

This work was written as part of one of the author's official duties as an Employee of the United States Government and is therefore a work of the United States Government. In accordance with 17 U.S.C. 105, no copyright protection is available for such works under U.S. Law.

Public Domain Mark 1.0

<https://creativecommons.org/publicdomain/mark/1.0/>

Access to this work was provided by the University of Maryland, Baltimore County (UMBC) ScholarWorks@UMBC digital repository on the Maryland Shared Open Access (MD-SOAR) platform.

Please provide feedback

Please support the ScholarWorks@UMBC repository by emailing scholarworks-group@umbc.edu and telling us what having access to this work means to you and why it's important to you. Thank you.

PROCEEDINGS OF SPIE

SPIDigitalLibrary.org/conference-proceedings-of-spie

Benchmarking domain adaptation for semantic segmentation

Masud Ahmed, Zahid Hasan, Naima Khan, Nirmalya Roy, Sanjay Purushotham, et al.

Masud Ahmed, Zahid Hasan, Naima Khan, Nirmalya Roy, Sanjay Purushotham, Aryya Gangopadhyay, Suyu You, "Benchmarking domain adaptation for semantic segmentation," Proc. SPIE 12124, Unmanned Systems Technology XXIV, 121240F (31 May 2022); doi: 10.1117/12.2618548

SPIE.

Event: SPIE Defense + Commercial Sensing, 2022, Orlando, Florida, United States

Benchmarking domain adaptation for semantic segmentation

Masud Ahmed^a, Zahid Hasan^a, Naima Khan^a, Nirmalya Roy^a, Sanjay Purushotham^a, Aryya Gangopadhyay^a, and Suyu You^b

^aUniversity of Maryland Baltimore Count, USA

^bArmy Research Lab, USA

ABSTRACT

Deep Learning (DL) requires a massive, labeled dataset for supervised semantic segmentation. Getting massive labeled data under a new setting (target domain) to perform semantic segmentation requires huge efforts in time and resources. One possible solution is domain adaptation (DA) where researchers transform the data distribution of existent annotated public data (source domain) to resemble the target domain. We develop a model on this transformed data. Nevertheless, this poses the questions of what source domain/s to utilize, and what types of transformation to perform on that domain/s. In this research work, we study those answers by benchmarking different data transformation approaches on source-only and single-source domain adaptation setups. We provide a new well-suited dataset using unmanned ground vehicle Husarion ROSbot 2.0 to analyze and demonstrate the relative performance of different DA approaches.

Keywords: Domain adaptation, semantic segmentation, adversarial learning

1. INTRODUCTION

Semantic segmentation provides the per-pixel prediction of objects to a given image, resulting in a detailed scene description that includes object class, position, and shape information. It has become an active research area in robotics, medical and other fields. The semantic segmentation process helps a robot to understand its surrounding autonomously. Recently, the availability of large-scale annotated segmentation datasets has facilitated the training of various Convolution Neural Network (CNN) based architectures to achieve state-of-the-art performance in the segmentation tasks. However, these trained models' generalization significantly depends on the selected training dataset distribution. For instance, it frequently fails to perform consistent semantic segmentation under new environments, whose distribution varies with the development dataset due to camera sensors, environments, applications, backgrounds, etc. The trivial solution for such cases involves developing semantic segmentation for encountered environments from scratch. It requires heavy resource collection and annotation of data in a specific domain. Further, it fails to comprehend the already learned knowledge from another data domain. Domain adaptation (DA) techniques overcome both of these issues by enabling high-performance model development from the already existing developed models.

DA approaches try to transfer the knowledge of a learned model on specific data (source domain) for fast adaption in a different but related new scenario (target domain) using minimal-labeled or unlabeled target domain data. Existing DA methods have shown their potential in various computer vision semantic segmentation tasks. However, it fails frequently during large domain shifts between source and a target domain and encounter performance degradation for both environments. For example, two available massive public datasets CityScape¹ and Rellis-3D² have substantial distribution differences due to environment and application scenarios. In our case, developing a baseline model using Rellis-3D as source domain and applying DA (Adversarial Domain Adaptation) for CityScape's target domain degrades the model performance and fails to adapt for CityScape environments⁷.

In this study, we empirically investigate these underlying reasons that contributes DA methods failures for CityScape,¹ Rellis-3D,² SemanticKITTI.³ We observe a fundamental concern raised by the domain adaptation

Further author information: (Send correspondence to Masud Ahmed)

E-mail: mahmed10@umbc.edu

technique about the proper knowledge transfer. Based on the statistical distance of the probability distribution between two domains, we encounter the unpredictable behavior of the DA methods. In these circumstances, before applying the domain adaptation technique, it requires specification and idea about the task, data distribution, and how much knowledge can pass from the source domain to the target domain. Based on the knowledge about quality and types of target and source data domain, dataset-specific changes can result in successful DA and train agents from the source domain to perform the objective task under the target environment.

Based on our experimental analysis, we put-forth some limitations and applicable boundary of different domain adaptation methods by investigating their capability for CityScape,¹ Rellis-3D,² SemanticKITTI³ and pose the following contributions to the DA for segmentation research.

- We investigate the reasons that bars the generalization of the domain adaptation application and propose data specific model changes to perform domain adaptation for multiple dataset that were previously not done.
- We provide extensive analysis of model performance with different dataset (CityScape,¹ Rellis-3D,² SemanticKITTI³) and with & without domain adaptation techniques. We also evaluate the network with our challenging dataset that contains both rural and urban environments sample with varying lighting conditions.

2. RELATED WORKS

Domain Adaptation is a well-studied issue for classification and detection tasks in the field of machine learning and deep learning. Besides, in image segmentation, the domain adaptation technique has been opted in other fields too, e.g., sensor-based human activity recognition,^{4,5} face recognition,⁶ image-to-image translation,^{7,8} etc. In the field of semantic segmentation, Chen et al.,⁹ Zhang et al.¹⁰ are one of the successfully implemented domain adaptation techniques at the earlier stage. The majority of the prior domain adaptation techniques on semantic segmentation are unsupervised, where the model is trained with the source image, source label, and target image. The major task of this approach is to align the feature of the target domain like the source domain. Researchers opted for several techniques to achieve the goal.

