This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Unsupervised Domain Adaptation for Semantic Segmentation of Urban Scenes

Matteo Biasetton Umberto Michieli Gianluca Agresti Pietro Zanuttigh
University of Padova, Italy
matteo.biasetton@gmail.com {umberto.michieli,gianluca.agresti,zanuttigh}@dei.unipd.it

Abstract

The semantic understanding of urban scenes is one of the key components for an autonomous driving system. Complex deep neural networks for this task require to be trained with a huge amount of labeled data, which is difficult and expensive to acquire. A recently proposed workaround is the usage of synthetic data, however the differences between real world and synthetic scenes limit the performance. We propose an unsupervised domain adaptation strategy to adapt a synthetic supervised training to real world data. The proposed learning strategy exploits three components: a standard supervised learning on synthetic data, an adversarial learning strategy able to exploit both labeled synthetic data and unlabeled real data and finally a self-teaching strategy working on unlabeled data only. The last component is guided by the segmentation confidence, estimated by the fully convolutional discriminator of the adversarial learning module, helping to further reduce the domain shift between synthetic and real data. Furthermore we weighted this loss on the basis of the class frequencies to enhance the performance on less common classes. Experimental results prove the effectiveness of the proposed strategy in adapting a segmentation network trained on synthetic datasets, like GTA5 and SYNTHIA, to a real dataset as Cityscapes.

1. Introduction

One of the key requirements for autonomous driving applications is an efficient semantic scene understanding algorithm able to recognize all the various objects and regions in the environment surrounding the car. This task is typically solved using deep learning techniques for semantic segmentation. Recent advances in deep neural networks have allowed to obtain an accurate semantic understanding of road scenes, however they typically require a huge amount of labeled data with pixel-level information for training and the generation of these annotations requires a huge effort. A recently proposed workaround for this issue is to use computer generated data for training the networks. In particular, very realistic rendering models have been realized by the video game industry. Modified version of the games software can be used to produce a large amount of high quality rendered road scenarios [28, 27]. However, despite the impressive quality of the video games graphics, there is still a domain shift between the synthetic video game data and the real world images acquired by video cameras placed on cars. This issue needs to be addressed to build a system able to obtain good and reliable performance in the real world scenario.

This paper proposes an unsupervised domain adaptation strategy based on adversarial learning to adapt an initial learning performed on synthetic data to real world scenes. This could potentially help shaping how autonomous vehicles face road scenarios [23]. We envisage a scenario where a large amount of annotated synthetic data is available but no labeled real world samples are available. The proposed method exploits an adversarial learning framework, where a segmentation network based on the DeepLab v2 framework [2] is trained using both labeled and unlabeled data thanks to the combination of three different losses. The first is a standard supervised cross-entropy loss exploiting ground truth annotations allowing to perform an initial supervised training phase on synthetic data. The second is an adversarial loss derived from previous methods [15, 19] developed in the context of semi-supervised semantic segmentation (i.e., for dealing with datasets only partially annotated). In this framework, we exploited a fully convolutional discriminator which takes in input the semantic segmentation from the generator network and the ground truth segmentation maps and produces a pixel-level confidence map distinguishing between the two types of data. It allows to train in an adversarial setting the segmentation network using both synthetic labeled data and real world scenes without ground truth information. Finally, the third term is based on a self-teaching framework inspired from [15], where the predicted segmentation is passed through the discriminator to obtain a confidence map and then high confidence regions are considered reliable and used as ground truth for self-teaching the network over them.

We trained the network on both synthetic labeled data (using the first and second component of the loss) and on unlabeled real world data (using the second and third component) thus being able to obtain accurate results on real world datasets even without using labeled real world data. Since the various classes have different frequencies, we improved the performance by weighting the loss coming from unlabeled data in proportion to the frequency of the classes in the dataset. This allowed to better balance the results between the different classes and to avoid a dramatic drop in performance on less common classes corresponding to small objects and structures that typically represent the critical elements in the autonomous driving scenario. The approach has been trained using the synthetic datasets SYNTHIA and GTA5 for the supervised part and the real dataset Cityscapes for the unsupervised components and then tested on the Cityscapes validation set, proving to achieve state-of-the-art results on the unsupervised domain adaptation task.

2. Related Work

Semantic segmentation of images, i.e., pixel-level labeling, is a very wide research field and a huge number of approaches have been proposed for this task. A very recent review can be found in [9]: current state-of-the-art approaches are mostly based on the Fully Convolutional Network (FCN) model [20], notable examples are DilatedNet [40], PSPNet [43] and DeepLab [2] which is the model employed for the generator network in this work. However, since this paper deals with adversarial learning techniques for semi-supervised training and on the problem of unsupervised domain adaptation from synthetic to real-world data, we will focus on techniques for these tasks in this section.

Semi-supervised learning. Semantic segmentation architectures are typically trained on huge datasets with pixel-wise annotations (e.g., the Cityscapes [5] or CamVid [1] datasets), which are highly expensive, time-consuming and error-prone to generate. To overcome this issue, semi-supervised methods are emerging, trying to exploit weakly annotated data (e.g., with only image labels or only bounding boxes) [25, 31, 37, 39, 13, 6, 14, 32] or completely unlabeled [24, 29, 15, 31, 19] data after a first stage of supervised training. In particular the works of [22, 31] have paved the way respectively to adversarial learning approaches for the semantic segmentation task and to their application to semi-supervised learning. The recent approaches of [15, 19] propose semi-supervised frameworks exploiting adversarial learning with a Fully Convolutional Discriminator (FCD) trying to distinguish the predicted probability maps from the ground truth segmentation distributions at pixel-level. These works targeted a scenario where the dataset is only partially labeled: in their settings, unlabeled data comes from the same dataset and shares the same domain data distribution of labeled data. We instead propose to tackle a scenario where unlabeled data refers to a different dataset with a inherently different domain distribution.

