FCNs in the Wild: Pixel-level Adversarial and Constraint-based Adaptation

Judy Hoffman
CS Department
Stanford University
jhoffman@cs.stanford

Dequan Wang
EECS Department
UC Berkeley
dqwang@cs.berkeley

Fisher Yu
CS Department
Princeton University
i@yf.io

Trevor Darrell
EECS Department
UC Berkeley
trevor@cs.berkeley

Abstract

Fully convolutional models for dense prediction have proven successful for a wide range of visual tasks. Such models perform well in a supervised setting, but performance can be surprisingly poor under domain shifts that appear mild to a human observer. For example, training on one city and testing on another in a different geographic region and/or weather condition may result in significantly degraded performance due to pixel-level distribution shift. In this paper, we introduce the first domain adaptive semantic segmentation method, proposing an unsupervised adversarial approach to pixel prediction problems. Our method consists of both global and category specific adaptation techniques. Global domain alignment is performed using a novel semantic segmentation network with fully convolutional domain adversarial learning. This initially adapted space then enables category specific adaptation through a generalization of constrained weak learning, with explicit transfer of the spatial layout from the source to the target domains. Our approach outperforms baselines across different settings on multiple large-scale datasets, including adapting across various real city environments, different synthetic sub-domains, from simulated to real environments, and on a novel large-scale dash-cam dataset.

1. Introduction

Semantic segmentation is a critical visual recognition task for a variety of applications ranging from autonomous agent tasks, such as robotic navigation and self-driving cars, to mapping and categorizing the natural world. As such, a significant amount of recent work has been introduced to tackle the supervised semantic segmentation problem using pixel-wise annotated images to train convolutional networks [20, 1, 23, 34, 19, 4, 33].

While performance is improving for segmentation models trained and evaluated on the same data source, there has yet been limited research exploring the applicability of these models to new related domains. Many of the challenges faced when considering adapting between visual domains for classification, such as changes in appearance, lighting, and pose, are also present when considering adapting for semantic segmentation. In addition, some new factors take on more prominence when considering recognition with localization tasks. In both classification and segmentation, the prevalence of classes may vary between different domains, but this variance can be more exaggerated with semantic segmentation applications as an individual object class may now appear many times within a single

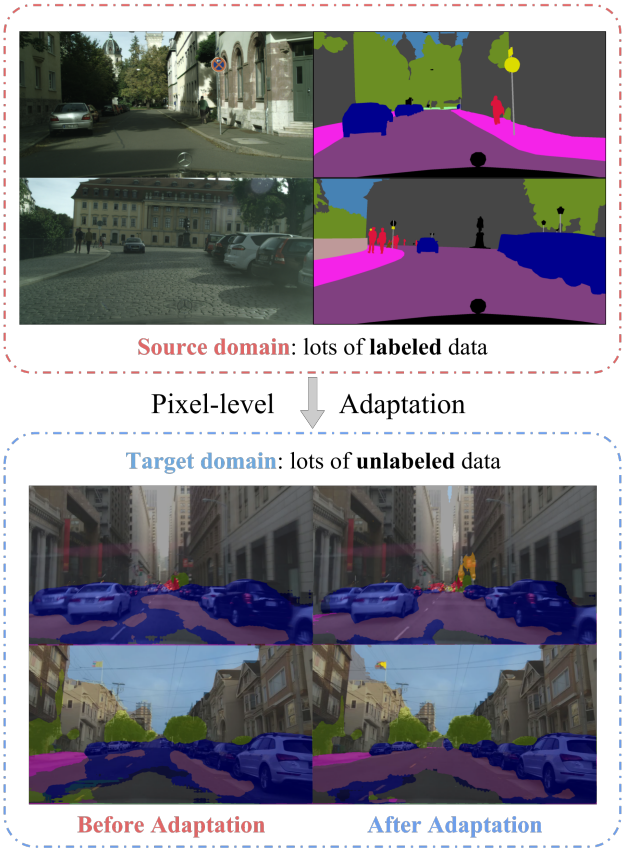


Figure 1: Unsupervised domain adaptation for pixel-level semantic segmentation.

scene. For instance, semantic segmentation for self-driving applications will focus on outdoor street scenes with objects of varying sizes, whose distribution may vary between cities or driving routes; in addition appearance statistics can vary considerably when, e.g., adapting a person recognition model trained only using indoor scene images. Moreover, pixel-wise annotations are expensive and tedious to collect, making it particularly appealing to learn to share and transfer information between related settings.

In this work, we propose the first unsupervised domain adaptation method for transferring semantic segmentation FCNs across image domains. A second contribution of our approach is the combination of global and local alignment methods, using global and category specific adaptation techniques that are themselves individually innovative contributions. We align the global statistics of our source and target data using a convolutional domain adversarial training technique, using a novel extension of previous image-level classification approaches [32, 6, 7]. Given a domain aligned representation space, we introduce a generalizable constrained multiple instance loss function, which expands on weak label learning [26, 25, 27, 24, 14], but can be applied to the target domain without any extra annotations and explicitly transfers category layout information from a labeled source dataset.

We evaluate our approach using multiple large scale datasets. We first make use of recently released synthetic drive-cam data from both the GTA5 [28] and SYNTHIA [29] datasets, in order to examine a large adaptation shift from simulated to the real images available in CityScapes [3]. Next, we explore the domain shift of cross season adaptation within the SYNTHIA dataset. We then focus on adaptation across cities in the real world. We perform a detailed quantitative analysis of cross-city adaptation within the CityScapes dataset.

