Semantic Segmentation of Indian Road Scenes

Saket Pradhan[1]

[1] Thakur College of Engineering & Technology, University of Mumbai, India
S1032190351@thakureducation.org

Abstract. Semantic segmentation classifies each pixel of an image into one of several predefined classes. In the context of road scenes, it is used for self-driving cars, traffic monitoring, and map generation. Recent advances have been driven by the development of large, publicly available road scene datasets, such as Cityscapes and SYNTHIA. The current SOTA is dominated by FCNs and encoder-decoder architectures that directly produce a dense, per-pixel prediction. Encoder decoder architectures, such as the U-Net, use a combination of convolutional and transposed convolutional layers to upsample a lower-resolution feature map to output a dense, per-pixel prediction. One area of particular interest is the development of models that handle variations in lighting and weather conditions. We use DAFormer for semantic segmentation, which is trained end- to-end with cross-entropy loss and the SGD optimizer. The use of data augmentation techniques such as random flipping, scaling, and cropping is also performed during training to improve the robustness of the model to different image variations. The performance is evaluated using several standard metrics, with mean IoU scores of over 95.7% on the Cityscapes and 84.5% on the SYNTHIA datasets.

Keywords: Semantic, Segmentation, DAFormer, SegFormer, Transformers, Self-Driving, Cityscapes, GTA5

1 Introduction

Semantic segmentation is a task in computer vision that involves classifying each pixel in an image into one of several predefined categories. This technology can be used in a wide range of applications, including self-driving cars, traffic monitoring and surveillance, and urban planning. In the context of Indian road scenes, semantic segmentation can be used to identify different types of vehicles, pedestrians, buildings, and road markings, which can be used to improve traffic safety and efficiency [1].

However, training semantic segmentation models on Indian road scenes can be challenging due to the wide variability in lighting, weather, and other factors that can affect the appearance of the scene. Additionally, the availability of labeled data for Indian road scenes is often limited, which can make it difficult to train high-performing models. Unsupervised domain adaptation is a method that can be used to improve the performance of semantic segmentation models on Indian road scenes. This approach involves training a model on one set of data, and then fine-tuning it to perform well on a different, but related, set of data. This can be done by using a technique called adversarial training, which involves training two models simultaneously: one that generates images that are similar to the target domain, and another that tries to distinguish between the generated images and the real images.

2 Related Works

2.1 Semantic Segmentation

Semantic segmentation is a computer vision task that involves classifying each pixel in an image into one of several predefined categories. The goal of semantic segmentation is to assign a semantic label, such as "car" or "person," to each pixel in an image. This technique is used to identify and locate various objects and features in an image, such as vehicles, pedestrians, traffic signs, and lane markings.

Semantic segmentation is typically performed using deep learning techniques, such as convolutional neural networks (CNNs) and fully convolutional networks (FCNs). These models are trained on large amounts of labelled data and are able to learn to recognize patterns and features in images that are relevant for semantic segmentation. One of the key advantages of semantic segmentation is its ability to provide a detailed understanding of the scene.

By assigning a label to each pixel, semantic segmentation can be used to identify and locate specific objects and features in an image, such as vehicles, pedestrians, traffic signs, and lane markings. This information can be used for a variety of applications, including self-driving cars, robotics, and surveillance systems.
Semantic segmentation is also used in the field of computer graphics, where it can be used to create high- quality images and animations. The technique can also be used in medical imaging and satellite imaging, where it can be used to identify and locate specific structures or regions in images.

2.2 Unsupervised Domain Adaptation

Unsupervised Domain Adaptation (UDA) is training a statistical model on labelled data to a target domain's data from a source domain more efficiently, having access to only unlabeled data in the target domain. It is a framework for learning that allows information to be transferred from source domains with lots of annotated training examples to target domains with just unlabeled data [2]. It aims to adapt a model trained on synthetic data to real-world data without requiring expensive annotations of real-world images.

In principle, UDA mainly focuses on the global distribution alignment between domains while not including the local distribution properties. The goal is to utilize characteristics from a categorized source area and apply them to an uncategorized target area that has a comparable but distinct data distribution. Typically, deep learning techniques for domain adaptation consist of two stages: Firstly, learning features that maintain a low level of risk on labeled samples (source domain); Secondly, making the features from both domains as similar as possible, to enable a classifier trained on the source domain to also be used on the target domain. However, many methods only function with reduced resolution images, as UDA techniques for semantic segmentation can be quite demanding on GPU memory.