Domain Adaptation. In addition to the aforementioned approaches to overcome the lack of data, an increasingly popular alternative is represented by domain adaptation from synthetic data. The development of sophisticated computer graphics techniques enabled the production of huge synthetic datasets for semantic segmentation purposes at a very low cost. To this end, several synthetic datasets have been built, e.g., GTA5 [27] or SYNTHIA [28] which have been employed in this work. The real challenge is then to address the cross-domain shift when a neural network trained on synthetic data needs to process real-world images since in this case training and test data are not drawn i.i.d. from the same underlying distribution as usually assumed [41, 33, 34, 10, 17]. A possible solution is to process synthetic images to reduce the inherent discrepancy between source and target domain distributions mainly using Generative Adversarial Networks (GANs) [30, 26, 38, 44, 16]

Unsupervised domain adaptation has been already widely investigated in classification tasks [7, 8, 21, 36]. On the other hand, its application to semantic segmentation is still a quite new research field. The first work to investigate cross-domain urban scene semantic segmentation is [12], where adversarial training is employed to align the features from the different domains. In [41], a curriculum-style learning approach is proposed where firstly the easier task of estimating global label distributions is learned and then the segmentation network is trained forcing that the target label distribution is aligned to the previously computed properties. Following these approaches, many works addressed the source to target domain shift problem with various techniques, such as cycle consistency [11], GANs [29], output space alignment [35], distillation loss [4], class-balanced self-training [46], conservative loss [45], geometrical guidance [3] and adaptation networks [42].

3. Architecture of the Proposed Approach

The proposed approach is based on two main modules, i.e., two different Convolutional Neural Networks (CNNs). The first network (i.e., the generator G in the adversarial learning framework) performs the semantic segmentation of the given color image. For this module, we exploited the Deeplab v2 network [2] based on the ResNet-101 model whose weights were pre-trained[1] on the MSCOCO dataset [18]. Although we considered the Deeplab v2, notice that our approach does not rely on specific properties of this network and any network for semantic segmentation can be fit inside the proposed learning framework. Figure 1 shows a

[1]We used the weights computed by V. Nekrasov available at: https://github.com/DrSleep/tensorflow-deeplab-resnet

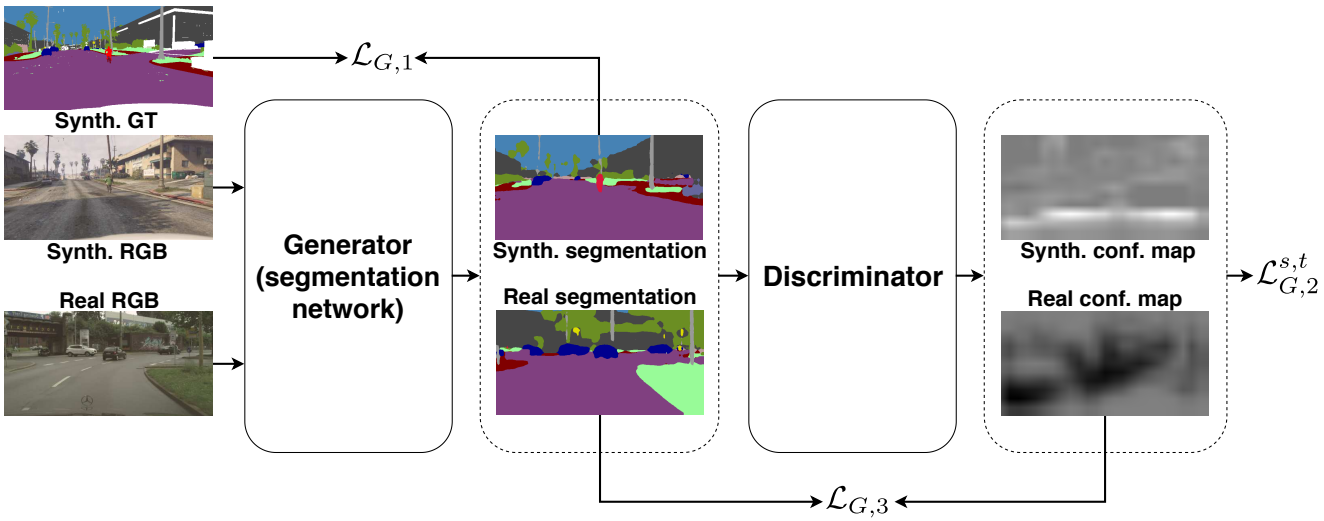

Figure 1: Architecture of the proposed framework for the training of the generator network. A first stage of supervised learning with synthetic data is followed by a second stage using also unlabeled real data to boost the performance of the segmentation network (i.e., the generator) through the combination of 3 losses. $\mathcal{L}_{G,1}$ is a standard cross-entropy loss computed on synthetic data, $\mathcal{L}^{s,t}_{G,2}$ is an adversarial loss referring to a fully-convolutional discriminator network, and $\mathcal{L}_{G,3}$ is a self-teaching loss for unlabeled real data.

general overview of the procedure used to train G exploiting 3 different losses.

