministry for Economic Affairs and Climate Action within the project "KI Delta Learning" (Forderkennzeichen 19A19013N), "KI Wissen – Entwicklung von Methoden für die Einbindung von Wissen in maschinelles Lernen" and the SynthBAD project. The authors would like to thank the consortium for the successful cooperation. Also funded by the Deutsche Forschungsgemeinschaft (DFG) - 401269959, 417962828.

References

[1] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors. *Semi-Supervised Learning*. The MIT Press, 2006.

[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. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[3] 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.

[4] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.

[5] M-P Dubuisson and Anil K Jain. A modified hausdorff distance for object matching. In *Proceedings of 12th international conference on pattern recognition*, volume 1, pages 566–568. IEEE, 1994.

[6] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style. *CoRR*, abs/1508.06576, 2015.

[7] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR, 2018.

[8] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016.

[9] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9924–9935, 2022.

[10] Daehan Kim, Minseok Seo, Jinsun Park, and Dong-Geol Choi. Source domain subset sampling for semi-supervised domain adaptation in semantic segmentation. *arXiv preprint arXiv:2205.00312*, 2022.

[11] Kamil Kowol, Stefan Bracke, and Hanno Gottschalk. A-eye: Driving with the eyes of ai for corner case generation. *arXiv preprint arXiv:2202.10803*, 2022.

[12] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional Learning for Domain Adaptation of Semantic Segmentation. In *Proc. of CVPR*, pages 6936–6945, Long Beach, CA, USA, June 2019.

[13] Robert A. Marsden, Alexander Bartler, Mario Döbler, and Bin Yang. Contrastive learning and self-training for unsupervised domain adaptation in semantic segmentation, 2021.

[14] Sudhanshu Mittal, Joshua Niemeijer, Jörg P. Schäfer, and Thomas Brox. Best practices in active learning for semantic segmentation. In *German Conference on Pattern Recognition (GCPR)*, 2023.

[15] Felix Moller, Diego Botache, Denis Huseljic, Florian Heidecker, Maarten Bieshaar, and Bernhard Sick. Out-of-distribution detection and generation using soft brownian offset sampling and autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 46–55, 2021.

[16] Joshua Niemeijer and Jörg Peter Schäfer. Domain adaptation and generalization: A low-complexity approach. In *6th Annual Conference on Robot Learning*, 2022.

[17] Joshua Niemeijer and Jörg P. Schäfer. Combining semantic self-supervision and self-training for domain adaptation in semantic segmentation. In *2021 IEEE Intelligent Vehicles Symposium Workshops (IV Workshops)*, pages 364–371, 2021.

[18] 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.

[19] 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.

[20] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10765–10775, 2021.

[21] Manuel Schwonberg, Joshua Niemeijer, Jan-Aike Termöhlen, Jörg P Schäfer, Nico M Schmidt, Hanno Gottschalk, and Tim Fingscheidt. Survey on unsupervised domain adaptation for semantic segmentation for visual perception in automated driving. *IEEE Access*, 2023.

[22] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

[23] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55, 2001.

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[25] Ruoqi Sun, Xinge Zhu, Chongruo Wu, Chen Huang, Jianping Shi, and Lizhuang Ma. Not all areas are equal: Transfer learning for semantic segmentation via hierarchical region selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[26] Tao Sun, Mattia Segu, Janis Postels, Yuxuan Wang, Luc Van Gool, Bernt Schiele, Federico Tombari, and Fisher Yu. Shift: A synthetic driving dataset for continuous multi-task domain adaptation, 2022.

[27] Jan-Aike Termöhlen, Marvin Klingner, Leon J. Brettin, Nico M. Schmidt, and Tim Fingscheidt. Continual unsupervised domain adaptation for semantic segmentation by online frequency domain style transfer. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 2881–2888, 2021.

[28] 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 *Proc. of CVPR*, pages 7472–7481, Salt Lake City, UT, USA, June 2018.

[29] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019.

[30] Yanchao Yang and Stefano Soatto. FDA: Fourier Domain Adaptation for Semantic Segmentation. In *Proc. of CVPR*, pages 4085–4095, Seattle, WA, USA, June 2020.

[31] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision*, pages 1–15, 2021.