2.3 Indian Road Scenes

Semantic segmentation of Indian road scenes is the process of classifying each pixel in an image of an Indian road scene into one of several predefined categories. This technique is used to identify and locate various objects and features such as vehicles, pedestrians, traffic signs, and lane markings. One of the key challenges in the semantic segmentation of Indian road scenes is the variability in the appearance of objects and features. Indian road scenes often contain a large number of pedestrians and other non- vehicular objects, which can further complicate the task of semantic segmentation.

To address these challenges, researchers have developed several approaches that leverage deep learning techniques such as convolutional neural networks (CNNs) and fully convolutional networks (FCNs) to improve the performance of semantic segmentation models. The application of semantic segmentation of Indian road scenes is mainly used in the development of advanced driver assistance systems (ADAS) for self-driving cars, road safety, traffic management, surveillance, and robotics [3].

ADAS systems use cameras and other sensors to perceive the environment and make decisions about how to control the vehicle. Semantic segmentation can be used to identify and locate objects and features such as vehicles, pedestrians, and traffic signs, which can be used to improve the performance of self-driving cars. Semantic segmentation can also be used to improve road safety by identifying and locating potential hazards such as pedestrians, bicycles, and motorcycles on the road. This information can be used to alert drivers or to control the speed and trajectory of self-driving cars [4].

3 Datasets Comparative Study

3.1 Cityscapes

Cityscapes is a comprehensive database that concentrates on comprehending urban street scenes semantically. It offers detailed annotations for 30 different classes, which are divided into 8 categories (such as flat surfaces, humans, vehicles, constructions, objects, nature, sky, and void), and includes both instance-wise and dense pixel annotations. The dataset encompasses around 5000 images with fine annotations and 20000 with coarse annotations. The data was captured in 50 cities over a period of several months and only during daylight hours and good weather conditions.

The footage was initially recorded as video, and specific frames were selected for inclusion based on the presence of a high number of dynamic objects, diverse scene layouts, and changing backgrounds. The dataset is comprised of a diverse set of stereo video sequences from 50 cities and boasts high-quality pixel-level annotations. This makes it significantly larger in scale compared to similar previous efforts. It is designed for evaluating the performance of vision algorithms in crucial tasks of semantic urban scene understanding, such as pixel-level, instance-level, and panoptic semantic labelling, and to support research that utilizes large amounts of (weakly) annotated data, like for training deep neural networks.

3.2 GTA5

The GTA5 dataset comprises 24966 synthetic images with pixel-level semantic labelling. The images were created through the open-world video game Grand Theft Auto 5 and feature a car's view of American-style virtual cities. There are 19 semantic categories, similar to those in the Cityscapes dataset. The dataset encompasses an extensive range of road scenes, from urban to rural, and provides comprehensive annotations for objects such as buildings, roads, vehicles, and pedestrians. Its diversity and large number of images make it a valuable resource for training models to handle diverse input and generalize effectively to new images.

3.3 SYNTHIA

The SYNTHIA dataset is a synthetic collection of images and annotations designed to support research in semantic segmentation and other scene-understanding tasks related to driving scenarios. The dataset contains an extensive amount of information, with accurate semantic annotations at the pixel level. This includes more than 200,000 high-definition images from video streams, as well as 20,000 images from separate snapshots. The scenes captured in this dataset are diverse, featuring a European-style town, modern city, highway, green areas, and a variety of dynamic objects, such as cars, pedestrians, and cyclists. Additionally, the dataset includes different seasons, lighting conditions, weather patterns, and a simulation of multiple sensors, including 8 RGB cameras and 8 depth sensors.

3.4 National Dataset for Indian Roads

The National Dataset for Indian Roads is a dataset created for research and development of autonomous systems— specifically for Indian road scene analysis, in partnership by IISc and Wipro under the WIRIN initiative. It was created to address the lack of publicly available datasets for Indian road scenes, having unique characteristics of such as crowded roads, diverse traffic, and a wide range of vehicle types. It includes 1000 images captured from various locations in India with a resolution of 1920*1080 pixels. Each picture is annotated with 29 classes, including roads, buildings, vehicles, pedestrians, and traffic signs.

The dataset is composed of images and their annotations which are useful for the semantic segmentation of Indian road scenes as well as traffic analysis. The dataset was created using real-world data captured in Bengaluru urban and suburban, including a wide variety of urban to rural scenes, which makes it unique and valuable for researchers working on autonomous systems in India.