A final contribution of our paper is the introduction of a new unconstrained drive-cam dataset for semantic segmentation, Berkeley Deep Driving Segmentation (BDDS). Below we demonstrate initial qualitative adaptation results from Cityscapes cities to the cities in BDDS. Across all of these studies, we show that our adaptation algorithm improves the target semantic segmentation performance without any target annotations.

2. Related Work

Semantic Segmentation Semantic segmentation is a key computer vision task and has been studied in a plethora of publications. Following the success of large-scale image classification, most current semantic segmentation models use some convolutional network architecture [5, 10] with many recent approaches using fully convolutional networks (FCNs) [20] to map the input RGB space to a semantic pixel space. These models are compelling because they allow a direct end-to-end function that can be trained using back propagation. The original FCN formulation has since been improved using dilated convolution [33] and post-processing techniques, such as Markov/conditional random fields [1, 19, 34].

Motivated by the high cost of collecting pixel level supervision, a related body of work has explored using weak labels (typically image-level tags defining presence / absence of each class), to improve semantic segmentation performance. Pathak *et al*. [26] and Pinheiro *et al*. [27] modeled this problem as multiple instance learning (MIL) and reinforce confident predictions during the learning process. An improved method was suggested by [24] who use an EM algorithm to better model global properties of the image segments. This work was in turn generalized by Pathak *et al*. who proposed a Constrained CNN which is able to model any linear constraints on the label space (*i.e*. presence / absence, percent cover) [25]. In another recent paper [15], Hong *et al*. used auxiliary segmentation to generalize semantic segmentations to categories where only weak label information was available.

From a domain adaptation perspective, these methods all assume that weak labels are present during training time for both source domain and target domain. In this work, we consider a related, but different learning scenario: strong supervision is available in the source domain, but that no supervision is available in the target domain.

Domain Adaptation Domain adaptation in computer vision has focused largely on image classification, with much work dedicated to generalizing across the domain shift between stock photographs of objects and the same objects photographed in the world [30, 17, 8]. Recent work includes [32, 6, 7] which all learn a feature representation which encourages maximal confusion between the two domains. Other work aims to align the features [21, 22] by minimizing the distance between their distributions in the two domains. Based on Generative Adversarial Network [9], Liu *et al*. proposed coupled generative adversarial network to learn a joint distribution of images from both source and target datasets [18].

Much less attention has been given to other important computer vision tasks such as detection and segmentation. In detection, Hoffman *et al*. proposed a domain adaptation system by explicitly modeling the representation shift between classification and detection models [11] along with a follow-up work which incorporated per-category adaptation using multiple instance learning [12]. The detection models were later converted into FCNs for evaluating semantic segmentation performance [13], but this work did not propose any segmentation specific adaptation approach. So far as we know, our method is the first to introduce domain adaptation techniques for semantic segmentation models.

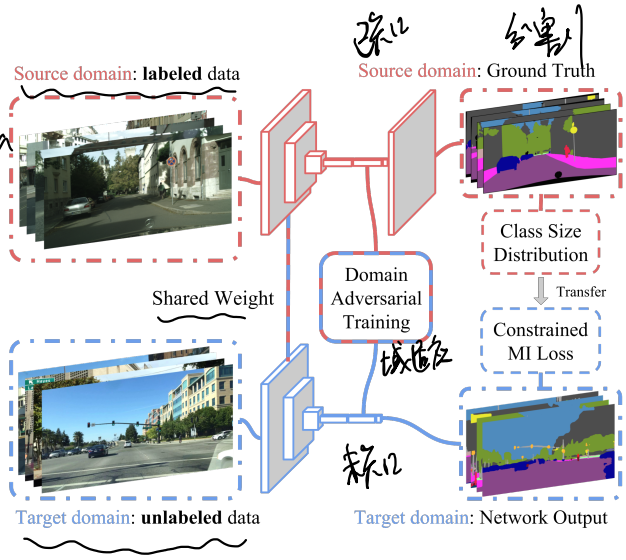


Figure 2: Overview of our pixel-level adversarial and constraint-based adaptation.

3. Fully Convolutional Adaptation Models

In this section, we describe our adaptation algorithm for semantic segmentation using fully convolutional networks (FCNs) across domains which share a common label space. Without loss of generality, our method can be applied to other segmentation models, though we focus here on FCNs due to their broad impact. We consider having access to a source domain, $\mathcal{S}$, with both images, $I_{\mathcal{S}}$, and labels, $L_{\mathcal{S}}$. We train a source only model for semantic segmentation which produces a pixel-wise per-category score map $\phi_{\mathcal{S}}(I_{\mathcal{S}})$.

Our goal is to learn a semantic segmentation model which is adapted for use on the unlabeled target domain, $\mathcal{T}$, with images, $I_{\mathcal{T}}$, but no annotations. We denote the parameters of such as network as $\phi_{\mathcal{T}}(\cdot)$. If there is no domain shift between the source and target domains then one could simply apply the source model directly to the target with no need for an adaptive approach. However, there is commonly a difference between the distribution of the source labeled domain and the target test domain.

Therefore, we present an *unsupervised adaptation* approach. We begin by noting that there are two main opportunities for domain shift. First, global changes may occur between the two domains resulting in a marginal distribution shift of corresponding feature space. This may occur between any two different domains, but will be most distinct in large shifts between very distinct domains, such as adapting between simulated and real domains. The second main shift occurs due to category specific parameter changes. This may result from individual categories having specific biases in the two domains. For example, when adapting between two different cities the distribution of cars and the appearance of signs may change.