The most often used method for semantic segmentation is the adversarial learning approach. This sort of model is consists of two networks. One is the segmentation network, and another is the discriminator network. The segmentation network generates the segmented image, whereas the discriminator predicts whether the segmented image is generated from the source domain or target domain and sends feedback to the segmentation network. The segmentation network improves its performance based on the feedback and tries to generate a more accurate segmented image. The loop keeps going on till the segmentation network can fool the discriminator network, resulting in a comparable distribution of features vectors between the two domains. In recent days several adversarial learning-based semantic segmentation approaches have been proposed, e.g., Luo et al.,¹¹ Vu et al.,¹² Mei et al.,¹³ Pan et al.¹⁴

Besides the adversarial approach, the generative approach is another popular technique in this field. CycleGAN,¹⁵ DCGAN¹⁶ are commonly used network in this approach. In this approach, a generative network is implemented before the segmentation network. The generative network reconstructs the target image/feature vector and aligns it with the source image/feature vector and vice versa. After the alignment, the segmentation network will segment the image. Several generative-based semantic segmentation approaches have been proposed, e.g., Choi et al.,¹⁷ Pizzati et al.¹⁸ Besides, some techniques combine both adversarial and generative approaches. Zhang et al.¹⁹ combines the generative approach for appearance adaptation and adversarial training for representation adaptation, a generative network is employed for appearance transformation and an adversarial network for representation transformation. We use adversarial approach in our experiments.

3. METHODOLOGY

In this section, we describe the unsupervised domain adaptation. The Unsupervised domain adaptation attempts to minimize the distance between the distributions of intermediate features from source and target domains. We

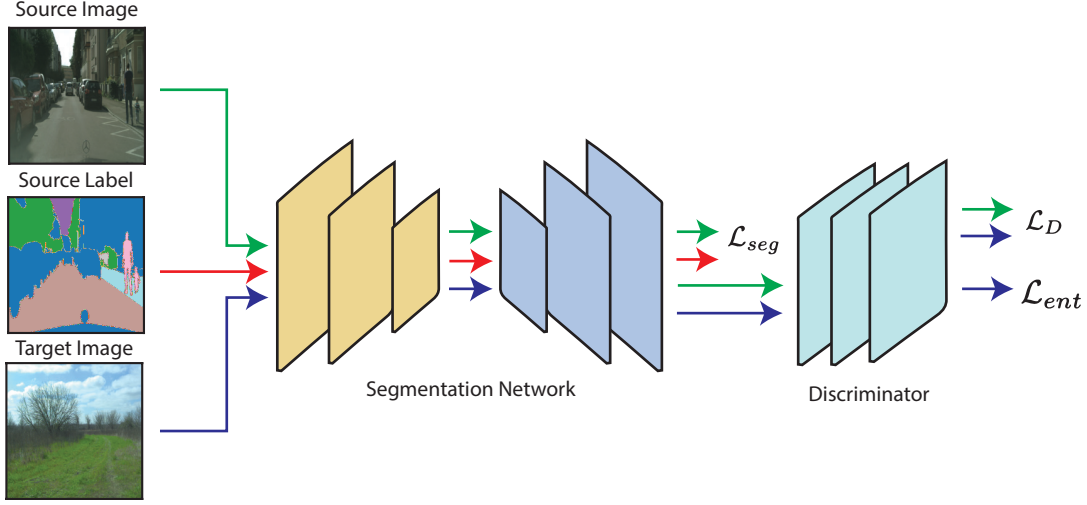


Figure 1. Overall architecture of the Model

integrate the entropy minimization principle of unsupervised cross-domain domain adaptation¹² as shown in Fig 1.

Assume the set of images from source domain \mathcal{X}_s and the associated segmentation labels \mathcal{Y}_s from C classes. Each of the RGB images x_s in \mathcal{X}_s is of $H \times T \times 3$ sizes for which we obtain the corresponding labels y_s as one-hot vector for each pixels at (h, w) . We learn a segmentation network \mathcal{S} which provides segmentation maps $[P_x^{h,w,c}]_{h,w,c}$ for images in \mathcal{X}_s . In this work, we considered U-net by Ronneberger et al.²⁰ as \mathcal{S} . However, the parameters θ_S of U-net network \mathcal{S} is learned to minimize the segmentation loss \mathcal{L}_{seg} . For an image x_s with true segmentation labels y_s can be expressed as following:

$$\mathcal{L}_{seg}(x_s, y_s) = - \sum_{h=1}^H \sum_{w=1}^W \sum_{c=1}^C y_s^{(h,w,c)} \log P_{x_s}^{(h,w,c)} \quad (1)$$

where, $P_{x_s}^{(h,w,c)}$ provides discrete distribution over all classes. If the score for one class is high, we obtain low entropy while for evenly spread scores, we obtain higher entropy. For target domain, the truth labels y_t are not available. Therefore, we try to predict segmentation labels with high confidence by minimizing the entropy of prediction for target domain. We compute entropy $E_{x_t} \in [0, 1]$ for pixel at (h, w) in a target domain image x_t as follows:

$$E_{x_t}^{(h,w)} = \frac{-1}{\log(C)} \sum_{c=1}^C P_{x_t}^{(h,w,c)} \log P_{x_t}^{(h,w,c)} \quad (2)$$