Starting from the first, the network produces a class probability map representing for each pixel the probability that it belongs to each class c inside the set of possible classes $\mathcal{C}$. This map can be directly used to train the network in a supervised way exploiting the semantic ground truth data: we used a standard cross-entropy loss ($\mathcal{L}_{G,1}$) for this task. More in detail, given the n-th input image $\mathbf{X}^s_n$ from the source (synthetic) domain, its one-hot encoded ground truth segmentation $\mathbf{Y}^s_n$ and the output of the segmentation network $G(\mathbf{X}^s_n)$, the loss $\mathcal{L}_{G,1}$ on the image $\mathbf{X}^s_n$ can be computed as:

$$\mathcal{L}_{G,1} = -\sum_{p \in \mathbf{X}^s_n} \sum_{c \in \mathcal{C}} \mathbf{Y}^{s\,(p)}_n[c] \cdot \log\left(G(\mathbf{X}^s_n)^{(p)}[c]\right) \quad (1)$$

where p is a generic pixel in the considered image $\mathbf{X}^s_n$, c is a particular class contained in the set $\mathcal{C}$ of possible classes and $\mathbf{Y}^{s\,(p)}_n[c]$ and $G(\mathbf{X}^s_n)^{(p)}[c]$ are respectively the value in the one-hot encoded ground truth and in the generator network estimate related to the pixel p and the class c.

Notice that this loss can be computed only on the source domain (i.e., on synthetic data) where the pixel-level semantic ground truth is available. However, our main target is to adapt the supervised synthetic training to the real world target domain in a unsupervised way. We exploited an adversarial learning framework: a second CNN is introduced, i.e., a discriminator network (D) that aims at distinguishing segmentation maps produced by the generator from the ground truth ones. Differently from other adversarial learning models, this network produces a per-pixel prediction instead of a single binary value for the whole input image. The discriminator D is made of a stack of 5 convolutional layers each with 4×4 kernels with a stride of 2 and Leaky ReLU activation function. The number of filters (from the first layer to the last one) is 64, 64, 128, 128, 1 and the cascade is followed by a bilinear upsampling to match the original input image resolution. The loss of the discriminator $\mathcal{L}_D$ is a standard cross-entropy loss between the produced map and the one-hot encoding related to the *fake* domain (class 0) or ground truth domain (class 1) depending on the fact that the input has been respectively drawn from the generator or from ground truth data. Mathematically, $\mathcal{L}_D$ is defined as:

$$\mathcal{L}_D = -\sum_{p \in \mathbf{X}^{s,t}_n} \log(1 - D(G(\mathbf{X}^{s,t}_n))^{(p)}) + \log(D(\mathbf{Y}^s_n)^{(p)}) \quad (2)$$

Notice that the discriminator has to label with 0 the segmentation maps produced by the generator using both synthetic data from the source domain s (denoted with $\mathbf{X}^s_n$) or real world data from the target domain t (i.e., $\mathbf{X}^t_n$). Thus, it allows to exploit also the real world data in an unsupervised way, and it tries to distinguish the segmentations produced by the generator G from ground truth segmentation data (that can be only synthetic in our framework). The usage of both types of data is made possible by the similar classes' statistics of source and target datasets. Notice also that, in principle, the task of the discriminator appears to be

trivially solvable by distinguishing a Dirac distributed input (i.e., the one-hot encoded annotations) from other prediction distributions. However, we have empirically observed that the generator network produces (and is forced to produce even more by the adversarial training process) segmentation maps which are very close to a Dirac distribution. The second loss term for the training of G is $\mathcal{L}_{G,2}^{s,t}$, that is computed on the generic image $\mathbf{X}_n^{s,t}$ from the discriminator output as:

$$\mathcal{L}_{G,2}^{s,t} = -\sum_{p \in \mathbf{X}_n^{s,t}} \log(D(G(\mathbf{X}_n^{s,t}))^{(p)}) \tag{3}$$

This term forces the training of the generator network in the direction of fooling the discriminator producing data that resembles the ground truth statistics. Notice that in this computation the image can be taken from both the source or the target dataset (i.e., it can be both a synthetic or a real world image): in the following of this paper, we are going to use $\mathcal{L}_{G,2}^{s}$ to refer to the loss function computed only on data extracted from the source dataset, while $\mathcal{L}_{G,2}^{t}$ refers to the loss computed on data from the target dataset. In particular, in the second case, $\mathcal{L}_{G,2}^{t}$ tries to force the generator to adapt to the target domain and to improve the performance by encouraging cleaner segmentations and global consistency with respect to the segment shapes.

Finally, starting from the idea in [15] we exploited the output of the discriminator D as a confidence measure representing the reliability of the estimations performed by G. This allows to perform a sort of self-training following the idea that the predictions of G are more reliable where D marks them as ground truth with an higher accuracy. This is represented by the third loss component of the generator, defined as:

$$\mathcal{L}_{G,3} = -\sum_{p \in \mathbf{X}_n^t} \sum_{c \in \mathcal{C}} I_{T_u}^{(p)} \cdot W_c^t \cdot \hat{\mathbf{Y}}_n^{(p)}[c] \cdot \log\left(G(\mathbf{X}_n^t)^{(p)}[c]\right) \tag{4}$$

where $\hat{\mathbf{Y}}_n$ is the one-hot encoded ground truth derived from the per-class argmax of the generated probability map $G(\mathbf{X}_n)$. W_c^t, instead, is the weighting function on the source domain defined as:

$$W_c^t = 1 - \frac{\sum_n |p \in \mathbf{X}_n^s \wedge p \in c|}{\sum_n |p \in \mathbf{X}_n^s|}, \tag{5}$$

where $|\cdot|$ represents the cardinality of the considered set.

This set of weights serves as a balancing factor when unlabeled data of the target set are used. Without this weighting factor, unlabeled data would lead the model to mislead rare and tiny objects (such as *traffic lights* or *pole*) as frequent and large ones (such as *road*, *building*). Notice that the term comes into play when using unlabeled data of the target domain but the class frequencies have to be computed on the labeled data of the source domain since we need the ground truth labels to evaluate it. This calculation has only to be performed *a priori* and it is not changed as the learning progresses.