We propose an unsupervised domain adaptation framework for adapting semantic segmentation models which directly tackles both the need for minimizing the global and the category specific shifts. For our model, we first make the necessary assumption that the source and target domains share the same label space and that the source model achieves performance greater than chance on the target domain. Then, we introduce two new semantic segmentation loss objectives, one to minimize the global distribution distance, which operates over both source and target images, $\mathcal{L}_{da}(I_{\mathcal{S}}, I_{\mathcal{T}})$. Another to adapt the category specific parameters using target images and transferring label statistics from the source domain $\mathcal{P}_{L_{\mathcal{S}}}$, $\mathcal{L}_{mi}(I_{\mathcal{T}}, \mathcal{P}_{L_{\mathcal{S}}})$. Finally, to ensure that we do not diverge too far from the source solution, which is known to be effective for the final semantic segmentation task, we continue to optimize the standard supervised segmentation objective on the source domain, $\mathcal{L}_{seg}(I_{\mathcal{S}}, L_{\mathcal{S}})$. Together, our adaptive learning approach is to optimize the following joint objective:

$$\begin{aligned}\mathcal{L}(I_{\mathcal{S}}, L_{\mathcal{S}}, I_{\mathcal{T}}) \quad &= \quad \mathcal{L}_{seg}(I_{\mathcal{S}}, L_{\mathcal{S}}) \\ &+\mathcal{L}_{da}(I_{\mathcal{S}}, I_{\mathcal{T}}) + \mathcal{L}_{mi}(I_{\mathcal{T}}, \mathcal{P}_{L_{\mathcal{S}}})\end{aligned} \tag{1}$$

We illustrate overall adaptation framework in Figure 2. Source domain data is used to update the standard supervised loss objective, trained using the source pixel-wise annotations. Both source and target data are used without any category annotations within fully-convolutional domain adversarial training to minimize the global distance of feature space between the two domains. Finally, category specific updates using a constrained pixel-wise multiple instance learning objective is performed on the target images, with source category statistics used to determine the constraints.

Note, our approach may be generally applied to any FCN-based semantic segmentation framework. For our experiments, we use the recently proposed front-end dilated fully convolutional network [33], based on 16 layers VGGNet [31], as our base model. There are 16 convolutional layers, where the last three convolutional layer converted from fully connected layers, called fc_6, fc_7, fc_8, followed by 8 times bilinear up-sample layer to produce segmentation in the same resolution as input image.

3.1. Global Domain Alignment

We begin by describing in more detail our global domain alignment objective, $\mathcal{L}_{da}(I_{\mathcal{S}}, I_{\mathcal{T}})$. Recall, that we seek to minimize the domain shift between representations of the source and target data. A recent line of research has shown that the domain discrepancy distance may be minimized through an adversarial learning procedure, whereby simultaneously a domain classifier is trained to best distinguish the source and target distributions and the representation space is updated according to the inverse objective [32, 2, 7]. The approaches heretofore have been introduced for classification models where each individual instance in the domain corresponds exactly to an image.

Here, we propose a new domain adversarial learning objective which may be applied for pixel-wise approaches to aid in learning domain invariant representations for semantic segmentation models. The first question to answer is what should comprise an instance within the dense prediction framework. Since recognition is sought at the pixel level alignment of full image representations will marginalize out too much distribution information limiting the alignment capability of the adversarial learning approach.

Instead, we consider the region corresponding to the natural receptive field of each spatial unit in the final representation layer (*e.g.* fc_7), as individual instances. In doing so, we directly supply our adversarial training procedure with the same information which is used to do final pixel prediction. Therefore, this provides a more meaningful view of the overall source and target pixel-space representation distribution distance which needs to be minimize.

Let $\phi_{\ell-1}(\theta, I)$ denote the output of the last layer before pixel prediction according to network parameters, θ. Then, our domain adversarial loss, $\mathcal{L}_{da}(I_{\mathcal{S}}, I_{\mathcal{T}})$ consists of alternating minimization objectives. One concerning the parameters of the representation space, θ, under which we would like to minimize the observed source and target distance, $\min d(\phi_{\ell-1}(\theta, I_{\mathcal{S}}), \phi_{\ell-1}(\theta, I_{\mathcal{T}}))$, for a given distance function, $d(\cdot)$. The second concerning estimating the distance function through training a domain classifier to distinguish instances of the source and target domains. Let us denote the domain classifier parameters as θ_D. We then seek to learn a domain classifier to recognize the difference between source and target regions and use that classifier to guide the distance minimization of the source and target representations.

Let $\sigma(\cdot)$ denote the softmax function and let the domain classifier predictions be indicated as $p_{\theta_D}(x) = \sigma(\phi(\theta_D, x)))$. Assuming the output of layer $\ell-1$ has $H \times W$ spatial units, then we can define the domain classifier loss, $\mathcal{L}_D$, as follows:

$$\mathcal{L}_D = -\sum_{I_{\mathcal{S}} \in \mathcal{S}} \sum_{h \in H} \sum_{w \in W} \log(p_{\theta_D}(R^{\mathcal{S}}_{hw})) \quad (2)$$

$$-\sum_{I_{\mathcal{T}} \in \mathcal{T}} \sum_{h \in H} \sum_{w \in W} \log(1 - p_{\theta_D}(R^{\mathcal{T}}_{hw})) \quad (3)$$

where $R^{\mathcal{S}}_{hw} = \phi_{\ell-1}(\theta, I_{\mathcal{S}})_{hw}$ and $R^{\mathcal{T}}_{hw} = \phi_{\ell-1}(\theta, I_{\mathcal{T}})_{hw}$ denote the source and target representation of each units, respectively.