We obtain the entropy loss \mathcal{L}_{ent} for an image x_t by combining the pixel-wise normalized entropies $\mathcal{L}_{ent} = \sum_{h,w} E_{x_t}^{(h,w)}$. Later, we combine the supervised segmentation loss \mathcal{L}_{seg} from source domain and unsupervised entropy loss \mathcal{L}_{ent} from target domain for optimization in the training phase as expressed in equation 3

$$\min_{\theta_S} \frac{1}{|\mathcal{X}_s|} \sum_{x_s} \mathcal{L}_{seg}(x_s, y_s) + \frac{\lambda_{ent}}{|\mathcal{X}_t|} \sum_{x_t} \mathcal{L}_{ent}(x_t) \quad (3)$$

where λ_{ent} is an weighting factor for \mathcal{L}_{ent} . We followed the iterative self-training approach to compute the entropy loss $\mathcal{L}_{ent}(x_t)$ by using pseudo-labels for target images. For self-training, a set of pixels is assumed to have higher prediction scores with higher probability compared to a fixed or scheduled threshold. This allows using cross-entropy loss with pseudo-labels on predictions for target images as the soft-assignment of $\mathcal{L}_{ent}(x_t)$. In self-training approach, the training objective expressed in equation 3 turns out as follows:

$$\min_{\theta_S} \frac{1}{|\mathcal{X}_s|} \sum_{x_s} \mathcal{L}_{seg}(x_s, y_s) + \frac{\lambda_{pl}}{|\mathcal{X}_t|} \sum_{x_t} \mathcal{L}_{seg}(x_t, \bar{y}_t) \quad (4)$$

where \bar{y}_t is the one-hot class prediction for x_t and similar as $\mathcal{L}_{seg}(x_s, y_s)$ as expressed in equation 1, $\mathcal{L}_{seg}(x_t, \bar{y}_t)$ can be expressed as follows:

$$\mathcal{L}_{seg}(x_t, \bar{y}_t) = - \sum_{h=1}^H \sum_{w=1}^W \sum_{c=1}^C \bar{y}_t^{(h,w,c)} \log P_{x_t}^{(h,w,c)} \quad (5)$$

However, only minimizing the pixel-wise prediction entropies overlooks the structural properties while adaptation on structured output space is useful for unsupervised domain adaptation. Minimizing the distance of weighted self-information maps between source, I_x and target I_{x_t} is equivalent to minimizing the entropy loss. Therefore, a unified adversarial training framework is trained to reduce the target entropy loss by making the entropy distribution of target similar to that of source. The weighted self-information maps I_x is passed to a fully-conventional discriminator network \mathcal{D} with parameters θ_D which classifies the domain as source (1) or target (0). The segmentation network \mathcal{S} is trained to adapt the weighted self-information space from source to target and fool discriminator in classifying the weight self-information maps as source or target while the discriminator network is trained to improve the classification. The training objective of the discriminator is as

$$\min_{\theta_D} \frac{1}{|\mathcal{X}_s|} \sum_{x_s} \mathcal{L}_D(I_x, 1) + \frac{1}{|\mathcal{X}_t|} \sum_{x_t} \mathcal{L}_D(I_{x_t}, 0) \quad (6)$$

and the adversarial objective to train the segmentation network is

$$\min_{\theta_S} \frac{1}{|\mathcal{X}_t|} \sum_{x_t} \mathcal{L}_D(I_{x_t}, 1) \quad (7)$$

During training, the segmentation and discriminator networks are optimized alternatively the final optimization problem is as follows:

$$\min_{\theta_S} \frac{1}{|\mathcal{X}_s|} \sum_{x_s} \mathcal{L}_{seg}(x_s, y_s) + \frac{\lambda}{|\mathcal{X}_t|} \sum_{x_t} \mathcal{L}_D(I_{x_t}, 1) \quad (8)$$

4. EXPERIMENTAL SETUP

To analyze the domain adaptation technique, we use four datasets, Rellis-3D, SemanticKITTI, CityScape, and our dataset. As the base network for semantic segmentation, we use the U-net model.²⁰ In the following subsections, we briefly introduce the datasets used for experiments and the architecture of U-net and adversarial network used for the domain adaptation.

4.1 Dataset

The description of the datasets is given as follows.

4.1.1 CityScape

CityScape¹ dataset is a large scale dataset containing video sequences of urban street scenes from 50 different cities. The dataset was created with the intention of better evaluation of deep learning based vision algorithms for semantic scene understanding in urban environment. The dataset contains pixel-level fine annotations for 5000 frames covering 30 classes from different time and season of the day and the year. Similar type of objects are grouped together into one class. For example, riders or passers-by are considered as 'human' class, all types of transports are in vehicle class, different types of infrastructure are included in construction class. The annotations provides the instance segmentation of vehicle and people with polygons. The labeling policy prioritized the foreground classes over the background classes. This refers when two or more classes are present in one region, background classes are ignored. One image from a video snippet is selected for annotation. There are also several metadata, such as, GPS coordinates, outside temperature from vehicle sensor etc. available.

4.1.2 SemanticKITTI

Based on KITTI Vision Benchmark,²¹ a large dataset named SemanticKITTI³ is constructed. It is also a multi-modal dataset with rgb and grey images including visual odometry and gps points. The dataset captures videos from rural areas and highways. Each image contains at most 15 cars and 30 pedestrians. There are 28 classes which annotates both moving and non-moving objects. Moving humans are annotated as bicyclist, motorcyclist. Among other functional classes there are road, sidewalk, parking, building, different types of objects e.g., pole, fence, traffic-signs etc.

4.1.3 Rellis-3D

Rellis-3D² is a multi-modal dataset consisting of off-road environment. This dataset contains 6235 images along with 13556 LiDAR scans. The dataset was created with the motivation of contributing to training of existing semantic segmentation algorithm by providing data from natural environment. It includes fine-grained annotations for various types of terrains. Along with the RGB images, this dataset also contains LiDAR point clouds, pairs of stereo images, GPS measurements with high precision. Overall the dataset annotates 20 classes. However, this dataset considers detailed annotations for earth surfaces and water sources. The classes includes mud, asphalt, gravel, mulch, rubble piles, puddle, deep water etc.

4.1.4 Our dataset

We collect novel data using small unmanned ground bot Husarion ROSbot 2.0 *. ROSbot 2.0 has Rockchip RK3288, Quad-core ARM Cortex-A17 32-bit processor with 2GB LPDDR3 RAM and ARM Mali-T764 MP2. Also, it has an Orbbec Astra RGB camera, which is used for data collection. Using this bot, we collect data from the University of Baltimore, Maryland County. The dataset contains both urban and rural environments. The variation of lighting conditions is also available in the dataset. Our dataset is still in progress of annotation and we are planning to make it publicly available soon.

