Examining the Effects of Supervision for Transfer from Synthetic to Real Driving Domains

Vashisht Madhavan
vashisht.madhavan@berkeley.edu

Abstract

With the surge of autonomous driving efforts in industry, semantic understanding of road scenes has become more commercially relevant than ever. Semantic segmentation, or dense pixel prediction, has become a popular approach for scene understanding, as the meteoric rise of deep convolutional networks (CNNs) has led to significant progress in recent years. However, these deep networks require large amounts of data, labeled at the pixel level, which can be quite cumbersome to collect for the driving domain. As a result, many have looked to realistic virtual world simulators (i.e. video games), for collecting large volumes of labeled data. Despite the visual similarity between real and synthetic domains, models trained on synthetic street scenes show poor results when evaluated on real-world data. To compensate for this visual shift, supervision from the real domain is necessary. In this work, I examine how much real-world supervision is appropriate for effective transfer of models pretrained on synthetic. Utilizing recent methods of supervised and semi-supervised transfer for fully convolutional networks (FCNs), I achieve promising results with a very small amount of labeled data(~ 50 images). By also quantitatively measuring different levels of domain shift, I reveal how simple metrics about a synthetic domain can be used to infer the ability of network features to transfer.

1. Introduction

Perception for autonomous vehicles is now at the forefront of many computer vision research efforts. With the automobile industry investing heavily in the autonomous driving market, computational systems for understanding road scenes are more commercially relevant than ever. One of the most promising directions for achieving vehicle perception is semantic segmentation, or dense pixel-level prediction. With the emergence of deep convolutional neural networks and large labeled datasets, numerous advances in image classification[21, 12, 26] and object detection[9, 20] have been made in recent years. In a similar manner, fully convolutional networks (FCNs)[13, 25] have proven successful for semantic segmentation tasks. However, these networks require large amounts of image data and corresponding pixel-level labels to learn effective features for segmentation. In the domain of driving, collecting and labeling a large number of driving scenes in not only cumbersome, but also very expensive. Additionally, existing public datasets, such as Cityscapes[5], contain a relatively small number of labeled images (e.g. 5000). Previous work by Bengio[1] and Yosinski et al[24] shows that training a deep network on a large, general purpose dataset(i.e. ImageNet) can alleviate the need for large amounts of supervision in other image domains. Essentially, the features learned from the large,source dataset provide a good baseline for a target domain of choice. Until recently, however, these large datasets contained images that were too different from those in driving domains, and thus large volumes of supervised driving data were still necessary.

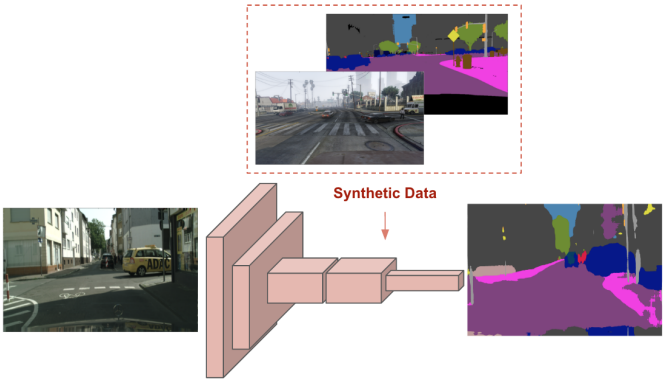

Figure 1. I take a network pretrained on synthetic data and transfer to the real domain

In recent months, learning representations from street scenes in virtual world simulators has become a promising area of research. As these simulations (i.e. video games) are now more realistic than ever and labeled data is very easy to collect within a game engine, researchers have looked to synthetic driving data for pretraining semantic segmentation models. Recently, two large scale synthetic driving datasets: GTAV[17] and SYNTHIA[18], have been released

for this very purpose. Along with the release of their respective datasets, Ros et al.[18] and Richter et al.[17], show that training of segmentation models with both real and synthetic data shows improvement over simply training with just real-world data. Although these results are promising, networks pretrained on only synthetic data perform poorly when evaluated in the real domain. This performance degradation can be attributed to high domain shift in the pixel space, which is often mild to a human observer. This visual shift consists of changes in lighting, pose, and appearance between objects in the synthetic and real domains. Since semantic segmentation also requires localization of multiple objects in an image, this shift can be more prominent if the size and distribution of objects between domains differ. As a result, adaptation of these networks with real world supervision is necessary if they are to be applied in practice.