For convenience let us also define the inverse domain loss, $\mathcal{L}_{Dinv}$ as follows:

$$\mathcal{L}_{Dinv} = -\sum_{I_{\mathcal{S}} \in \mathcal{S}} \sum_{h \in H} \sum_{w \in W} \log(1 - p_{\theta_D}(R^{\mathcal{S}}_{hw})) \quad (4)$$

$$-\sum_{I_{\mathcal{T}} \in \mathcal{T}} \sum_{h \in H} \sum_{w \in W} \log(p_{\theta_D}(R^{\mathcal{T}}_{hw})) \quad (5)$$

Finally, with these definitions, we may now describe the alternating minimization procedure.

$$\min_{\theta_D} \quad \mathcal{L}_D \quad (6)$$

$$\min_{\theta} \quad \tfrac{1}{2}\left[\mathcal{L}_D + \mathcal{L}_{Dinv}\right] \quad (7)$$

Optimizing these two objectives iteratively amounts to learning the best possible domain classifier for relevant image regions (Eq (6)) and then using the loss of that domain classifier to inform the training of the image representations so as to minimize the distance between the source and target domains (Eq (7)).

3.2. Category Specific Adaptation

Given our representation which has minimized the global domain distribution distance through our fully convolutional adversarial training objective, the next step is to further adapt our source model through modifying the category specific network parameters. In order to do this, we draw upon recent weak learning literature [25, 26], which introduced a fully convolutional constrained multiple instance learning objective. This work used size and existence constraints to produce a predicted target labeling to use for further training. We present the novel application of such approaches for domain adaptation and generalize the technique for use in our unlabeled setting.

First, we consider new constraints which are useful for our pixel-wise unsupervised adaptation problem. In particular, we begin by computing per image labeling statistics in the source domain, $\mathcal{P}_{L_{\mathcal{S}}}$. Specifically, for each source image which contains class c, we compute the percentage of image pixels which have a ground truth label corresponding to this class. We can then compute a histogram over these percentages and denote the lower 10% boundary as, α_c, the average value as δ_c, and the upper 10% as γ_c. We may then use this distribution to inform our target domain size constraints, thereby explicitly transferring scene layout information from the source to the target domain. For example, in a driving scenario, often the road occupies a large portion of the image while street signs occupy relatively little image real estate. This information is critical to the constrained multiple instance learning procedure. In contrast, prior work used a single size threshold across classes known to be in the image.

We begin by presenting our constrained multiple instance loss for the case where image-level labels are known. Thus, for a given target image for which a certain class c is present, we impose the following constraints on the output prediction map, $p = \arg\max \phi(\theta, I_{\mathcal{T}})$.

$$\delta_c \leq \sum_{h,w} p_{hw}(c) \leq \gamma_c \quad (8)$$

Thus, our constraint encourages pixels to be assigned to class c such that the percentage of the image labeled with class c is within the expected range observed in the source domain. Practically, we optimize this objective with lower bound slack to allow for outlier cases where c simply occupies less of the image than is average in the source domain. However, we do not allow slack on the upper bound constraint as it is important that no single class occupies too much of any given image. Notice that our updated constraint is general and can be equivalently applied to all classes regardless if they correspond with traditional object notion (*e.g*. bikes or people) or stuff notion (*e.g*. sky or vegetation).

Given this constraint we may now optimize for a new class prediction space to use for future learning. For the specific optimization details we refer the reader to Pathak *et al*. [25]. We provide one important modification. As we seek to optimize over both object and stuff categories, we note that the relative number of pixels devoted to each may vary significantly which could cause the model to diverge, over-fitting to those classes which are highly represented in the images. Instead, we use a simple size constraint that if the lower 10% of the source class distribution, α_c, is greater than 0.1, then we down-weight the gradients due to these classes by a factor of 0.1. This is re-weighting approach can be viewed as a re-sampling of the classes so as to come closer to a balanced set, allowing the relatively small classes potential to inform the learning objective.

While the approach described above describes a generalized constrained multiple instance objective, it relies on known image-level labels. Since we lack such information in our unsupervised adaptation setting, we now describe our procedure for predicting image level labels. Thus, our complete approach can be described as first predicting image-level labels and then optimizing for pixel predictions which satisfy the source transferred class size constraints.

In contrast to weakly-supervised settings, we do not learn a segmentation model from scratch with known image-level annotations. Instead, we have access to a fully supervised source dataset and use domain transferred constraints to facilitate transfer to an unsupervised target domain. Thus we both have a stronger initial model with a fully supervised using pixel-level annotations from source domain and are additionally able to regularize the learning procedure by training with weak label loss on target domain. Again, given a target image, $I_{\mathcal{T}}$, we compute the current output class prediction map, $p = \arg\max \phi(\theta, I_{\mathcal{T}})$. For each class we compute the percentage of pixels assigned to that class in our current prediction, $d_c = \frac{1}{H \cdot W} \sum_{h \in H} \sum_{w \in W} (p_{hw} = c)$. Finally, we assign an image-level label to class c if $d_c > 0.1 * \alpha_c$, meaning if we currently label at least as many pixels as 10% of the expected number for a true class appearing in the image.