One of its key features is that it includes detailed annotations and labels for different types of roads (such as highways, residential streets, and dirt roads), as well as for different types of vehicles (such as cars, trucks, and buses) and different types of buildings (such as houses and factories). These annotations can be used to train models to accurately identify and segment different objects in the images, which is an important step in developing models for autonomous vehicles and other applications that rely on understanding the environment. This dataset also includes annotations for unique and specific Indian road conditions such as traffic lights, traffic signs, autorickshaws, and guard rails. This allows for the development of models that are specifically tailored to Indian road conditions and regulations.

4 Methodology

In this paper, we implement two models for semantic segmentation, namely SegFormer [5] and DAFormer [11], and discuss their results and performance.

4.1 SegFormer

SegFormer [5] introduces a new transformer encoder structure with a hierarchical design that generates multi-scale features. Unlike other methods, SegFormer does not need positional encoding, which helps prevent the interpolation of positional codes that can lead to reduced performance when the resolution differs between training and testing. Additionally, it has shown to disprove the necessity of complex decoders. The MLP decoder merges local and global attention from different layers, resulting in effective representations. The SegFormer architecture includes two modules: a hierarchical transformer encoder to capture coarse and fine features, and a decoder that combines these multi-level features to predict the segmentation mask. The two key concepts of SegFormer are the use of an encoder that generates multi-scale features, and the MLP-based decoder that aggregates information from different layers to produce a segmentation map.

Encoder. An input image is first divided into 4*4 patches (similar to vision transformers), but in vision transformers, it generally uses a patch size of 16*16. A smaller patch size used to make better dense prediction tasks. The transformer block is composed of three sub-modules:

i. An efficient self-attention,
ii. A mixed feed-forward network, and
iii. An overlapping patch merging module.

The first efficient self-attention module works like the original multi-head self-attention transformer, but it uses a sequence reduction technique in order to lower the computational costs. The next block, which is a mixed feed- forward network, is used to solve the fixed resolution problem. Instead of using fixed sized positional encoding, layers of convolution and multi-layer perceptron (MLP) are used to implement data-driven positional encoding. The last module, which is the overlapping patch merging block, is used to reduce the size of the feature map. The size of the feature is reduced as it goes to the higher part of the network.

Decoder. The decoder is quite simple compared to the modules in the encoder, because there are different features of different sizes in each layer of the encoder. The full MLP layer in the decoder takes the features from the encoder and fuses them together. This all MLP decoder is composed of four main steps:

i. Firstly, multi-level features from the encoder are fed into the multi-layer perceptron layer to unify in the channel dimension.

ii. Next, the features are up-sampled to ¼ of its size and are concatenated together.

iii. Thirdly, a MLP layer is adapted to fuse the concatenated features.

iv. Finally, another MLP layer takes these fused features to predict the segmentation mask.

4.2 DAFormer

Training a neural network for semantic segmentation usually requires expensive pixel-wise annotations of real-world images. Therefore, it is more desirable to exploit other domains that are easier to annotate, such as synthetic data. However, a model trained on the source domain typically experiences a performance drop when applied to the target domain. The goal of Unsupervised Domain Adaptation (UDA) is to increase the performance over the target domain by using unlabeled target images.

Many state-of-the-art UDA methods are based on self- training. The network is trained using ground truth labels for source images and pseudo-labels for target images. The pseudo-labels are generated by taking confident predictions of a teacher. The teacher network is an exponential moving average of the student for temporally stable predictions. In that way, the networks are iteratively adapted to the target domain. Previous UDA methods mostly use a DeepLabV2 network architecture. However, DeepLabV2 is significantly outperformed by more recent networks in supervised semantic segmentation.

DAFormer [11] studies the influence of the network architecture on UDA and compiles a more sophisticated architecture, representing a major advance in UDA. It improves the state-of-the-art performance by a significant margin of 10.8 mIoU on GTA→Cityscapes. A hierarchical transformer is utilized for the encoder, which is revealed to be more domain-robust than the predominant CNNs. For the decoder, a context-aware feature fusion is used, which utilizes domain-robust context clues from different encoder levels. Compared to the DeepLabV2 architecture, DAFormer significantly reduces the performance gap between UDA and the supervised oracle. In many cases, the source dataset is imbalanced as some rare classes appear only in a few images. Consequently, the performance of the infrequent classes is highly influenced by the random selection of data. By frequently sampling images that contain these rare classes, the network can learn them more effectively, leading to more stable results. This, in turn, enhances the quality of pseudo-labels and mitigates confirmation bias.