4.2 U-net Model Architecture

The U-net model consists of an encoder and a decoder network. The architecture of the model is depicted in Fig 2. First, the encoder layer encodes the features of the image. 3×3 convolution filters with a stride size 1

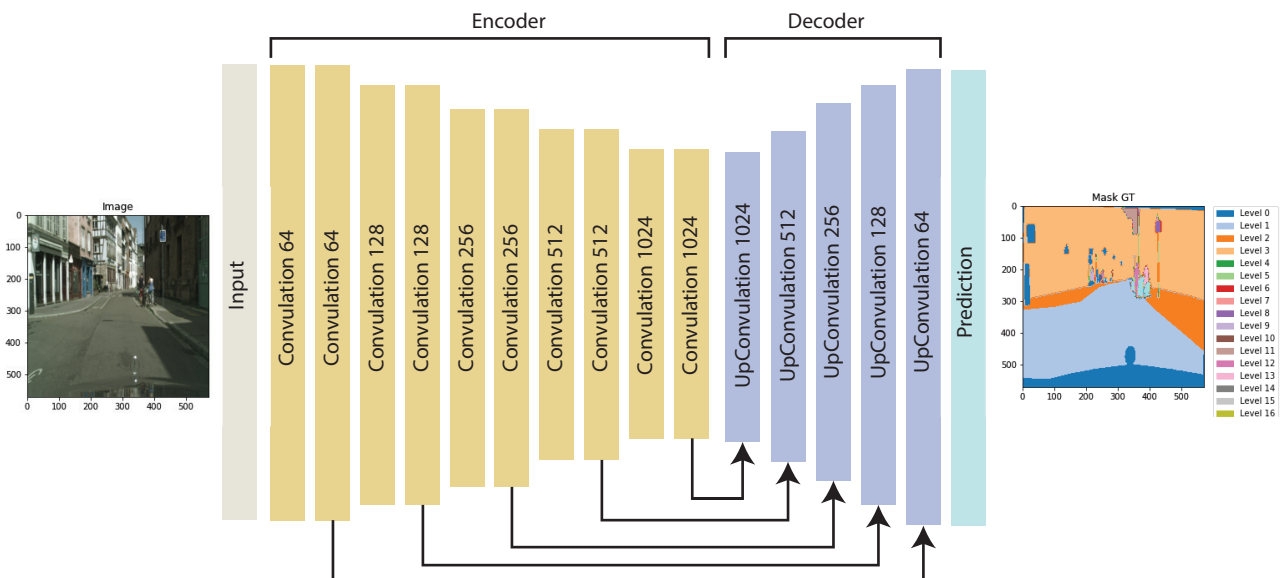


Figure 2. Architecture of U-net model

are piled on top of each other in the Encoder layer. Following each convolution, a rectified linear unit (ReLU)

*<https://store.husarion.com/collections/dev-kits/products/rosbot>

and a 2×2 max-pooling operation with stride size 2 are used to downsample the data. On the other hand, the decoder layer decodes the feature information and constructs segmented images. It consists of transpose convolution layers with filter size 2×2 and strides size 2. Unlike the encoder network, the decoder network does not contain any max-pooling layer. A convolution with filter size 1×1 and stride size 1 is employed at the final layer to convert each feature vector to the desired number of classes and generate the segmented image.

4.3 Adversarial Network Architecture

For adversarial learning, we use a similar discriminator network from the DCGAN¹⁶ shown in Fig 3. Segmented

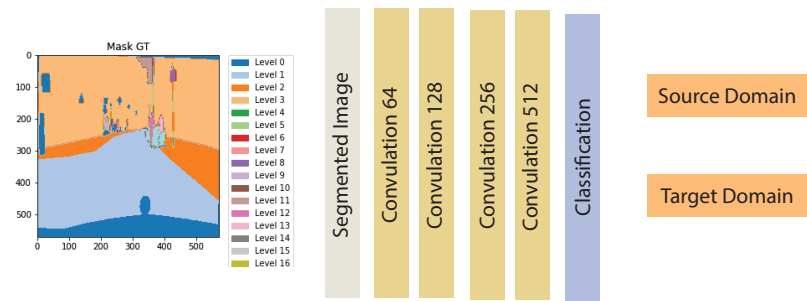


Figure 3. Architecture of discriminator network

image from the U-net model is fed to the network. The discriminator network consists of 4 convolution filters with a filter size 4×4 and stride 2. After each convolution filter, a LeakyReLU activation function with a fixed slope of -0.2 is used. Finally, a classifier layer generates classification outputs that indicate whether the segmented image belongs to the source or target domain. This adversarial network helps the U-net model to train in a more general way.

5. RESULT AND ANALYSIS

All of the dataset mentioned above has different sort of distribution. We utilize KL (Kullback–Leibler)²² and JS (Jensen–Shannon)²³ divergence to determine the probability distribution distance between the two datasets. We also consider only the same pixel labels in both datasets for the semantic segmentation task. Otherwise, the model will ignore that pixel label as there is no extra knowledge to transfer. The model will not learn these unique instances.

5.1 CityScope & SemanticKITTI

Both of these datasets are captured in urban environment. The KL divergence from CityScope to SemanticKITTI is 0.35, and from SemanticKITTI to CityScope is 0.55. The JS divergence is 0.36. Common classes between the two datasets are 19. There are some unique classes in CityScope dataset that we ignore. The ignored data instances from the CityScope dataset are Ego Vehicle, Rectification Border, Guard Rail, Bridge, Tunnel, Trailer, Caravan, etc.

The results of the domain adaptation technique on the SemanticKITTI dataset of models trained on Cityscope are shown in Table 1 and output of the model on SemanticKITTI dataset is depicted in Fig 4.