4. Experiments

In this section, we report our experimental results on three different domain adaptation tasks: $cities \rightarrow cities$, $season \rightarrow season$, and $synthetic \rightarrow real$, studied across four different datasets. We analyze both our overall adaptation approach as well as the sub-components to verify that both our global and category specific alignment offer meaningful contributions.

For all experiments we use the front-end dilated fully convolutional network [33] as both the initialization for our method and as the baseline model for comparison. All code and models are trained and evaluated in the Caffe [16] framework and will be made available before camera-ready.

For fair comparison, we use the Intersection over Union (IoU) evaluation metric for all experiments. For $cities \rightarrow cities$ and $synthetic \rightarrow real$ tasks, we followed the evaluation protocol of [3] and train our models with 19 semantic labels of Cityscapes. For $season \rightarrow season$ task, we use 13 semantic labels of SYNTHIA instead.

4.1. Datasets

Cityscapes contains 34 categories in high resolution, 2048×1024. The whole dataset is divided into three parts: $2,975$ training samples, 500 validation samples and $1,525$ test samples. The split of this dataset is city-level, which covers individual European cities in different geographic and population distribution.

SYNTHIA contains 13 classes with different scenarios and sub-conditions. As for $season \rightarrow season$ task, we regard SYNTHIA-VIDEO-SEQUENCES as play ground. There are 7 sequences, covering different scenarios (highway, roundabout, mountain path, New York City, Old European Town) with several sub-sequences, such as seasons(Spring, Summer, Fall, Winter), weathers(Rain, Soft-Rain, Fog), and illuminations(Sunset, Dawn, Night). These frames are captured by 8 RGB cameras forming a binocular 360° visual field. In order to minimize the impact of viewpoint, we only pick up the dashcam-like frames for all the time. As for $synthetic \rightarrow real$ task, we take SYNTHIA-RAND-CITYSCAPES, providing $9,000$ random images from all the sequences with Cityscape-compatible annotations, as source domain data.

GTA5 contains $24,966$ high quality labeled frames from realistic open-world computer games, Grand Theft Auto V (GTA5). Each frame, with high resolution 1914×1052, is generated from fictional city of Los Santos, based on Los Angeles in Southern California. We take the whole dataset with labels compatible to Cityscapes categories for $synthetic \rightarrow real$ adaptation.

BDDS contains thousands of dense annotated dashcam video frames and hundreds of thousands of unlabeled frames. Each sample, with high resolution 1280×720, provides 34 categories compatible to Cityscapes label space.

GTA5 → Cityscapes																				
Method	road	sidewalk	building	wall	fence	pole	t light	t sign	veg	terrain	sky	person	rider	car	truck	bus	train	mbike	bike	mIoU
Dialation Frontend [33]	31.9	18.9	47.7	7.4	3.1	**16.0**	10.4	1.0	76.5	13.0	58.9	36.0	1.0	67.1	**9.5**	3.7	0.0	0.0	0.0	21.1
Our Method (GA only)	67.4	29.2	**64.9**	**15.6**	**8.4**	12.4	9.8	**2.7**	74.1	12.8	**66.8**	38.1	2.3	63.0	9.4	5.1	0.0	3.5	0.0	25.5
Our Method (GA + CA)	**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**

SYNTHIA → Cityscapes																				
Dialation Frontend [33]	6.4	17.7	29.7	1.2	0.0	15.1	0.0	7.2	30.3	0.0	66.8	51.1	1.5	47.3	0.0	3.9	0.0	0.1	0.0	14.7
Our Method (GA only)	**11.5**	18.3	**33.3**	**6.1**	0.0	**23.1**	0.0	11.2	**43.6**	0.0	**70.5**	45.5	1.3	45.1	0.0	**4.6**	0.0	0.1	0.5	16.6
Our Method (GA + CA)	**11.5**	**19.6**	30.8	4.4	0.0	20.3	0.1	**11.7**	42.3	0.0	68.7	**51.2**	**3.8**	**54.0**	0.0	3.2	0.0	**0.2**	**0.6**	**17.0**

Table 1: **Adaptation from synthetic to real.** We study the performance using GTA5 and SYNTHIA as source labeled training data adapted and Cityscapes *train* as an unlabeled target domain, while evaluating our adaptation algorithm on Cityscapes *val*. Meanwhile, we show an ablation of the components of our method and how each contributes to the overall performance of our approach. Here GA represents global domain alignment and CA indicates category specific adaptation.

The majority of our data comes from New York and San Francisco, which are the representative of eastern and western coasts. Different from the other existing driving datasets, this dataset covers diverse driving scenarios under different conditions, such as urban street view at night, highway scene in rain and so on, providing challenging domain adaptation settings.

4.2. Quantitative and Qualitative Results

We broadly study three types of shifts. First we study a large distribution shift, as seen when adapting from simulated to real imagery. Next, we study a medium sized shift, through adaptation across season patterns observed within the SYNTHIA dataset. Finally, we explore situations of relatively smaller domain shift, though exploring adaptation between different cities within the CityScapes dataset.