Finally, $I_{T_u}^{(p)}$ is an indicator function defined as:

$$I_{T_u}^{(p)} = \begin{cases} 1, & \text{if } D(G(\mathbf{X}_n^t))^{(p)} > T_u \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

with T_u being a threshold for the pixel-wise confidence maps generated by the discriminator in response to the data produced by the generator. We empirically set $T_u = 0.2$ being a reasonable value. This term is intended to enhance the learning process in a self-taught manner using unlabeled data of the target domain.

To conclude, a weighted average of the three losses is used to train the generator exploiting the proposed adversarial learning framework, i.e.:

$$\mathcal{L}_{full} = \mathcal{L}_{G,1} + w^{s,t}\mathcal{L}_{G,2}^{s,t} + w'\mathcal{L}_{G,3} \tag{7}$$

We set the weighting parameters empirically to balance between the three components as $w^s = 0.01$, $w^t = 0.001$ to give less weight in case of unlabeled data and $w' = 0.1$.

The discriminator is fed both with ground truth labels and with the generator output computed on a mixed batch containing both labeled and unlabeled data and is trained aiming at minimizing $\mathcal{L}_D$. Concerning the generator, instead, during the first 5000 steps $\mathcal{L}_{G,3}$ is disabled (i.e., w' is set to 0) thus allowing the discriminator to enhance its capabilities to produce higher quality confidence maps before using them. After this, the training process continues up to 20000 steps with all the three components of the loss enabled.

4. Datasets

In this section, we introduce the datasets used to evaluate the performance of the proposed unsupervised domain adaptation framework. Our target is to show how it is possible to train a semantic segmentation network in a supervised way on synthetic datasets and then apply unsupervised domain adaptation to real data in autonomous driving scenarios. Thus, we used two publicly available synthetic datasets, namely GTA5 [27] and SYNTHIA [28] for the supervised part of the training, while the unsupervised adaptation and the result evaluation have been performed on the real world Cityscapes [5] dataset. In general we followed the same evaluation scenarios of the competing approaches for fair comparison [12, 29, 41].

GTA5 [27] is a huge dataset composed by 24966 photo-realistic synthetic images with pixel level semantic annotation. The images have been recorded from the prospective of a car in the streets of virtual cities (resembling the ones

in California) in the open-world video game *Grand Theft Auto 5*. Being taken from a high budget commercial production they have an impressive visual quality and are very realistic. In our experiments, we used 23966 images for the supervised training and 1000 images for validation purposes. There are 19 semantic classes which are compatible with the ones of the Cityscapes dataset. The original resolution of the images is 1914×1052 px but we rescaled and cropped them to the size of 375×750 px for memory constraints before being fed to the architecture.

SYNTHIA [28] is a very large dataset of photo-realistic images. It has been produced with an ad-hoc rendering engine, allowing to obtain a large variability of the images. On the other hand, the visual quality is not the same of the commercial video game GTA5. We used the *SYNTHIA-RAND-CITYSCAPES* version of the dataset, which contains 9400 images with annotations compatible with 16 of the 19 classes of Cityscapes. These images have been captured on the streets of a virtual European-style town in different environments under various light and weather conditions. As done in previous approaches, we randomly extracted 100 images for validation purposes from the original training set, while the remaining part, composed by 9300 images, is used for the supervised training of our networks. Again, the images have been rescaled and cropped from the original size of 760×1280 px to 375×750 px. For the evaluation of the proposed unsupervised domain adaptation on the Cityscapes dataset, only the 16 classes contained in both datasets are taken into consideration.

Cityscapes [5] is the target dataset for our domain adaptation framework. It is composed by 2975 high resolution color images captured on the streets of 50 different European cities. They have pixel level semantic annotation with 34 classes overall. Since the labels of the original test set are not available, we exploited the original training set (without the labels) for unsupervised training and used the 500 images in the original validation set as a test set, as done also by other recent approaches.

More in detail, the semantic labels have been used just for testing purposes, while the labels of training data have not been used since we aim at proposing a fully unsupervised adaptation strategy. As for the other datasets, the original high resolution images have been resized to 375×750 px for memory constraints. The testing was instead carried out on the original resolution of 2048×1024 px.

5. Experimental Results

The target of the proposed approach is to adapt a deep network trained on synthetic data to real world scenes. To evaluate the performance on this task we performed two different sets of experiments. In the first experiment we trained the network using the scenes from the GTA5 dataset to compute the supervised loss $\mathcal{L}_{G,1}$ and the adversarial loss $\mathcal{L}^s_{G,2}$. Then we used the training scenes of the Cityscapes dataset for the unsupervised domain adaptation: no labels from Cityscapes have been used and when dealing with this dataset we only computed the losses $\mathcal{L}^t_{G,2}$ and $\mathcal{L}_{G,3}$. Finally we evaluated the performance on the validation set of Cityscapes. In the second experiment we performed the same procedure but we replaced the GTA5 dataset with the SYNTHIA one.

The proposed architecture has been implemented using TensorFlow and more material is available at http://lttm.dei.unipd.it/paper_data/semanticDA. The generator network G (that is a Deeplab v2 network) has been trained as proposed in [2] using the Stochastic Gradient Descent (SGD) optimizer with momentum set to 0.9 and weight decay to 10^{-4}. The discriminator D has been trained using the Adam optimizer. The learning rate employed for both G and D started from 10^{-4} and was decreased up to 10^{-6} by means of a polynomial decay with power 0.9. We trained the two networks for 20000 iterations on a NVIDIA Titan X GPU. The longest training inside this work, i.e., the one with all the losses enabled, took about 10 hours to complete.

To measure the performance, we compared the predictions on the Cityscapes validation set with the ground truth labels and computed the mean Intersection over Union (mIoU) as done by most competing approaches [12, 4, 35].