Table 1. From Cityscope to SemanticKITTI

Model	Domain	Road	Sidewalk	Building	Wall	Fence	Pole	Traffic Light	Traffic Signal	Vegetation	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motorcycle	Bicycle	Pixel Accuracy
Without DA	Source	98.22	90.70	92.52	63.24	68.85	44.27	22.24	48.50	92.27	74.45	91.68	65.50	16.64	92.41	59.55	81.36	84.25	28.78	63.14	93.36
	Target	50.74	14.10	40.39	1.49	4.91	12.55	1.79	5.76	61.45	24.71	54.00	9.36	0.62	33.72	2.75	1.50	1.13	0.58	6.21	53.68
With DA	Source	98.14	90.09	92.63	62.31	67.69	45.37	17.71	47.54	92.48	74.39	91.95	65.43	8.20	93.84	61.37	81.36	84.69	10.71	59.14	93.03
	Target	56.69	11.97	43.65	2.33	6.79	13.89	2.20	6.47	68.97	29.69	38.89	13.27	0.63	46.98	0.84	1.57	1.51	0.33	5.49	57.20

The results of the domain adaptation technique on the Cityscope dataset of models trained on SemanticKITTI are shown in Table 2 and output of the model on SemanticKITTI dataset is depicted in Fig 5.

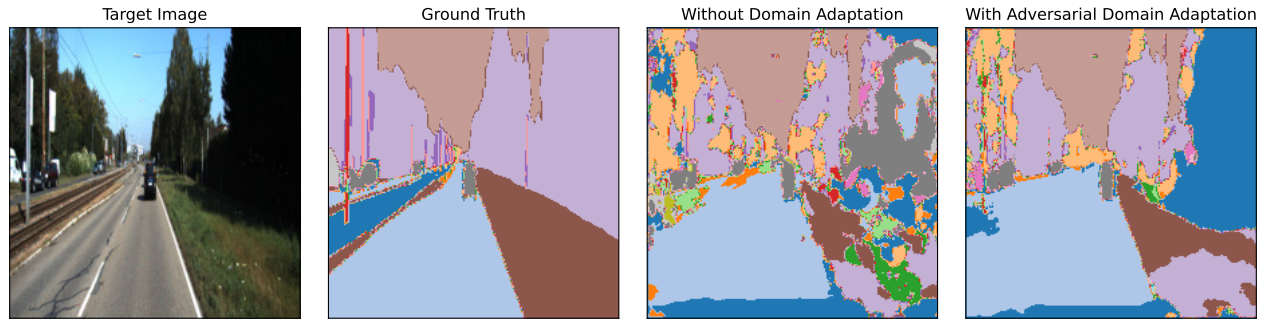


Figure 4. Output on Target dataset from Cityscape to SemanticKITTI

Table 2. From SemanticKITTI to Cityscape

Model	Domain	Road	Sidewalk	Building	Wall	Fence	Pole	Traffic Light	Traffic Signal	Vegetation	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motorcycle	Bicycle	Pixel Accuracy
Without DA	Source	95.14	90.16	93.36	59.99	77.98	52.85	35.02	55.89	96.19	89.28	96.15	65.19	19.73	93.57	88.40	85.29	82.98	0.00	78.18	95.31
	Target	47.80	12.86	42.00	1.03	1.03	6.31	1.29	3.98	59.29	19.39	39.37	7.61	0.06	26.85	1.58	0.93	0.59	0.00	5.49	47.00
With DA	Source	95.57	91.46	94.37	59.97	78.26	52.87	33.63	55.72	96.02	89.03	95.72	66.91	0.90	92.75	87.48	84.50	83.79	0.00	77.08	95.27
	Target	54.17	18.90	43.43	0.71	1.06	8.89	0.64	6.40	61.14	29.35	71.23	5.37	0.07	42.77	0.41	0.41	0.14	0.00	2.69	56.78

5.2 CityScope & Rellis-3D

In this case, data are captured in two different configurations. Rellis-3D data are captured in the rural environment, whereas CityScope captured data in the urban environment. The KL divergence from CityScope to Rellis-3D is 0.84, and from Rellis-3D to CityScope is 0.71. The JS divergence is 0.51. Common classes between the two datasets are 9. There are unique classes in both datasets, e.g., log, barrier, mud, puddle, rubble, etc., in the Rellis-3D dataset. The unique classes from the Cityscope dataset are building, all sorts of vehicles, traffic signals, lights, bridges, border guards, etc. Also, the same instances in the Rellis-3D are labeled with a different name; in this case, we follow the label pattern of the Cityscope dataset.

The results of the domain adaptation technique on the Rellis-3D dataset of models trained on Cityscape are shown in Table 3 and output of the model on Rellis-3D dataset is depicted in Fig 6.

Table 3. From Cityscape to Rellis-3D

Model	Domain	Terrain	Vegetation	Pole	Sky	Road	Person	Fence	Sidewalk	Pixel Accuracy
Without DA	Source	75.66	93.40	44.89	92.41	98.16	71.56	69.39	90.67	95.55
	Target	13.03	9.22	0.12	11.69	0.22	0.48	0.15	3.63	26.75
With DA	Source	74.96	93.06	46.13	92.53	98.24	71.43	68.97	91.21	95.52
	Target	1.26	4.72	0.07	0.69	0.14	1.82	0.04	2.04	31.63

The results of the domain adaptation technique on the Cityscape dataset of models trained on Rellis-3D are shown in Table 4 and output of the model on Cityscape dataset is depicted in Fig 7.