In this work, I explore the effect of adding real-world supervision to FCNs pretrained on synthetic data. As the goal is to leverage the representations learned from synthetic data, I vary the level of real domain supervision to determine how much data is sufficient for reasonable performance. Specifically, I explore this impact with respect to the popular supervised transfer method of fine-tuning[24], in which network weights pretrained on an arbitrary source domain are learned towards a target domain. In addition, I experiment with a novel method of semi-supervised transfer, extended from very recent work in unsupervised domain adaptation[10]. Unlike fine-tuning, the semi-supervised method also utilizes the raw,unsupervised data in the real domain, which provides additional information for transferring representations.

As a final extension to my experiments, I quantitatively measure domain shift between real and synthetic domains in two distinct ways, which I discuss further in Section 3. By looking at these measures for both synthetic datasets, along with their corresponding results after transfer, I analyze how transfer performance can be inferred with simple domain knowledge. Since the domain shift for semantic segmentation is more nuanced than that of image classification, this analysis provides interesting insights for adapating FCNs not only for driving domains, but for general purpose datasets.

2. Related Work

Semantic segmentation is one of the most studied fields in computer vision, with applications most prominent for autonomous driving and robot navigation. Following the success of convolutional network architectures, fully convolutional networks have (FCNs)[13] have emerged as a popular approach for solving semantic segmentation tasks. Although already effective for dense prediction, FCNs have been further improved with dilated convolutions[25] and conditional random fields[3] for post-processing. Effective learning of FCN parameters, however, requires large amounts of labeled data, which can be hard to obtain in practice. For the domain of autonomous driving, data collection difficulties are further aggravated by the high cost of setting up camera rigs for vehicles. As a result, large, labeled datasets of synthetic street scenes have been explored as a proxy for real-world supervision.

Work from Ros et.al[18] outlines the creation of the SYNTHIA dataset and evaluates how Balanced Gradient Contribution (BGC), which combines real and synthetic data in each training batch, improves segmentation results on the CamVid[2] and KITTI[8] datasets. Although this method shows improvements over models trained only on real driving data, CamVid and KITTI are too small($\sim$ 300 images) for stable training of network parameters and do not contain diverse driving scenarios. In addition, the results only show improvements for the T-Net architecture[19] and not the more commonly used FCN. In this work, I examine how the performance of pretrained FCNs changes as more data from the real domain becomes available. In addition, I remedy the drop in FCN performance from BGC by using the dilation FCN model from Yu and Koltun[25].

Richter et.al[17] also showed improved segmentation results on CamVid and KITTI by training dilation FCNs with a combination of real and synthetic data. As with Ros et al., KITTI and CamVid do not have the diversity or size of Cityscapes, which limits their ability to train convolutional networks. This dearth of images for training also makes the results from the paper very misleading. In addition, their work did not analyze how real domain performance changes with supervision or explore the impact of domain shift on transfer.

Adapting visual classifiers between domains has been been extensively studied in computer vision literature[7, 24, 1]. In the context of CNNs, effective methods adapting network weights trained on one dataset to a different domain/task is an area of active research. One of the most promising approaches in recent years has been domain adversarial training[22, 23, 10], which encourages maximal confusion between feature representations of the two domains. This is achieved by an adversarial learning mechanism between a domain classifier and a domain alignment module, further described in Tzeng et al[22]. Very recent work by Hoffman et al.[10] applied domain adversarial learning for semantic segmentation and work by Pathak et.al[16, 15] uses Multiple-Instance Learning (MIL) to learn segmentation maps from weak image-level labels. Despite the promise of these works for adapting semantic segmentation models, most research in domain adaptation focuses on the task of image classification. In my experiments, I examine how these novel adaptation methods can be improved with more target domain supervision.