Table 1 refers to the first experiment (i.e., using GTA5 for the supervised training). It shows the accuracy of the proposed approach when exploiting different domain adaptation strategies and compares it with some state-of-the-art approaches. By simply training the network in a supervised way on the GTA5 dataset and then performing inference on real world data from the Cityscapes dataset we obtained a mIoU of 27.9%. The adversarial learning framework on synthetic data (i.e., the contribution of $\mathcal{L}^s_{G,2}$) allows to improve the mIoU to 29.3%. By looking more in detail to the various class accuracies it is possible to see that the accuracy has increased on some of the most common classes corresponding to large structures, while the behaviour on low frequency classes corresponding to small objects is more unstable (some improve but others have a lower accuracy). For this reason in the third loss component related to the self-teaching, the class weights have been taken into account. Thanks to this when using the full framework with all the losses the mIoU increases to 30.4% and in particular it is possible to appreciate a large performance boost on many uncommon classes corresponding to small objects and structures.

By comparing with state-of-the-art approaches, it is possible to see how the method of Hung et al. [15], based on a similar framework, achieves an accuracy of 29%, lower than our approach mostly because it struggles with small structures and uncommon classes. The method of [12] has even

	road	sidewalk	building	wall	fence	pole	t light	t sign	veg	terrain	sky	person	rider	car	truck	bus	train	mbike	bike	mean
Ours ($\mathcal{L}_{G,1}$ only)	45.3	20.6	50.1	9.3	12.7	19.5	4.3	0.7	81.9	21.1	63.3	52.0	1.7	77.9	26.0	39.8	0.1	4.7	0.0	27.9
Ours ($\mathcal{L}_{G,1}$, $\mathcal{L}^s_{G,2}$ only)	61.0	18.5	51.6	15.4	12.3	20.5	1.4	0.0	82.6	24.7	61.0	52.1	2.2	78.5	25.9	41.5	0.4	8.0	0.1	29.3
Ours ($\mathcal{L}_{full}$)	54.9	23.8	50.9	16.2	11.2	20.0	3.2	0.0	79.7	31.6	64.9	52.5	7.9	79.5	27.2	41.8	0.5	10.7	1.3	30.4
Hoffman et al. [12]	70.4	32.4	62.1	14.9	5.4	10.9	14.2	2.7	79.2	21.3	64.6	44.1	4.2	70.4	8.0	7.3	0.0	3.5	0.0	27.1
Hung et al. [15]	81.7	0.3	68.4	4.5	2.7	8.5	0.6	0.0	82.7	21.5	67.9	40.0	3.3	80.7	34.2	45.9	0.2	8.7	0.0	29.0

Table 1: Mean intersection over union (mIoU) on the different classes of the original Cityscapes validation set. The approaches have been trained in a supervised way on the GTA5 dataset and then the unsupervised domain adaptation has been performed using the Cityscapes training set.

	road	sidewalk	building	wall	fence	pole	t light	t sign	veg	sky	person	rider	car	bus	mbike	bike	mean
Ours ($\mathcal{L}_{G,1}$ only)	10.3	20.5	35.5	1.5	0.0	28.9	0.0	1.2	83.1	74.8	53.5	7.5	65.8	18.1	4.7	1.0	25.4
Ours ($\mathcal{L}_{G,1}$, $\mathcal{L}^s_{G,2}$ only)	9.3	19.3	33.5	0.9	0.0	32.5	0.0	0.5	82.3	76.9	54.7	5.5	64.9	17.0	5.7	3.9	25.4
Ours ($\mathcal{L}_{full}$)	78.4	0.1	73.2	0.0	0.0	16.9	0.0	0.2	84.3	78.8	46.0	0.3	74.9	30.8	0.0	0.1	30.2
Hoffman et al. [12]	11.5	19.6	30.8	4.4	0.0	20.3	0.1	11.7	42.3	68.7	51.2	3.8	54.0	3.2	0.2	0.6	20.1
Hung et al. [15]	72.5	0.0	63.8	0.0	0.0	16.3	0.0	0.5	84.7	76.9	45.3	1.5	77.6	31.3	0.0	0.1	29.4

Table 2: Mean intersection over union (mIoU) on the different classes of the original Cityscapes validation set. The approaches have been trained in a supervised way on the SYNTHIA dataset and then the unsupervised domain adaptation has been performed using the Cityscapes training set.

lower performance, however it is also based on a different generator network with lower accuracy (i.e, the method of [40]).

Figure 2 shows the output of the different versions of our approach and of the method of [15] on some sample scenes. The supervised training leads to reasonable results but some small objects get lost or have a wrong shape (e.g., the riders in row 1). Furthermore, some regions of the street and of structures like the walls are corrupted by noise (see the street in the last two rows or the fence on the right in row 3). The adversarial loss $\mathcal{L}^s_{G,2}$ reduces these artifacts but there are still issues on the small objects (e.g., the rider in the fifth row) and the boundaries are not always very accurate (see the fence in the third row). The complete model leads to better performance, for example in the images of Figure 2 the people are better preserved and the structures have better defined edges. Finally the approach of [15] seems to lose some structures (e.g., the fence in the third row) and has issues with the small objects (the riders in row 5 get completely lost) as pointed out before.