4.2.1 Large Shift: Synthetic to Real Adaptation

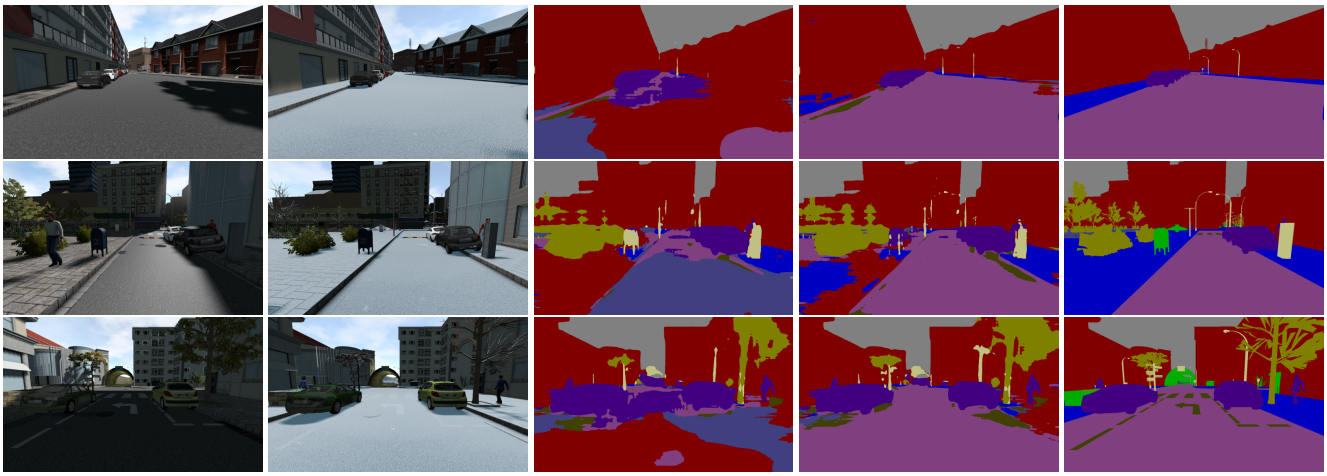
We begin the evaluation of our method by studying the large domain shift of adapting between simulated driving data and real world drive-cam data. Table 1 shows semantic segmentation performance for the shift between GTA5 to CityScapes and between SYNTHIA to CityScapes. This illustrates that even with this large domain difference our unsupervised adaptation solution is capable of improving the performance of the source dilation model. Notice that for this larger shift setting, such as GTA5→Cityscapes, the domain adversarial training contributes 4.4% raw and $\sim 20\%$ relative percentage mIoU improvement and multiple instance loss contributes yet another 1.6% raw and $\sim 6\%$ relative percentage mIoU improvement. As for SYNTHIA→Cityscapes, our method also offers a measurable improvement.

4.2.2 Medium Shift: Cross Seasons Adaptation

As our next experiment, we seek to analyze adaptation across season patterns. To this end, we use the SYNTHIA dataset which has synthetic images available along with season annotations. We first produce one domain per each of the season labels available: Summer, Fall and Winter. We then perform adaptation across each of the 6 shifts and report the performance of our method in comparison to the source dilation model in Table 2. On average we get ~ 3 percentage mIoU improvement for $season \rightarrow season$ adaptation and find that for 12/13 object categories our adaptation method provides higher mIoU. The one class we saw no improvement after adaptation is for `car`. We presume this results from the fact that cars have little or no appearance difference across seasons in this synthetic dataset. For example, consider the qualitative results shown in Figure 3 for the shift of fall to winter. While the roads and side-walks have been rendered in a white to simulate snow in winter, the cars are rendered in the same appearance as in fall. In fact some of the largest performance improvements we saw from our method we in categories like `road` in the shift of fall to winter, and our method is able to overcome this large appearance shift.

4.2.3 Small Shift: Cross City Adaptation

For our third quantitative experiment we move towards studying cross city adaptation within the CityScapes dataset. In Table 3 we report performance on the task of adapting between the labeled cities in the CityScapes *train* to the unlabeled cities in either the Cityscapes *val*. The top row shows the performance of the dilation frontend model [33]. We report performance after only global alignment through domain adversarial training (indicated



(a) Fall Image (b) Winter Image (c) Before Adaptation (d) After Adaptation (e) Ground Truth

Figure 3: Qualitative results on adaptation from cities in SYNTHIA *fall* to cities in SYNTHIA *winter*.