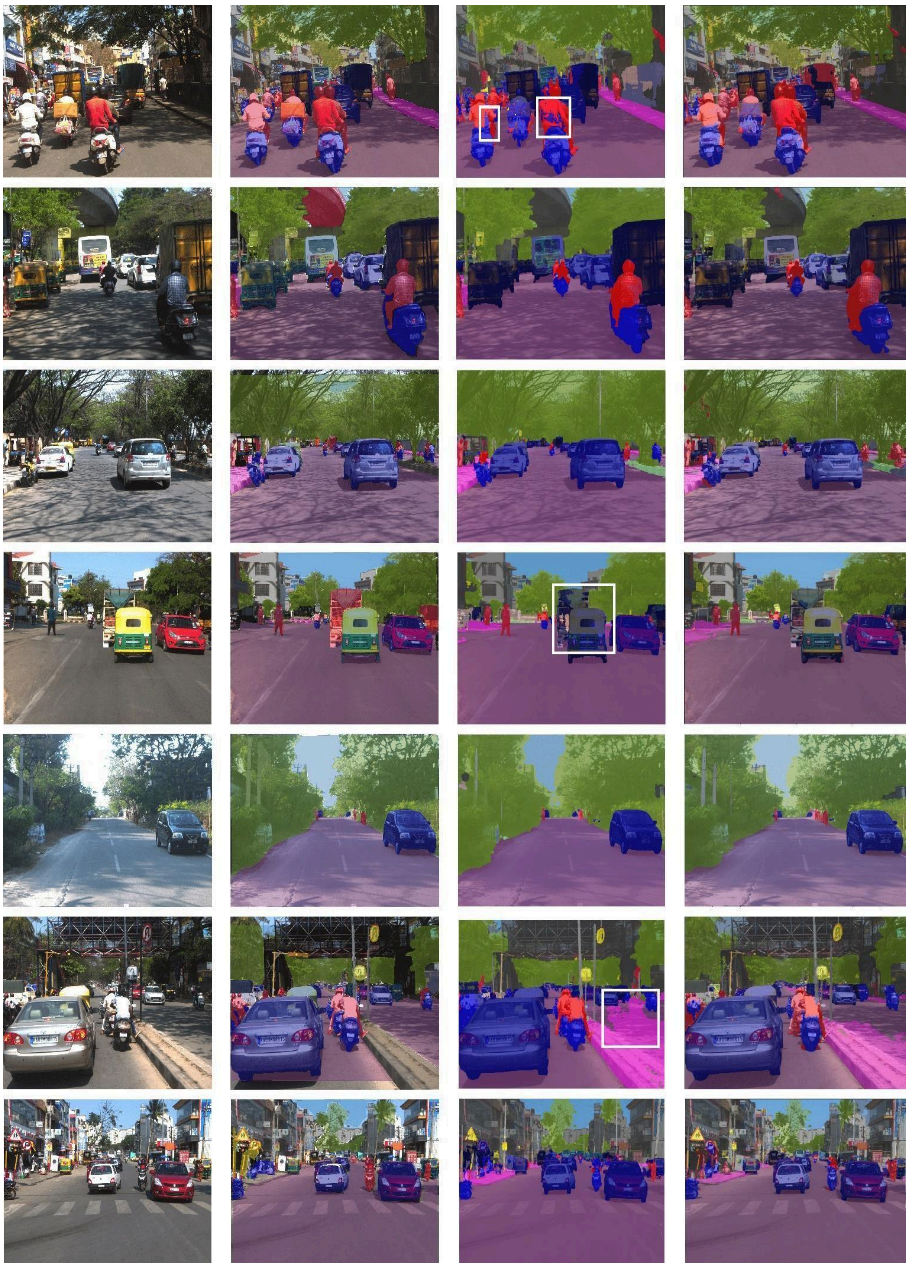

Fig. 1. Qualitative results for semantic segmentation on the National Dataset for Indian Roads. Example predictions showing a better recognition of most classes by DAFormer as opposed to SegFormer. Compared to SegFormer, the DAFormer model predicts the Ground Truth masks with substantially finer details near object boundaries. It also reduces long-range errors as highlighted in white in images 1, 4, and 6.