By using the SYNTHIA dataset as source dataset, the domain adaptation task is even more challenging if compared with the GTA5 case since the computer generated graphics are less realistic. Table 2 shows that by training the network G in a supervised way on the SYNTHIA dataset and then performing inference on the real world Cityscapes dataset, a mIoU of 25.4% can be obtained. This value is smaller than the mIoU of 27.9% obtained by training G on the GTA5 dataset. This result confirms that the GTA5 dataset has a smaller domain shift with respect to real world data, when compared with the SYNTHIA dataset (GTA5 data, indeed, have been produced by a more advanced rendering engine with more realistic graphics). Under this training scenario, the proposed adversarial loss $\mathcal{L}^s_{G,2}$ does not bring to noteworthy improvements in the domain adaptation task, indeed the mIoU is equal to the *baseline*. On the other hand, by adding the self-taught loss $\mathcal{L}_{G,3}$, a noticeable improvement to a mIoU of 30.2% can be obtained.

Our domain adaptation framework is able to outperform the compared state-of-the-art approaches. The method of Hung et al. [15], that exploits the same generator architecture of our approach, obtains a mIoU equal to 29.4%, lower than our method. The method of [12] appears to be again the less performing approach. In this comparison, it is even less accurate than our *baseline*, but it employs a different segmentation network.

Figure 3 shows the output of the different versions of our approach and of the method of [15] on some sample scenes. The first thing that can be noticed by looking at the qualitative results of the *baseline* supervised version is that

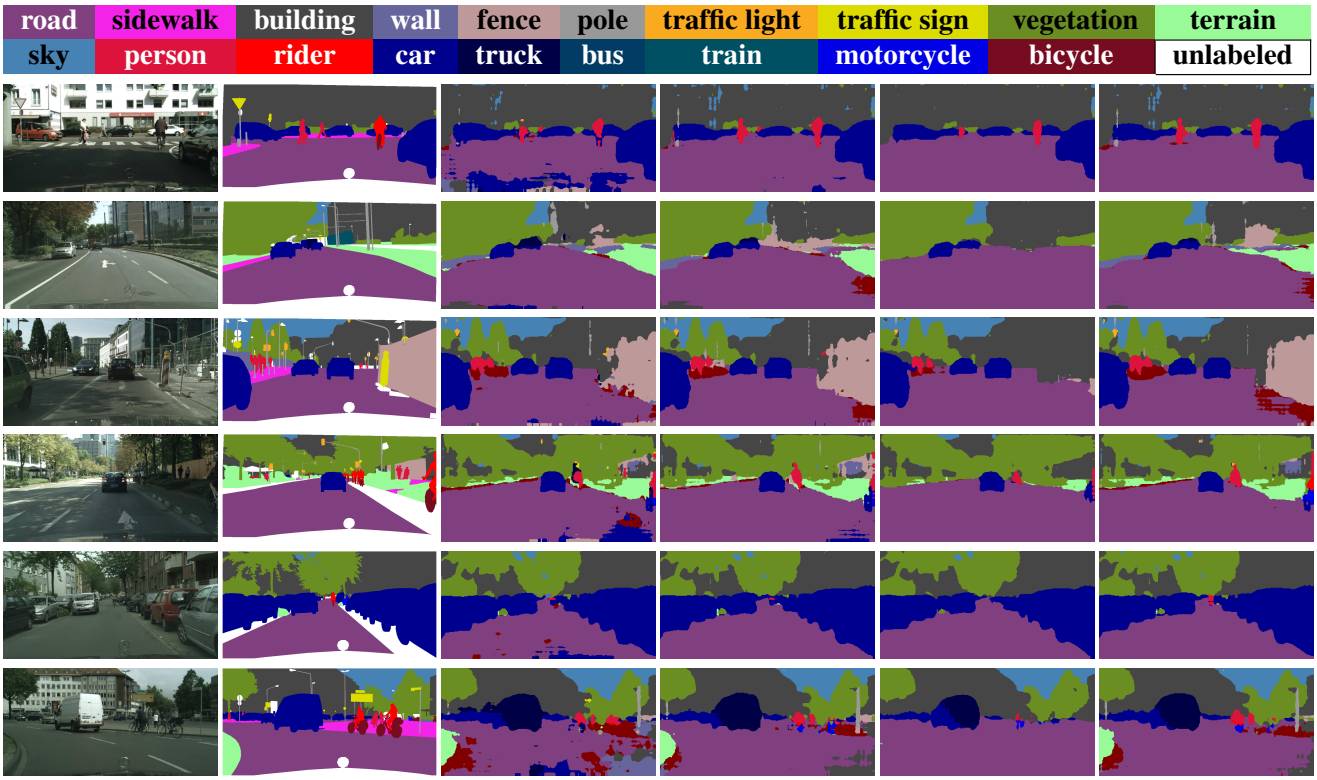

Figure 2: Semantic segmentation of some sample scenes extracted from the Cityscapes validation dataset. The network has been trained using GTA5 with annotations and Cityscapes for the unsupervised part (*best viewed in colors*).