Method	Source	Target (test)	sky	building	road	sidewalk	fence	vegetation	pole	car	t sign	pedestrian	bicycle	lanemarking	t light	mIoU
Before Adapt	Summer	Fall	94.7	91.6	95.2	90.4	86.9	71.6	50.0	87.0	52.9	64.3	57.8	72.3	18.8	71.8
After Adapt	Summer	Fall	95.0	91.6	94.6	90.3	86.8	71.7	49.4	87.3	54.7	64.3	54.2	69.7	20.8	71.6
Before Adapt	Fall	Summer	95.0	92.9	96.7	91.9	90.2	71.8	53.4	93.2	50.6	62.3	48.1	82.7	14.7	72.8
After Adapt	Fall	Summer	95.4	93.5	97.2	92.9	91.6	73.7	56.7	92.3	57.4	68.5	55.4	85.3	31.7	76.4
Before Adapt	Summer	Winter	91.4	72.5	80.1	8.6	66.1	51.6	43.1	62.9	55.5	53.9	39.4	47.8	22.3	53.5
After Adapt	Summer	Winter	90.7	71.1	78.4	9.2	64.0	50.9	42.8	60.5	56.1	53.5	40.7	49.2	23.0	53.1
Before Adapt	Winter	Summer	91.2	90.5	82.5	34.2	53.3	59.1	49.2	85.7	44.7	62.8	46.3	44.6	28.9	59.5
After Adapt	Winter	Summer	94.8	90.4	81.8	46.0	59.6	65.1	51.8	87.2	48.4	62.3	47.9	42.0	35.1	62.5
Before Adapt	Fall	Winter	92.0	78.8	81.8	15.5	31.7	52.3	43.4	63.8	41.3	57.8	48.6	56.7	11.3	51.9
After Adapt	Fall	Winter	92.1	86.7	91.3	20.8	72.7	52.9	46.5	64.3	50.0	59.5	54.6	57.5	26.1	59.6
Before Adapt	Winter	Fall	94.3	88.0	85.5	32.0	56.2	60.9	48.7	77.2	47.0	57.9	49.8	46.6	27.1	59.3
After Adapt	Winter	Fall	94.5	87.0	84.0	46.5	62.7	65.8	51.0	68.1	55.7	58.5	53.9	48.3	30.8	62.0
Before Adapt	Avg	Avg	93.1	85.7	87.0	45.4	64.1	61.2	48.0	**78.3**	48.7	59.8	48.3	58.5	20.5	61.5
After Adapt	Avg	Avg	**93.8**	**86.7**	**87.9**	**51.0**	**72.9**	**63.4**	**49.7**	76.6	**53.7**	**61.1**	**51.1**	**58.7**	**27.9**	**64.2**

Table 2: **Adaptation across seasons.** We study the cross season performance using sub-sequences of SYNTHIA dataset. We report quantitative comparisons of performance before and after adaptation for training on one season and evaluating on another unannotated novel season. (Avg: the average performance of adaptation from one to another.)

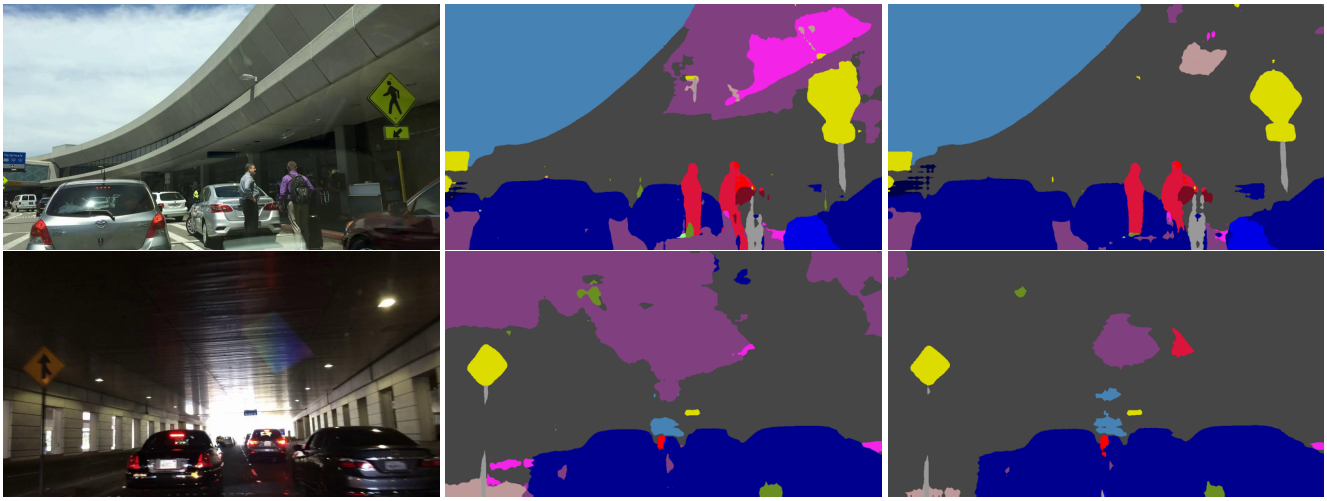
as Our method (GA only)) and after the category specific alignment with the constrained multiple instance loss (indicated as Our method (GA+CA)). We note that for this adaptation experiment the majority of the improvement from our method is as a result of the domain adversarial training (3.6 percentage mIoU) whereas after category specific alignment only offers a noticeable improvement on the categories of `traffic light`, `rider` and `train`. One reason could be that the domain shift between *train* and *val* mainly results from a change in the global appearance, due to the difference in city, whereas the specific category appearance may not change that significantly. Since performance on this within dataset adaptation is already quite high, the primary improvements arise from producing more consistent within object segmentations.

4.3. BDDS Adaptation

Finally, we analyze another real world $cities \rightarrow cities$ adaptation using our new large scale driving image dataset BDDS. To understand this difficulty and evaluate our methods more extensively, we create a new image dataset based on dash cam videos. Although CityScapes covers various cities in Germany and neighboring countries, we observe that the cities in the other places have different visual ap-

Cityscapes *train* → Cityscapes *val*																				
Method	road	sidewalk	building	wall	fence	pole	t light	t sign	veg	terrain	sky	person	rider	car	truck	bus	train	mbike	bike	mIoU
Dialation Frontend [33]	96.2	76.0	88.4	32.5	46.4	53.5	52.0	68.7	88.6	46.6	91.0	74.8	46.0	90.5	46.9	58.0	44.7	45.2	70.3	64.0
Our Method (GA Only)	**97.0**	**79.6**	89.6	**42.8**	**49.9**	55.0	55.2	**70.2**	**91.2**	**59.8**	92.5	75.4	46.5	91.6	**51.4**	**66.0**	49.3	**48.9**	**71.6**	67.6
Our Method (GA + CA)	**97.0**	**79.6**	**89.8**	42.2	49.0	**55.4**	**56.3**	70.1	**91.2**	**59.8**	**92.6**	**75.5**	**48.1**	**91.7**	50.4	65.8	**53.2**	48.0	**71.6**	**67.8**

Table 3: **Adaptation across cities.** We study the performance using Cityscapes *train* cities as source labeled training data adapted and evaluate our adaptation algorithm on Cityscapes *val* as unlabeled target domains. Meanwhile, we show an ablation of the components of our method and how each contributes to the overall performance of our approach. Here GA indicates global domain alignment in section 3.1 and CA represents category specific adaptation in section 3.2.