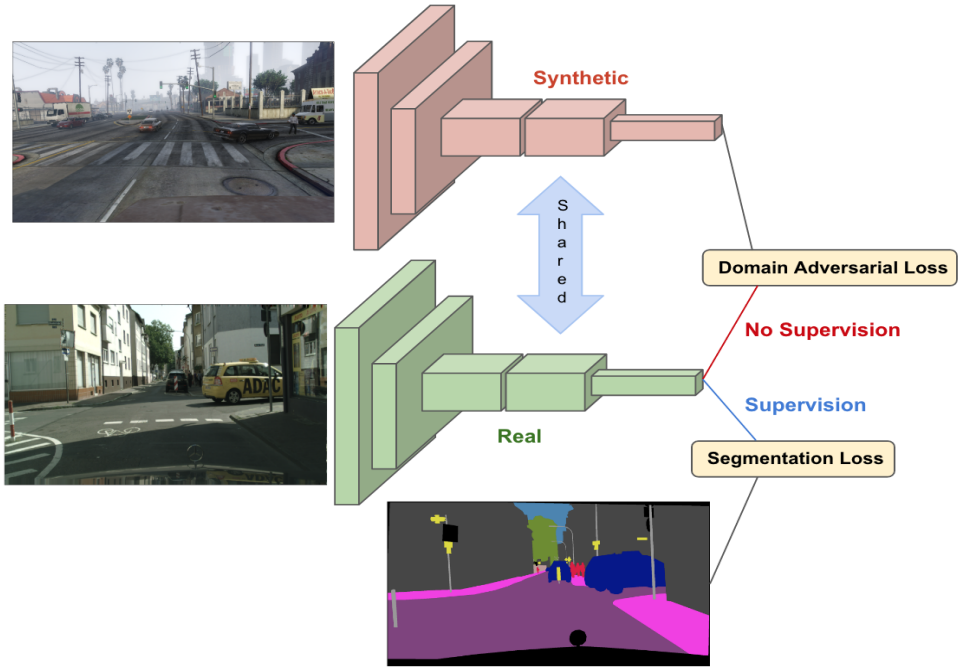

Figure 2. Architecture of Semi-Supervised Transfer Method with Adversarial Loss

Specifically, I propose a semi-supervised extension to the model outline by Hoffman et al., by including a strongly supervised loss for target domain images with labels. Chu and Madhavan et. al[4] explored image classification transfer with respect to quantitative measures of domain shift (i.e. classification accuracy between two domains) and the amount of target domain data available. I take the same analytical approach in this work, except for the task of semantic segmentation. Since localization is as equally important as recognition in segmentation tasks, I consider a more nuanced view of domain shift between source and target domains.

3. Experiments

In experiments, I evaluate the performance of dilation FCN models, pretrained on the SYNTHIA and GTA datasets, on the real-world driving dataset, Cityscapes. All results reported below are from evaluation on the Cityscapes validation set, which contains 500 images of diverse driving scenes. After training dilation FCNs for each synthetic dataset, varying amounts of labeled data from the Cityscapes training set are used for fine-tuning and semi-supervised adaptation. All hyperparameters are kept the same between experiments the Intersection over Union (IoU) metric is used for fair comaparison. All models are implemented in the Caffe[11] deep learning framework. Although learning rates slightly vary between experiments, each network is trained for 100K iterations. Vanilla SGD is used to optimize each network, using a momentum of 0.9 and L2 regularization to prevent overfitting. Details about each experiment are outlined in the sections below.

3.1. Supervised Fine-Tuning

One of the most popular methods of supervised transfer is fine-tuning[24]. As described previously, weights from a network trained on a large,labeled source dataset, are copied to a target network and trained towards the target domain/task. I take the models pretrained on the GTA and SYNTHIA datasets and train them towards the target domain of Cityscapes. I also vary the amount of supervised target data available, similar to the approach in Chu and Madhavan[4], to not only determine the critical amount of supervision necessary for effective transfer, but to also understand the rate at which target domain performance changes with more data. The results are shown in Table 3

3.2. Semi-Supervised Adaptation

To extend my analysis of supervisory effects, I explored transfer from a semi-supervised perspective. I simply extend the architecture from Hoffman et al.[10] to include a Softmax loss for the target domain. When labeled data is available for the real driving domain, a Softmax loss is used to update network weights. When data from Cityscapes has no supervision, however, the domain adversarial loss, described in [22, 23], is used to update network parameters. Figure 2 provides a visualization of this modified architec-

ture. This method allows the network to learn from both supervised and unsupervised image data in the target domain. Naturally, the results from these experiments provide a better estimate of the target supervision needed for good performance. As with the supervised experiment, I vary the level of supervised Cityscapes data to evaluate the impact of increasing target supervision. The results are shown in Table 4.