Table 4. From Rellis-3D to Cityscape

Model	Domain	Terrain	Vegetation	Pole	Sky	Road	Person	Fence	Sidewalk	Pixel Accuracy
Without DA	Source	96.51	94.23	0.00	97.69	0.00	30.06	0.00	91.65	96.40
	Target	2.12	23.11	0.00	27.31	0.00	0.03	0.00	7.63	20.18
With DA	Source	96.45	94.20	0.00	97.65	39.59	30.78	0.00	90.82	96.48
	Target	1.11	26.60	0.00	50.52	0.0006	0.0002	0.00	0.11	22.13

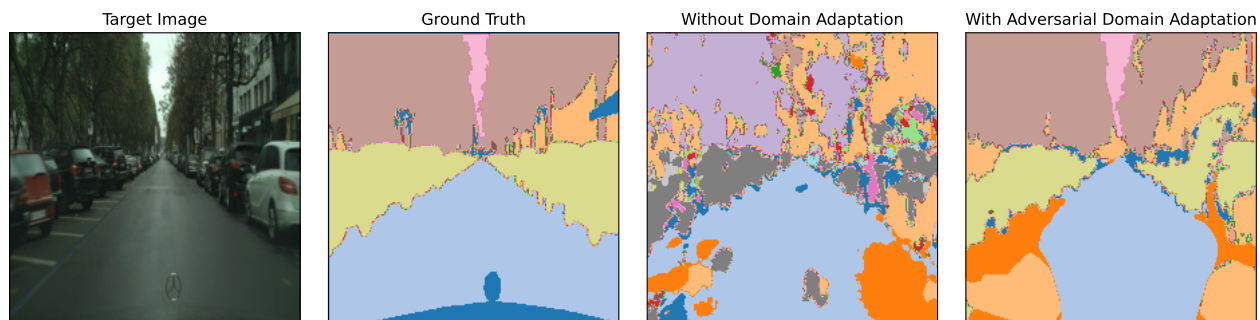


Figure 5. Output on Target dataset from SemanticKITTI to Cityscape

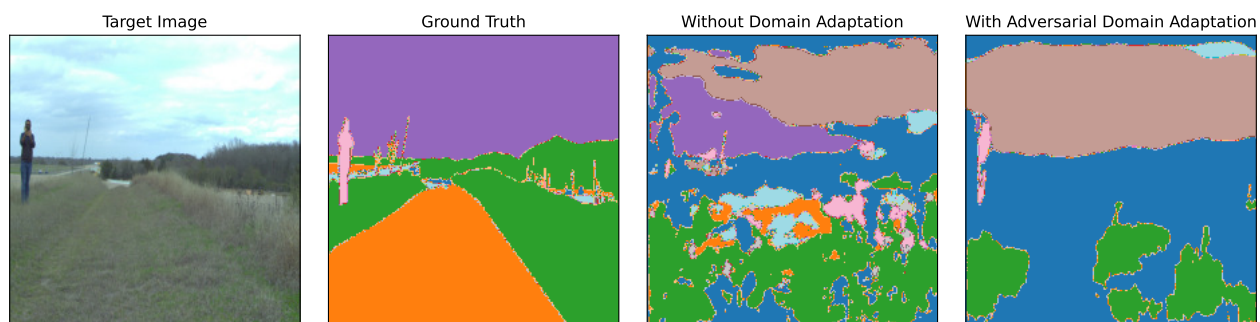


Figure 6. Output on Target dataset from Cityscape to Rellis-3D

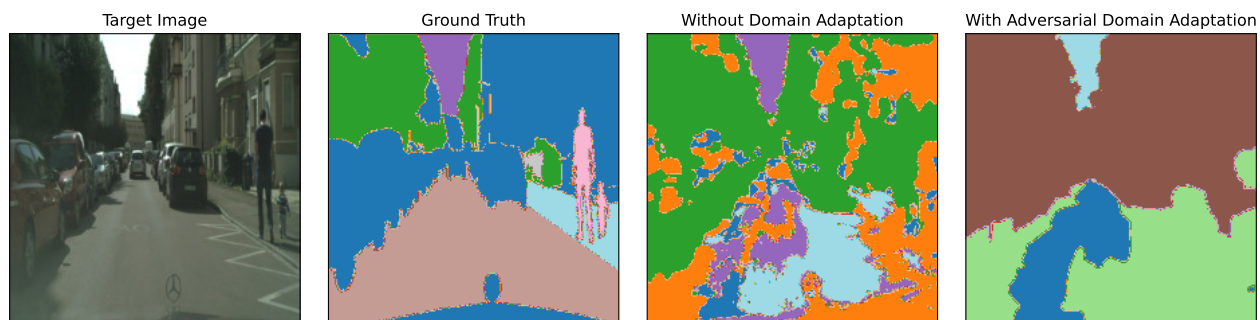


Figure 7. Output on Target dataset from Rellis-3D to Cityscape

5.3 Rellis-3D & SemanticKITTI

This case is similar to the previous case. Rellis-3D data are captured in the rural environment, and like CityScape, SemanticKITTI captured data in the urban environment. The KL divergence from SemanticKITTI to Rellis-3D is 1.09, and from Rellis-3D to SemanticKITTI is 0.51. The JS divergence is 0.68. Common classes between the two datasets are 9. Like the previous case, unique classes from Rellis-3D are log, barrier, mud, puddle, rubble, etc., and unique classes from the SemanticKITTI dataset are building, all sorts of vehicles, traffic signals, lights, bridges, border guards, etc. Also, the same instances in the Rellis-3D are labeled with a different name; in this case, we follow the label pattern of the SemanticKITTI dataset.

The results of the domain adaptation technique on the SemanticKITTI dataset of models trained on Rellis-3D are shown in Table 5 and output of the model on SemanticKITTI dataset is depicted in Fig 8.

Table 5. From Rellis-3D to SemanticKITTI

Model	Domain	Terrain	Vegetation	Pole	Sky	Road	Person	Fence	Sidewalk	Pixel Accuracy
Without DA	Source	96.31	94.28	0.00	97.57	0.00	0.00	0.01	91.15	96.28
	Target	13.05	41.70	0.00	58.07	0.00	0.00	0.00	7.46	26.07
With DA	Source	96.27	94.25	0.00	97.36	30.49	0.01	0.00	91.11	96.26
	Target	11.42	41.99	0.00	66.72	0.0003	0.00	0.00	0.03	29.69