Table 1. GTA5 TO CITYSCAPES

Classes	Models						
	CBST [8]	DACS [7]	CorDA [8]	BAPA [9]	ProDA [10]	DAFormer [11]	HRDA [12]
Road	91.8	89.9	94.7	94.4	87.8	95.7	96.4
Sidewalk	53.5	39.7	63.1	61	56	70.2	74.4
Building	80.5	87.9	87.6	88	79.7	89.4	91
Wall	32.7	30.7	30.7	26.8	46.3	53.5	61.6
Fence	21	39.5	40.6	39.9	44.8	48.1	51.5
Pole	34	38.5	40.2	38.3	45.6	49.6	57.1
Traffic Light	28.9	46.4	47.8	46.1	53.5	55.8	63.9
Sign	20.4	52.8	51.6	55.3	53.5	59.4	69.3
Vegetation	83.9	88	87.6	87.8	88.6	89.9	91.3
Terrain	34.2	44	47	46.1	45.2	47.9	48.4
Sky	80.9	88.8	89.7	89.4	82.1	92.5	94.2
Person	53.1	67.2	66.7	68.8	70.7	72.2	79
Rider	24	35.8	35.9	40	39.2	44.7	52.9
Car	82.7	84.5	90.2	90.2	88.8	92.3	93.9
Truck	30.3	45.7	48.9	60.4	45.5	74.5	84.1
Bus	35.9	50.2	57.5	59	59.4	78.2	85.7
Train	16	0	0	0	1	65.1	75.9
Motorbike	25.9	27.3	39.8	45.1	48.9	55.9	63.9
Bike	42.8	34	56	54.2	56.4	61.8	67.5
mIoU	91.8	89.9	94.7	94.4	87.8	95.7	96.4

Table 2. SYNTHIA TO CITYSCAPES

Classes	Models						
	CBST [8]	DACS [7]	CorDA [8]	BAPA [9]	ProDA [10]	DAFormer [11]	HRDA [12]
Road	68	80.6	91.7	93.3	87.8	84.5	85.2
Sidewalk	29.9	25.1	53.8	61.6	45.7	40.7	47.7
Building	76.3	81.9	83.9	85.3	84.6	88.4	88.8
Wall	10.8	21.5	22.4	19.6	37.1	41.5	49.5
Fence	1.4	2.9	0.8	5.1	0.6	6.5	4.8
Pole	33.9	37.2	34.9	37.8	44	50	57.2
Traffic Light	22.8	22.7	30.5	36.6	54.6	55	65.7
Sign	29.5	24	42.8	42.8	37	54.6	60.9
Vegetation	77.6	83.7	86.6	84.9	88.1	86	85.3
Terrain	-	-	-	-	-	-	-
Sky	78.3	90.8	88.2	90.4	84.4	89.8	92.9
Person	60.6	67.6	66	69.7	74.2	73.2	79.4
Rider	28.3	38.3	34.1	41.8	24.3	48.2	52.8
Car	81.6	82.9	86.6	85.6	88.2	87.2	89
Truck	-	-	-	-	-	-	-
Bus	23.5	38.9	51.3	38.4	51.1	53.2	64.7
Train	-	-	-	-	-	-	-
Motorbike	18.8	28.5	29.4	32.6	40.5	53.9	63.9
Bike	39.8	47.6	50.5	53.9	45.6	61.7	64.9
mIoU	68	80.6	91.7	93.3	87.8	84.5	85.2

Table 3. NATIONAL DATASET FOR INDIAN ROADS

Classes	Models			
	CBST [8]	BAPA [9]	ProDA [10]	DAFormer [11]
Road	79.3	84.6	82.6	90.9
Sidewalk	45.4	54.2	52	66.9
Building	69.2	79.2	74.2	84.3
Wall	27.9	23.8	43.9	51.3
Fence	17.7	36.1	43.4	45.7
Pole	28.1	34	42.8	47.6
Traffic Light	26.6	42.9	47.5	53.2
Sign	17.4	49.7	49.5	57.1
Vegetation	73.5	77.4	82.3	85.9
Terrain	30	41.2	42	45.3
Sky	67.5	78.6	76.3	87.6
Person	42.1	62.3	65.7	69.1
Rider	20.4	35.6	36.5	42.8
Car	72.9	80.8	82.2	87.8
Truck	27.1	51.9	42.3	70.7
Bus	31.5	52.1	55.4	74.4
Train	13.4	0	0.9	61.8
Motorbike	22.8	41	45.7	53.4
Bike	38.3	48.5	52.4	59
mIoU	78	84.6	81.6	90.7

During UDA, the network overfits the source domain, and difficult classes of the target domain are not distinguished clearly. Therefore, a Thing-Class ImageNet Feature Distance is introduced to regularize the source training. DAFormer shows significant successive improvements of the proposed components over a strong UDA baseline. In particular, it learns even difficult classes that previous methods struggled with.