3.3. Domain Shift: Global and Local

The experiments described above analyze transfer with respect to the amount of target supervision, yet they only consider transfer performance with respect to one variable: the level of supervision. Target domain performance, regardless of the transfer method, is also dependent on the similarity between the source and target domain. If the source and target data exhibit high domain shift, it becomes more difficult to transfer learned representations, thus increasing amount of target supervision necessary for good results[24, 4]. To quantify this domain shift, I trained simple domain classifiers on the $fc7$ features of a VGG-16 classification architecture[21] pretrained on the PASCAL VOC dataset[6]. By passing images from both synthetic and real domains through the network, I obtain features for domain classification SVMs, which give an estimate of how different two domains are[4]. The classification accuracy also aligns with more common measures of domain shift, such as Maximum Mean Discrepancy (MMD)[14]. However, this accuracy is only representative of global differences between images in each dataset, and fail to account for visual differences between individual categories. Recent weak adaptation and weak learning literature[15, 16] suggests that these category-specific shifts, such as the distribution of a cars in an image, cause performance drops for semantic segmentation tasks. To measure domain shift at the category-specific level, I calculated the KL divergence between category distributions in each domain, similar to the approach in [10]. These distributions are estimated from the percentage of pixels in each image belonging to each category. The domain classification results are shown in Table 1 and the KL divergence results are shown in Table 2.

4. Data

Cityscapes. The Cityscapes dataset[5] is a real-world dataset for urban scene segmentation. It contains street scenes from many different cities in Germany, all taken from a stereo camera setup mounted on a car. The dataset contains full-pixel labels for 5000 high resolution images with 2975 set aside for training, 500 set aside for validation, and 1525 set aside for testing. Each pixel is assigned to one of 34 categories, which include roads, pedestrians, traffic signs, and other object categories related to driving.

SYNTHIA. The SYNTHIA dataset[18] is a synthetic dataset generated from a virtual world simulator. It contains 9000 images with full pixel-level labels, using the same categories as Cityscapes. As the virtual world can simulate different weather,traffic, and driving conditions, it contains a more varied set of images than that of Cityscapes. However, the visual field is unconstrained, with many images not taken from the driver's point of view. An example of this is seen in Figure 3. The lack of a constrained visual field introduces difficulties when applying a pretrained network to Cityscapes data. For my experiments, I split the SYNTHIA data into 7000 images fro training, 1000 images for validation, and 1000 images for testing.

Figure 3. Top is an example of data from SYNTHIA with an unconstrained field of view. Bottom is a sample from GTA, which only contains images from the driver's point of view

GTA5 The GTA5 dataset[17] contains about 25000 fully labeled synthetic street scenes, all taken from the popular video game Grand Theft Auto V. At almost 5 times the size of Cityscapes, the dataset contains realistic street scenes solely from the driver's point of view. This dataset also contains the same set of labels as Cityscapes making it much easier to evaluate and adapt pretrained models. For these experiments, I set aside 22000 images for training, 1000 images for validation, and 2000 images for testing.

5. Results and Analysis

5.1. Domain Shift Results

Table 1. reports the global domain shift between the two synthetic domains and City, relying on three different measures for thoroughness. The consistency between these results not only suggests that domain classification accuracy is a good measure of global domain shift, but also that both

Experiment	GTA	SYNTHIA
Linear SVM	92.0%	96.5%
Polynomial Kernel SVM	94.5%	96.0%
Max Mean Discrepancy (MMD)	0.46	0.73

Table 1. **Global Domain Shift**.MMD is widely used metric to measure domain shift, but has no context in terms of upper and lower bounds. SVM accuracy provides a bounded measure of domain shift and exhibits the same ordering as MMD.