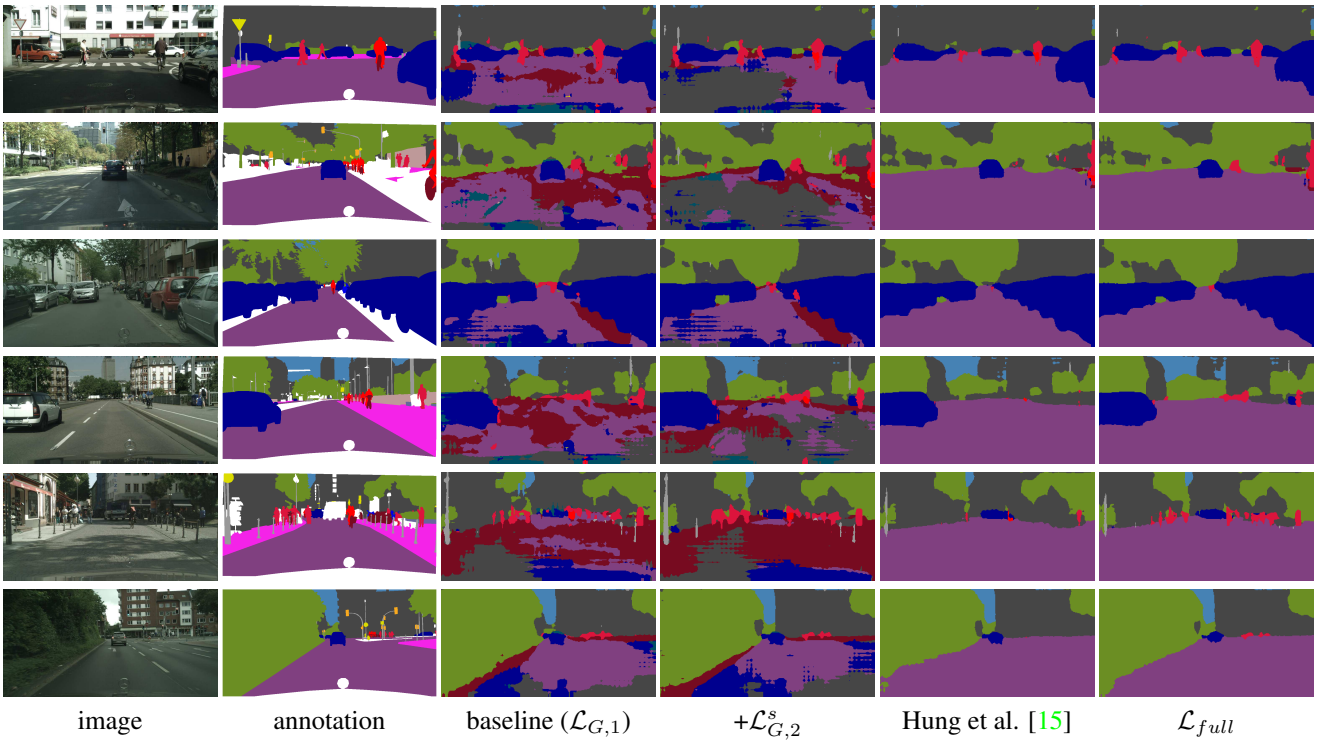

Figure 3: Semantic segmentation of some sample scenes extracted from the Cityscapes validation dataset. The network has been trained using SYNTHIA with annotations and Cityscapes for the unsupervised part (*best viewed in colors*).

by training on the SYNTHIA dataset some classes as *sidewalk* and *road* are highly corrupted. It is evident that a simple synthetic supervised training starting from this dataset would bring to a network which can not be used in an autonomous vehicle scenario. This is probably caused by the not completely realistic representation of streets and sidewalks in the SYNTHIA dataset, where their textures are often very unrealistic. Additionally, while the positioning of the camera in the Cityscapes dataset is always fixed and mounted on-board inside the car, in SYNTHIA the camera is placed in different positions. For example, the pictures can be captured from inside the car, from cameras looking from the top or from the side of the road.

Similarly to the *baseline* approach, the adversarial loss $\mathcal{L}^s_{G,2}$ is unable to adapt the network to the real domain, indeed the class *road* remains very badly detected also after its usage. Differently, Figure 3 shows how unsupervised data and the self-teaching component of the third loss allows to avoid all the artifacts on the *road* surface by reinforcing the segmentation network to capture the real nature of this class in the Cityscapes dataset. Also Hung's method [15] is able to correctly reconstruct the class *road*, avoiding the noise present in the *baseline*, but it suffers on small classes where it is outperformed by the proposed method. This is clearly visible on rows 4 and 5 of Figure 3, where our method is able to locate more precisely small classes as *person*.

5.1. Ablation Study

In this section, we are going to analyze the contributions of the various terms controlling the optimization in the proposed framework. Table 3 collects the results of this analysis on the Cityscapes validation split when using GTA5 as source dataset for the supervised part.

As it is possible to notice from Table 3, the generator network trained in a supervised way with the standard cross entropy loss (i.e., using only $\mathcal{L}_{G,1}$) is the less performing strategy achieving a mIoU of 27.9%. Some improvements can be obtained by adding the adversarial term $\mathcal{L}^s_{G,2}$ in the loss function, that is by exploiting also adversarial learning on the source dataset. In this case, the segmentation network is more accurate achieving a mIoU of 29.3%. The domain adaptation using adversarial learning on the target dataset only, i.e., $\mathcal{L}^t_{G,2}$ in combination with $\mathcal{L}_{G,1}$ obtains results very similar to the *baseline* approach. Instead, the exploitation of the self-teaching module $\mathcal{L}_{G,3}$ (without adversarial learning) allows to perform some adaptation to the segmentation network obtaining a mIoU of 28.7% (the main issue is the low performance on the road class since it is not able to remove the noise of the baseline method on it). The last row contains the results of the complete version of our approach, where all the aforementioned components are taken in consideration. We can appreciate that the full combination is able to outperform the exploitation of each of the single components and achieves a mIoU of 30.4%.

$\mathcal{L}_{G,1}$	$\mathcal{L}^s_{G,2}$	$\mathcal{L}^t_{G,2}$	$\mathcal{L}_{G,3}$	mIoU
✓				27.9
✓	✓			29.3
✓		✓		27.9
✓			✓	28.7
✓	✓	✓	✓	30.4

Table 3: Mean intersection over union (mIoU) of some configurations of our framework on the Cityscapes validation set using GTA5 as source dataset.