5 Results and Discussion

The observations made while applying the DAFormer model are as follows: The DAFormer struggles to deal with shadows, particularly if they cover the part of the road right in front of the vehicle. It is not effective at dividing the road into proper segments if the sidewalks or walkways are not well- defined. The model is vulnerable to creating inaccurate segmentation maps in situations where there is too much light exposure, particularly on roads that are overexposed to sunlight. The DAFormer also fails to recognize or segment road signs incorrectly. The observations made while using the SegFormer model are: It frequently makes incorrect identifications and classifications for objects that are far away in the frame, particularly for vehicles and humans. The model struggles to recognize the road surface or sky if the image is not clear enough, as seen in an example where the ground is partially segmented. Shadows or bright patches of light present in the image can cause the model to leave the affected area unclassified. In some cases, a single vehicle may be divided into multiple classes due to sharp boundaries or margins in the image. The results of the benchmark of previous models on various classes and shown in Table 1, Table 2, and Table 3. Fig. 1 shows the inference results for some images from the National Dataset for Indian Roads.

6 Conclusion

The task of semantic segmentation of Indian road scenes is difficult due to the complexity and variety of these scenes, as well as the lack of sufficient labeled data. Unsupervised domain adaptation has proven to be a promising solution by fine-tuning pre-trained models on the target domain using limited labeled data. This approach has shown positive results for different types of Indian road scenes, with the ability to accurately segment road surfaces and objects like vehicles, pedestrians, and buildings. The unsupervised domain adaptation also increases the generalization of the models to new scenes, which is crucial for practical applications. However, there are still some challenges to be overcome, such as the models' incapacity to fully capture variations like weather conditions, etc.

Furthermore, further research is necessary to assess the robustness and scalability of unsupervised domain adaptation for semantic segmentation of Indian road scenes. In conclusion, unsupervised domain adaptation is a hopeful method for semantic segmentation of Indian road scenes. It effectively uses pre-trained models and limited labeled data to enhance the performance and generalization of the models. Despite the existing limitations, this field continues to evolve with advancements in computer vision, deep learning, and the availability of large labeled datasets, which could result in even better and more reliable models for applications like self-driving cars, traffic monitoring, and smart city planning.

References

1. Dewangan, D. K., & Sahu, S. P. (2021). Road detection using semantic segmentation-based convolutional neural network for intelligent vehicle system. In *Data engineering and communication technology* (pp. 629-637). Springer, Singapore.
2. Baheti, B., Gajre, S., & Talbar, S. (2019, October). Semantic scene understanding in unstructured environment with deep convolutional neural network. In *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)* (pp. 790-795). IEEE.
3. Hong, Y., Pan, H., Sun, W., & Jia, Y. (2021). Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes. *arXiv preprint arXiv:2101.06085*.
4. Chen, Y., Li, W., & Van Gool, L. (2018). Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7892-7901).
5. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, *34*, 12077-12090.
6. Zou, Y., Yu, Z., Kumar, B., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: ECCV. pp. 289–305 (2018).

7. Tranheden, W., Olsson, V., Pinto, J., Svensson, L.: DACS: Domain Adaptation via Cross-domain Mixed Sampling. In: WACV. pp. 1379– 1389 (2021).
8. Wang, Q., Dai, D., Hoyer, L., Fink, O., Van Gool, L.: Domain adaptive semantic segmentation with self-supervised depth estimation. In: ICCV. pp. 8515–8525 (2021).
9. Liu, Y., Deng, J., Gao, X., Li, W., Duan, L.: Bapa-net: Boundary adaptation and prototype alignment for cross-domain semantic segmentation. In: ICCV. pp.8801–8811 (2021).
10. Zhang, P., Zhang, B., Zhang, T., Chen, D., Wang, Y., Wen, F.: Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In: CVPR. pp. 12414–12424 (2021).
11. Hoyer, L., Dai, D., Van Gool, L.: DAFormer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In: CVPR (2022).
12. Hoyer, L., Dai, D., & Van Gool, L. (2022). HRDA: Context-Aware High-Resolution Domain-Adaptive Semantic Segmentation. arXiv preprint arXiv:2204.13132.