synthetic datasets are quite distant from Cityscapes. SYNTHIA shows greater global domain shift than GTA, a notion further confirmed by the baseline results for each domain in Table 3.. In terms of category-specific shift, the results in Table 2. suggests more of the same. For the most prevalant object categories ,like Road, Car and Building, SYNTHIA has a relatively higher KL divergence with Cityscapes. This means that these object categories occupy a much different percentage of each image in SYNTHIA than they do in Cityscapes. This distribution shift can be attributed to unconstrained visual field of the SYNTHIA dataset, whereas Cityscapes and GTA are limited to the first-person driver view. This explains why GTA shows lower KL divergence than SYNTHIA for large, prevalant object categories.

Despite the high category-level domain shift for prevalant object categories, smaller and more rare objects (i.e. Traffic Lights, Traffic Signs, Motorcycles,etc.) in SYNTHIA show relatively lower distribution shift than those in GTA. Since these smaller object categories occupy less of each image, their impact on global domain shift is marginal. However, variations in pose and object size between domains, for these difficult categories, can negatively impact segmentation performance during transfer. As a result, robust visual representations for these difficult categories is necessary, making SYNTHIA still an attractive option for pretraining despite high global domain shift. Subsection 5.2 sheds light on the interaction between global and category-specific domain in affecting transfer performance in FCNs.

5.2. Supervised Results

In line with domain shift results from the previous section, baseline FCN models trained on SYNTHIA and GTA perform poorly when applied to the Cityscapes domain. The GTA pretrained model (G) shows much better initial performance than the SYNTHIA model (S), mostly due the lower global domain shift between GTA and Cityscapes. Even with a very low amount of supervised target data (e.g. 75 images), performance surges for both pretrained models. A closer look at the results reveals that most of this performance boost comes from large object categories (i.e. Road, Sidewalk, Building, Car,etc.), which are prevalent in almost all real domain images. The performance improvement implies that these prevalent object categories are easier to learn and thus models can achieve very high accuracy with little supervision. More difficult categories like Traffic Light and Traffic Sign don't see the same improvements and require a lot more real domain labels to learn reasonable representations. Even with more target supervision, however, the model does not significantly improve in segmenting the difficult categorie, causing the overall performance improvement plateaus after a certain point. Even with 85% of Cityscapes training data (e.g. 2000 images), the dilation FCN model achieves performance on par with a model trained on all Cityscapes training data. Although fine-tuning has been shown to exceed the performance of models trained from scratch[9, 4], the Cityscapes baseline model (C) is fine-tuned from weights pretrained on PASCAL, resulting in more stable learning and higher performance.

One surprising result from these experiments is that a fine-tuned S network transfers much better than a fine-tuned G network, despite the former having much lower baseline performance. What makes this result particularly interesting is that it sheds light on the relationship between fine-tuning, global domain shift, and category-specific domain shift. The GTA dataset has lower global domain shift than SYNTHIA, which is largely impacted by the easy, prevalent object categories. With small amounts of real supervision, the dilation FCN quickly remedies the global domain shift and achieves very high accuracy for these easy categories. The category-specific shift results from Table 3 show that SYNTHIA provides better alignment with Cityscapes for tougher, more rare categories such as Traffic Lights and Motorcycles. Since performance on these difficult objects does not improve significantly with large amounts of target supervision, good baseline representations are necessary for successful transfer. If poor initial representations are learned, they can directly conflict with the representations learned from real world supervision. The conflicting nature of these representations is evidenced by the performance gap for difficult categories between fine-tuned SYNTHIA and GTA models. Fine-tuned SYNTHIA models quickly learn the easy categories from real world supervision and provide better initial representations for hard categories. Pretrained GTA models, however, learn conflicting representations for these difficult categories, leading to worse overall transfer performance.