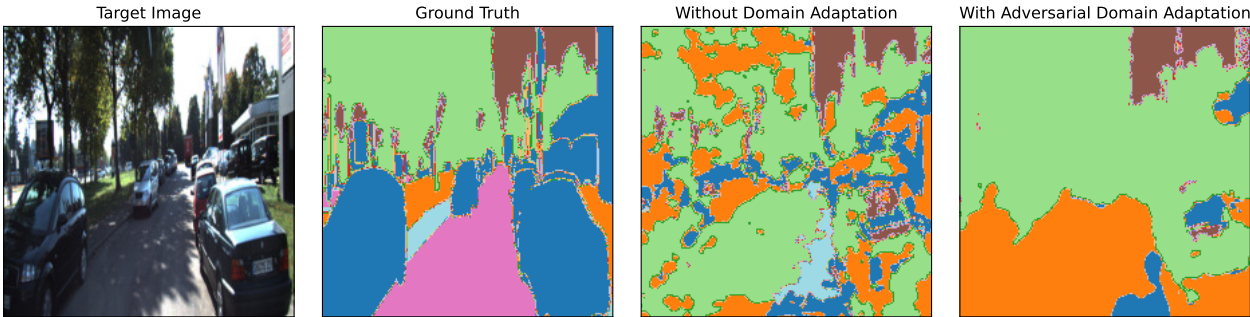


Figure 8. Output on Target dataset from Rellis-3D to SemanticKITTI

The results of the domain adaptation technique on the SemanticKITTI dataset of models trained on Rellis-3D are shown in Table 6 and output of the model on Rellis-3D dataset is depicted in Fig 9.

Table 6. From SemanticKITTI to Rellis-3D

Model	Domain	Terrain	Vegetation	Pole	Sky	Road	Person	Fence	Sidewalk	Pixel Accuracy
Without DA	Source	89.53	96.36	53.43	96.25	95.81	67.13	78.87	91.30	94.60
	Target	32.93	14.76	0.21	36.05	0.28	0.17	0.11	0.99	28.85
With DA	Source	88.69	96.06	53.39	96.26	95.71	68.17	77.68	91.12	94.67
	Target	6.64	10.90	0.25	32.21	0.16	0.07	0.08	4.34	36.85

5.4 Evaluation on Our Dataset

We implement the domain adaptation technique on our dataset using all three datasets as the source domain. We consider all sorts of possible classes from these three datasets. Although we have not reported any quantitative analysis using our dataset, we provide a qualitative analysis of our DA approach. As shown in Fig 10, our DA approach performs well on diverse classes of our dataset for both rural and urban environments. The performance looks visually promising, even though we use only the unannotated images of our dataset.

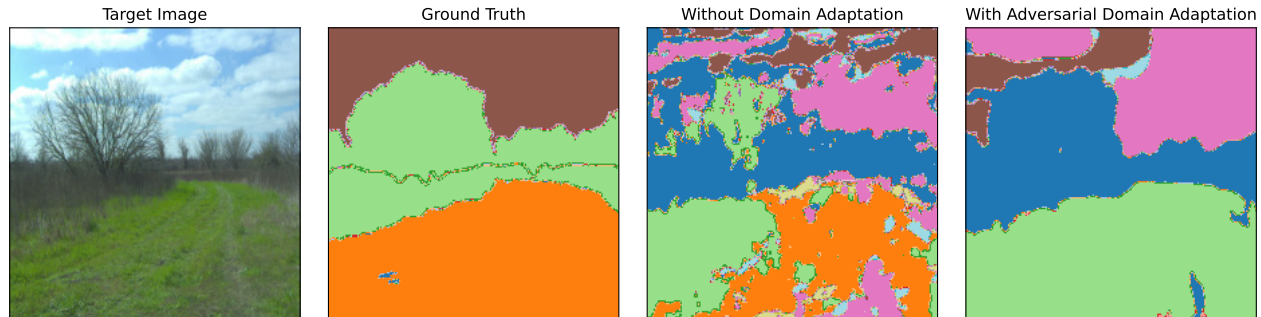


Figure 9. Output on Target dataset from SemanticKITTI to Rellis-3D

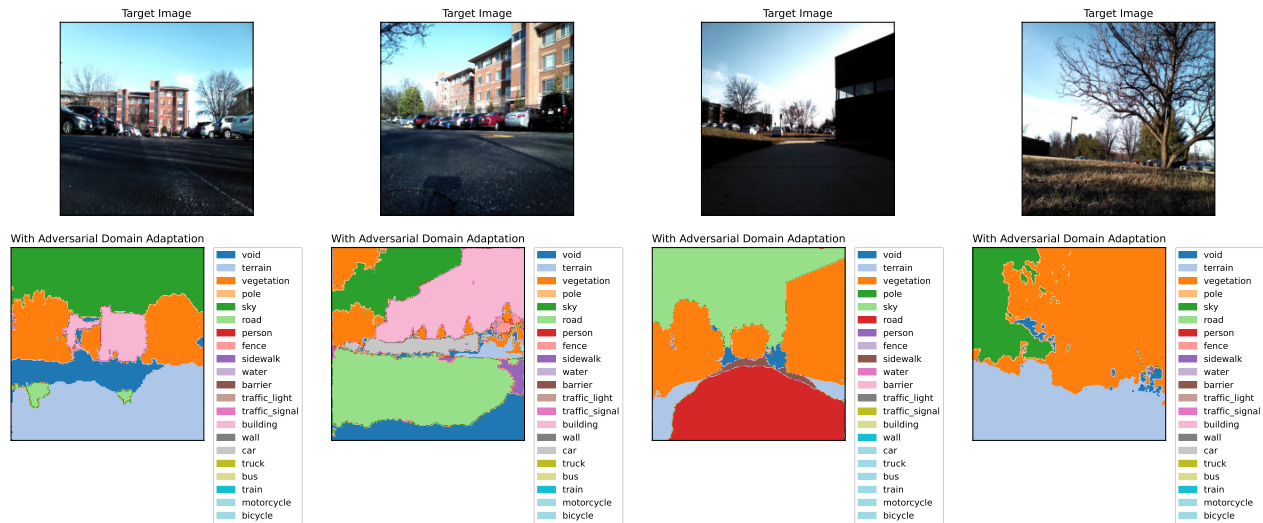


Figure 10. Output on our dataset from CityScape, Rellis-3D & SemanticKITTI

5.5 Discussion

We present the quantitative (i.e., IoU per class and overall pixel accuracy) and qualitative results of our DA technique in the above tables and figures. From the pictorial demonstration, we can infer that the DA approach works as smoothing segmentation compared to approaches without DA. Our DA approach increases pixel-accuracy which demonstrates the improvement in segmentation quality. The IoU of different classes shows that the DA approach struggles to segment the objects which have very small instances present in the dataset. For example, in the Rellis-3D dataset some classes (e.g., road, sidewalk, pole, fence, etc.) have lower number of instances compared to other classes. In those cases, even by training the model on a single domain data, it fails to segment out those skewed classes in test instances of same dataset. Thus it is quite challenging to transfer the semantic segmentation knowledge of those classes to other datasets from Rellis-3D dataset. Further, we empirically conclude that the DA approach performance is inversely related to the JS divergence.