(a) Original Image (b) Before Adaptation (c) After Adaptation

Figure 4: Qualitative results on adaptation from cities in Cityscapes to cities in BDDS.

pearance and street layout. They may post serious challenges to the models learned from CityScapes. Up to now, we have collected more than $100,000$ images covering outdoor scenes at different time and places. Based on the current annotation progress, there would $5,000{\sim}10,000$ images with fine segmentation annotation before CVPR 2017. We aim for $10,000{\sim}20,000$ finely segmented street scene images eventually.

We take $\sim 60,000$ images in area of San Francisco from BDDS and study how well we can adapt the model learned on Cityscapes to San Francisco. Because our methods don't require labels in the target domain, we can use all the new images in our training the adaptation. Some results are shown in Figure 4. From these qualitative results, we observe that there is a significant segmentation quality drop when a model trained on Cityscapes is used in BDDS directly. It usually appears as noisy segmentation or wrong context. After adaptation, the segmentation results usually become much cleaner. We expect to conduct extensive quantitative evaluation when annotations are ready.

5. Conclusion

In this paper, we present an unsupervised domain adaptation framework with fully convolutional networks for semantic segmentation. We propose fully convolutional networks with domain adversarial training for global domain alignment, while leveraging class-aware constrained multiple instance loss for transferring spatial layout. We demonstrate the effectiveness of our method on domain shifts between different cities, seasons and from synthetic to real, and we offer a new large-scale real-city driving image dataset. While the task of image classification has seen the bulk of the effort in developing domain adaptation methods, our experiments demonstrate the importance of adaptation in pixel-level dense prediction as well. Our approach is the first step in this direction.

References

[1] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, 2015. 1, 2

[2] W. Chen, H. Wang, Y. Li, H. Su, D. Lischinsk, D. Cohen-Or, B. Chen, et al. Synthesizing training images for boosting human 3d pose estimation. In *3DV*, 2016. 3

[3] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 2, 5

[4] J. Dai, K. He, and J. Sun. Instance-aware semantic segmentation via multi-task network cascades. In *CVPR*, 2016. 1

[5] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *TPAMI*, 2013. 2

[6] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015. 2

[7] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 2016. 2, 3

[8] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2012. 2

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014. 2

[10] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *ECCV*, 2014. 2

[11] J. Hoffman, S. Guadarrama, E. S. Tzeng, R. Hu, J. Donahue, R. Girshick, T. Darrell, and K. Saenko. Lsda: Large scale detection through adaptation. In *NIPS*, 2014. 2

[12] J. Hoffman, D. Pathak, T. Darrell, and K. Saenko. Detector discovery in the wild: Joint multiple instance and representation learning. In *CVPR*, 2015. 2

[13] J. Hoffman, D. Pathak, E. Tzeng, J. Long, S. Guadarrama, T. Darrell, and K. Saenko. Large scale visual recognition through adaptation using joint representation and multiple instance learning. *JMLR*, 2016. 2

[14] S. Hong, H. Noh, and B. Han. Decoupled deep neural network for semi-supervised semantic segmentation. In *NIPS*, 2015. 2

[15] S. Hong, J. Oh, B. Han, and H. Lee. Learning transferrable knowledge for semantic segmentation with deep convolutional neural network. In *CVPR*, 2016. 2

[16] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, 2014. 5

[17] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *CVPR*, 2011. 2

[18] M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. In *NIPS*, 2016. 2

[19] Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang. Semantic image segmentation via deep parsing network. In *ICCV*, 2015. 1, 2

[20] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1, 2

[21] M. Long, Y. Cao, J. Wang, and M. Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015. 2

[22] M. Long, J. Wang, and M. I. Jordan. Unsupervised domain adaptation with residual transfer networks. In *NIPS*, 2016. 2

[23] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *ICCV*, 2015. 1

[24] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *ICCV*, 2015. 2

[25] D. Pathak, P. Krahenbuhl, and T. Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *ICCV*, 2015. 2, 4, 5

[26] D. Pathak, E. Shelhamer, J. Long, and T. Darrell. Fully convolutional multi-class multiple instance learning. In *ICLR Workshop*, 2015. 2, 4

[27] P. O. Pinheiro and R. Collobert. From image-level to pixel-level labeling with convolutional networks. In *CVPR*, 2015. 2

[28] S. R. Richter, V. Vineet, S. Roth, and V. Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016. 2

[29] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016. 2

[30] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*, 2010. 2

[31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 3

[32] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In *ICCV*, 2015. 2, 3

[33] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016. 1, 2, 3, 5, 6, 8

[34] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr. Conditional random fields as recurrent neural networks. In *ICCV*, 2015. 1, 2