5.3. Semi-Supervised Results

Though fine-tuning shows good results given enough target supervision, it does not leverage information from the raw,unlabeled data in the target domain. The benefit of us-

Dataset	Road	Sidewalk	Building	Wall	Fence	Pole	T. Light	T. Sign	Vegetation	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motorcycle	Bicycle
GTA	10.7	2.9	6.1	2.9	6.7	0.27	0.68	5.0	13.9	0.27	18.6	1.1	1.3	12.1	1.2	5.2	9.6	1.4	n/a
SYNTHIA	14.8	15.4	10.7	2.9	6.1	0.21	0.51	4.8	14.4	n/a	15.5	0.68	1.4	12.6	n/a	5.1	n/a	0.35	1

Table 2. **KL Divergence** measures the difference between two distributions, which in this case are the label distributions for each domain. The table shows KL divergence measures between the listed dataset and Cityscapes, with higher values implying greater category-specific shift. The values vary between classes as different objects occupy different proportions of the image.

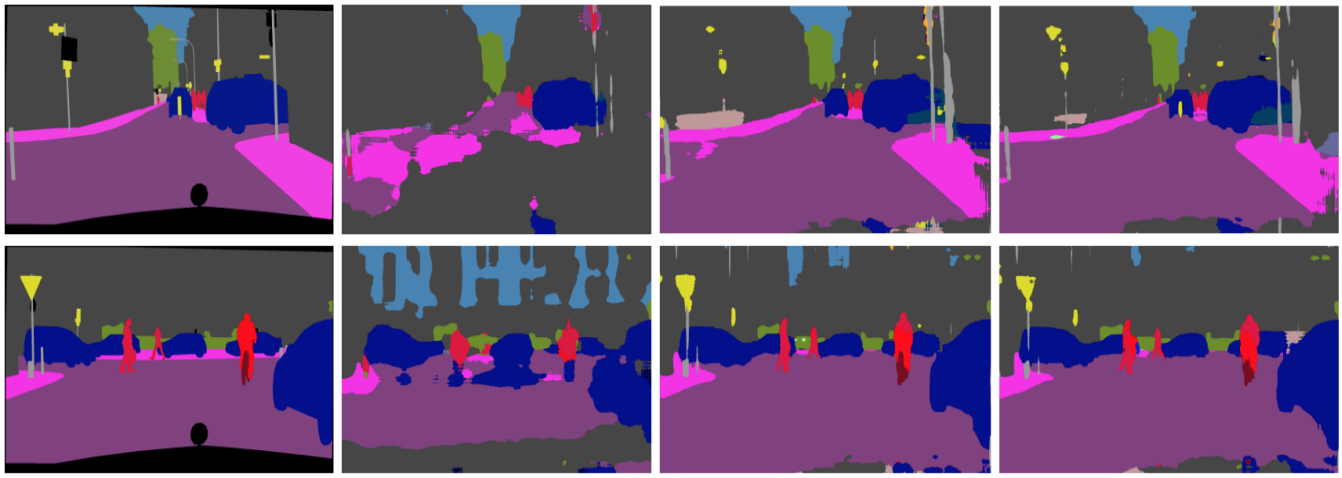

Figure 4. Results of using supervision to adapt a SYNTHIA pretrained model. From left to right: ground truth, S Baseline, S+C Low, S-DA Very Low

.

ing unsupervised image data is apparent from the results in Table 4. Even with only 50 labeled images from Cityscapes, the semi-supervised domain adversarial method achieves accuracy on par with networks fine-tuned with 1000 labeled images. This is a dramatic reduction in the number of labels needed for good performance, suggesting that domain adversarial learning provides a significant benefit when paired with target supervision. This performance improvement, however, plateaus as more data is added. As with fine-tuning, the more difficult categories are harder to localize and predict correctly, while roads and sidewalks are easily predicted with small amounts of Cityscapes data. The domain-adversarial training improves upon the baseline performance on difficult categories, but fails build on that improvement as more target supervision is added. Consequently, the mean IoU does not improve significantly between the experiments. As domain classification accuracy is a measure of global domain shift, the domain classification based, domain adversarial approach mitigates the effect of this global shift. To improve performance on difficult categories, more nuanced ways of addressing category-specific shift through network loss functions is necessary.

6. Conclusions and Future Work

In this work, I investigated the role of supervision from real driving domains in improving transfer for models pretrained on synthetic datasets. The results from my experiments show that current methods of supervised and semi-supervised transfer are able to achieve great performance with a small fraction of the labeled data needed to train FCNs. However, these transfer methods do show shortcomings, as the results show marginal benefit to adding more real domain supervision after a certain point. Categories that are easy to identify, such as roads and sidewalks, are quickly segmented with small amounts of real domain supervision. However, performance on difficult object categories (i.e. traffic lights, fences, rider,etc.), does not improve as more target data becomes available.