6. CONCLUSION AND FUTURE WORK

In this research work, we investigate the problems of unsupervised domain adaption for semantic segmentation. We experimented both with and without domain adaptation-based semantic segmentation on the urban and rural environment. We evaluate these approaches on three publicly available dataset (i.e., Rellis-3D, SemanticKITTI, Cityscape) and our collected dataset from challenging environments. We empirically observe that the DA approach performs well in similar settings (i.e. urban to urban) and fails to transfer the knowledge efficiently from urban to rural and vice versa. We further observe the problems with class imbalance cases. In future, we will investigate methods to develop model to transfer cross-environment. We will also explore the semantic segmentation approaches which can perform well with low instances in order to resolve class-imbalance problem.

Acknowledgement

This research is supported by U.S. Army grant W911NF2120076.

REFERENCES

- [1] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B., “The cityscapes dataset for semantic urban scene understanding,” in [*Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], (2016).
- [2] Jiang, P., Osteen, P., Wigness, M., and Saripalli, S., “Rellis-3d dataset: Data, benchmarks and analysis,” (2020).
- [3] Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., and Gall, J., “SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences,” in [*Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV)*], (2019).
- [4] Faridee, A. Z. M., Chakma, A., Misra, A., and Roy, N., “Strangan: Adversarially-learned spatial transformer for scalable human activity recognition,” *Smart Health* **23**, 100226 (2022).
- [5] Chakma, A., Faridee, A. Z. M., Khan, M. A. A. H., and Roy, N., “Activity recognition in wearables using adversarial multi-source domain adaptation,” *Smart Health* **19**, 100174 (2021).
- [6] Hong, S., Im, W., Ryu, J., and Yang, H. S., “Ssp-dan: Deep domain adaptation network for face recognition with single sample per person,” in [*2017 IEEE International Conference on Image Processing (ICIP)*], 825–829, IEEE (2017).
- [7] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A., “Image-to-image translation with conditional adversarial networks,” in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 1125–1134 (2017).
- [8] Tzeng, E., Devin, C., Hoffman, J., Finn, C., Abbeel, P., Levine, S., Saenko, K., and Darrell, T., “Adapting deep visuomotor representations with weak pairwise constraints,” in [*Algorithmic Foundations of Robotics XII*], 688–703, Springer (2020).
- [9] Chen, Y.-H., Chen, W.-Y., Chen, Y.-T., Tsai, B.-C., Frank Wang, Y.-C., and Sun, M., “No more discrimination: Cross city adaptation of road scene segmenters,” in [*Proceedings of the IEEE International Conference on Computer Vision*], 1992–2001 (2017).
- [10] Zhang, Y., David, P., and Gong, B., “Curriculum domain adaptation for semantic segmentation of urban scenes,” in [*Proceedings of the IEEE international conference on computer vision*], 2020–2030 (2017).
- [11] Luo, Y., Liu, P., Guan, T., Yu, J., and Yang, Y., “Adversarial style mining for one-shot unsupervised domain adaptation,” *Advances in Neural Information Processing Systems* **33**, 20612–20623 (2020).
- [12] Vu, T.-H., Jain, H., Bucher, M., Cord, M., and Pérez, P., “Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation,” in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 2517–2526 (2019).
- [13] Mei, K., Zhu, C., Jiang, L., Liu, J., and Qiao, Y., “Cross-stained segmentation from renal biopsy images using multi-level adversarial learning,” in [*ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*], 1424–1428, IEEE (2020).
- [14] Pan, F., Shin, I., Rameau, F., Lee, S., and Kweon, I. S., “Unsupervised intra-domain adaptation for semantic segmentation through self-supervision,” in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 3764–3773 (2020).
- [15] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A., “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in [*Proceedings of the IEEE international conference on computer vision*], 2223–2232 (2017).
- [16] Radford, A., Metz, L., and Chintala, S., “Unsupervised representation learning with deep convolutional generative adversarial networks,” *arXiv preprint arXiv:1511.06434* (2015).
- [17] Choi, J., Kim, T., and Kim, C., “Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation,” in [*Proceedings of the IEEE/CVF International Conference on Computer Vision*], 6830–6840 (2019).

- [18] Pizzati, F., Charette, R. d., Zaccaria, M., and Cerri, P., “Domain bridge for unpaired image-to-image translation and unsupervised domain adaptation,” in [*Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*], 2990–2998 (2020).
- [19] Zhang, Y., Qiu, Z., Yao, T., Liu, D., and Mei, T., “Fully convolutional adaptation networks for semantic segmentation,” in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 6810–6818 (2018).
- [20] Ronneberger, O., Fischer, P., and Brox, T., “U-net: Convolutional networks for biomedical image segmentation,” in [*International Conference on Medical image computing and computer-assisted intervention*], 234–241, Springer (2015).
- [21] Geiger, A., Lenz, P., and Urtasun, R., “Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite,” in [*Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*], 3354–3361 (2012).
- [22] Kullback, S. and Leibler, R. A., “On information and sufficiency,” *The annals of mathematical statistics* **22**(1), 79–86 (1951).
- [23] Manning, C. and Schutze, H., [*Foundations of statistical natural language processing*], MIT press (1999).