6. Conclusions

In this paper, a novel scheme to perform unsupervised domain adaptation from synthetic urban scenes to real world ones has been proposed. Two different strategies have been used to exploit unlabeled data: firstly an adversarial learning framework based on a fully convolutional discriminator and secondly a self-teaching strategy based on the assumption that predictions labeled as highly confident by the discriminator are reliable. Experimental results on the Cityscapes dataset prove the effectiveness of the proposed approach. In particular, we obtained good results also on challenging uncommon classes thanks to the class frequency dependent weighting of the self-teaching loss. This could enhance autonomous navigation in scenarios with tiny objects which can help characterizing the environment.

Further research will be devoted to test the proposed framework on other datasets and to the exploitation of different backbone networks. Additionally, we will investigate some improvements to the self-teaching strategy and to the exploitation of generative models to produce more realistic and refined synthetic training data to be fed to the framework.

Acknowledgments

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

References

[1] G.J. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, pages 88–97, 2009. 2

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. 40(4):834–848, 2018. 1, 2, 5

[3] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. *arXiv preprint arXiv:1812.05040*, 2018. 2

[4] Yuhua Chen, Wen Li, and Luc Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. pages 7892–7901, 2018. 2, 5

[5] 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. pages 3213–3223, 2016. 2, 4, 5

[6] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1635–1643, 2015. 2

[7] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189, 2015. 2

[8] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016. 2

[9] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, Pablo Martinez-Gonzalez, and Jose Garcia-Rodriguez. A survey on deep learning techniques for image and video semantic segmentation. *Applied Soft Computing*, 70:41 – 65, 2018. 2

[10] Boqing Gong, Fei Sha, and Kristen Grauman. Overcoming dataset bias: An unsupervised domain adaptation approach. In *NIPS Workshop on Large Scale Visual Recognition and Retrieval*, volume 3. Citeseer, 2012. 2

[11] 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 *Proceedings of the 35th International Conference on Machine Learning*, 2018. 2

[12] 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. 2, 4, 5, 6

[13] Seunghoon Hong, Hyeonwoo Noh, and Bohyung Han. Decoupled deep neural network for semi-supervised semantic segmentation. In *Advances in neural information processing systems*, pages 1495–1503, 2015. 2

[14] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7014–7023, 2018. 2

[15] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou34, Yen-Yu Lin, and Ming-Hsuan Yang15. Adversarial learning for semi-supervised semantic segmentation. 2018. 1, 2, 4, 5, 6, 7, 8

[16] Daniel Jiwoong Im, Chris Dongjoo Kim, Hui Jiang, and Roland Memisevic. Generating images with recurrent adversarial networks. *arXiv preprint arXiv:1602.05110*, 2016. 2

[17] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *European Conference on Computer Vision*, pages 158–171. Springer, 2012. 2

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[19] Xiaoming Liu, Jun Cao, Tianyu Fu, Zhifang Pan, Wei Hu, Kai Zhang, and Jun Liu. Semi-supervised automatic segmentation of layer and fluid region in retinal optical coherence tomography images using adversarial learning. *IEEE Access*, 7:3046–3061, 2019. 1, 2

[20] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. pages 3431–3440, 2015. 2

[21] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pages 97–105, 2015. 2

[22] Pauline Luc, Camille Couprie, Soumith Chintala, and Jakob Verbeek. Semantic segmentation using adversarial networks. In *NIPS Workshop on Adversarial Training*, 2016. 2

[23] Umberto Michieli and Leonardo Badia. Game theoretic analysis of road user safety scenarios involving autonomous vehicles. In *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pages 1377–1381. IEEE, 2018. 1

[24] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1742–1750, 2015. 2

[25] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. pages 1796–1804, 2015. 2

[26] Xingchao Peng and Kate Saenko. Synthetic to real adaptation with generative correlation alignment networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1982–1991. IEEE, 2018. 2

[27] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. volume 9906 of *LNCS*, pages 102–118. Springer International Publishing, 2016. 1, 2, 4

[28] 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. pages 3234–3243, 2016. 1, 2, 4, 5

[29] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. pages 3752–3761, 2018. 2, 4

[30] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. pages 2107–2116, 2017. 2

[31] Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi and weakly supervised semantic segmentation using generative adversarial network. *arXiv preprint arXiv:1703.09695*, 2017. 2

[32] Fengdong Sun and Wenhui Li. Saliency guided deep network for weakly-supervised image segmentation. *Pattern Recognition Letters*, 2019. 2

[33] Tatiana Tommasi, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. A deeper look at dataset bias. In *Domain Adaptation in Computer Vision Applications*, pages 37–55. Springer, 2017. 2

[34] A Torralba and AA Efros. Unbiased look at dataset bias. pages 1521–1528. IEEE Computer Society, 2011. 2

[35] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7472–7481, 2018. 2, 5

[36] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4068–4076, 2015. 2

[37] Alexander Vezhnevets and Joachim M Buhmann. Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning. pages 3249–3256. IEEE, 2010. 2

[38] Xiaolong Wang and Abhinav Gupta. Generative image modeling using style and structure adversarial networks. In *European Conference on Computer Vision*, pages 318–335. Springer, 2016. 2

[39] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. STC: A simple to complex framework for weakly-supervised semantic segmentation. 39(11):2314–2320, 2017. 2

[40] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *International Conference on Learning Representations (ICLR)*, 2016. 2, 6

[41] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. pages 2020–2030, 2017. 2, 4

[42] Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, and Tao Mei. Fully convolutional adaptation networks for semantic segmentation. pages 6810–6818, 2018. 2

[43] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. pages 2881–2890, 2017. 2

[44] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *European Conference on Computer Vision*, pages 597–613. Springer, 2016. 2

[45] Xinge Zhu, Hui Zhou, Ceyuan Yang, Jianping Shi, and Dahua Lin. Penalizing top performers: Conservative loss for semantic segmentation adaptation. pages 568–583, 2018. 2

[46] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 289–305, 2018. 2