The semi-supervised domain adversarial method provides a step in the right direction, performing well on difficult categories with a very small amount of images. To scale this improvement with larger amounts of real supervision, future work may consider weighting the supervised loss by the category distribution. This would help rare and small object categories have as much influence on the predictions as large, prevalent categories. Furthermore, minimization of

Experiment	Road	Sidewalk	Building	Wall	Fence	Pole	T. Light	T. Sign	Vegetation	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motorcycle	Bicycle	**Mean**
C Baseline	96.3	74.3	88.2	39.6	44.0	48.0	47.2	60.8	90.0	55.3	90.9	70.6	38.8	89.6	41.9	58.4	29.1	43.3	65.3	**61.7**
G Baseline	49.1	23.9	58.2	12.8	6.0	16.0	10.4	4.0	77.5	12.0	67.6	36.3	2.1	69.8	13.9	12.6	0.6	8.3	n/a	**24.5**
G+C Very Low	92.6	56.9	81.8	22.6	26.3	24.8	27.1	32.4	85.9	42.1	84.1	48.6	16.9	80.9	15.6	8.0	19.0	18.4	48.9	**43.8**
G+C Low	94.4	64.4	84.5	25.9	34.8	31.6	32.0	39.5	87.1	48.1	88.3	58.5	21.3	83.8	19.2	34.7	20.9	25.4	53.9	**49.9**
G+C Medium	95.3	69.0	85.6	33.3	38.8	37.5	36.6	47.2	88.1	51.0	88.9	61.8	28.2	86.0	33.7	47.3	30.4	31.0	57.4	**55.1**
G+C High	95.4	69.3	85.8	32.2	39.8	37.0	36.2	47.0	88.3	52.1	88.4	62.0	26.6	86.0	33.4	42.7	29.1	29.3	57.3	**54.6**
S Baseline	7.4	15.5	31.7	1.2	0.0	9.9	0.1	0.8	47.3	n/a	66.5	46.5	2.9	54.1	n/a	4.5	n/a	0.0	1.0	**15.2**
S+C Very Low	92.1	53.7	82.3	17.4	27.8	38.8	29.3	39.9	86.4	40.6	84.0	59.8	24.0	81.2	12.5	9.7	17.4	20.8	55.0	**46.0**
S+C Low	95.2	68.4	86.0	24.7	31.4	45.2	44.0	55.9	89.3	49.6	88.9	68.6	34.5	86.8	18.2	48.9	27.9	36.1	63.3	**55.5**
S+C Medium	95.9	71.7	87.5	32.9	40.1	46.6	45.5	59.4	89.6	52.8	90.6	69.0	39.7	89.0	39.4	54.9	29.5	40.5	64.0	**59.0**
S+C High	96.2	74.0	88.0	31.7	45.1	49.4	49.5	62.8	89.9	54.4	90.9	70.9	41.0	89.4	37.0	51.1	29.4	45.5	66.5	**61.2**

Table 3. **Supervised Results**. The GTA (G) and SYNTHIA (S) Baseline models are evaluated directly on the Cityscapes (C) validation set. Then, some supervised data, denoted by Very Low (75 images), Low (450 images) , Medium (1000 images), and High (2500 images), from Cityscapes is used to fine-tune these baseline models.

Experiment	Road	Sidewalk	Building	Wall	Fence	Pole	T. Light	T. Sign	Vegetation	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motorcycle	Bicycle	**Mean**
C Baseline	96.3	74.3	88.2	39.6	44.0	48.0	47.2	60.8	90.0	55.3	90.9	70.6	38.8	89.6	41.9	58.4	29.1	43.3	65.3	**61.7**
G Baseline	49.1	23.9	58.2	12.8	6.0	16.0	10.4	4.0	77.5	12.0	67.6	36.3	2.1	69.8	13.9	12.6	0.6	8.3	n/a	**24.5**
G-DA Low	95.0	68.6	85.2	26.0	37.2	38.2	39.7	48.1	88.3	49.5	88.3	63.9	16.3	85.9	30.9	45.5	27.9	25.8	56.8	**53.5**
G-DA Medium	94.9	67.9	85.3	27.7	37.0	38.0	39.0	46.9	88.2	50.0	88.3	64.3	22.6	85.8	31.4	45.0	25.0	29.9	57.0	**53.9**
G-DA High	94.8	67.4	85.4	28.5	37.2	38.0	39.6	47.3	88.0	49.4	88.3	63.8	15.2	85.7	29.6	43.7	20.7	22.5	56.8	**54.4**
S Baseline	7.4	15.5	31.7	1.2	0.0	9.9	0.1	0.8	47.3	n/a	66.5	46.5	2.9	54.1	n/a	4.5	n/a	0.0	1.0	**15.2**
S-DA Low	95.6	71.4	85.8	28.2	37.3	39.6	33.7	55.5	88.4	50.0	88.3	67.6	29.5	86.6	20.2	39.5	26.7	31.8	63.6	**54.7**
S-DA Medium	95.5	71.3	85.7	29.1	37.8	39.6	32.7	55.6	89.2	49.6	87.6	67.9	29.6	86.4	19.4	41.6	26.2	32.5	63.8	**55.0**
S-DA Medium	95.5	71.0	85.8	29.2	37.9	39.4	32.4	55.2	89.8	49.6	88.0	67.7	29.9	86.3	19.3	42.0	24.5	32.3	64.2	**55.3**

Table 4. **Semi-Supervised Results**. Adapted GTA (G) and SYNTHIA (S) models are evaluated on the Cityscapes (C) val set. The amount of target supervision, denoted by Low (50 images), Medium (150 images) , High (450 images), used for strong segmentation loss varies. All data is used for unsupervised domain adversarial learning

the KL divergence between category distributions in each domain, may prove useful in addressing category-specific domain shift.

References

[1] Y. Bengio. Deep learning of representations for unsupervised and transfer learning. In *Unsupervised and Transfer Learning - Workshop held at ICML 2011, Bellevue, Washington, USA, July 2, 2011*, pages 17–36, 2012.

[2] G. J. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, xx(x):xx–xx, 2008.

[3] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *CoRR*, abs/1412.7062, 2014.

[4] B. Chu, V. Madhavan, O. Beijbom, J. Hoffman, and T. Darrell. Best practices for fine-tuning visual classifiers to new domains. *CoRR*, 2016.

[5] 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. *CoRR*, abs/1604.01685, 2016.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

[7] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle,

F. Laviolette, M. Marchand, and V. S. Lempitsky. Domain-adversarial training of neural networks. *CoRR*, abs/1505.07818, 2015.

[8] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *Int. J. Rob. Res.*, 32(11):1231–1237, Sept. 2013.

[9] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524, 2013.

[10] J. Hoffman, D. Wang, F. Yu, and T. Darrell. FCNs in the Wild: Pixel-level Adversarial and Constraint-based Adaptation. *ArXiv e-prints*, Dec. 2016.

[11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *CoRR*, abs/1408.5093, 2014.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[13] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *CoRR*, abs/1411.4038, 2014.

[14] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. In *Proceedings of the 21st International Jont Conference on Artifical Intelligence*, IJCAI'09, pages 1187–1192, San Francisco, CA, USA, 2009. Morgan Kaufmann Publishers Inc.

[15] D. Pathak, P. Krähenbühl, and T. Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *ICCV*, 2015.

[16] D. Pathak, E. Shelhamer, J. Long, and T. Darrell. Fully convolutional multi-class multiple instance learning. *CoRR*, abs/1412.7144, 2014.

[17] S. R. Richter, V. Vineet, S. Roth, and V. Koltun. Playing for data: Ground truth from computer games. *CoRR*, abs/1608.02192, 2016.

[18] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[19] G. Ros, S. Stent, P. F. Alcantarilla, and T. Watanabe. Training constrained deconvolutional networks for road scene semantic segmentation. *CoRR*, abs/1604.01545, 2016.

[20] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, abs/1312.6229, 2013.

[21] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[22] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. *CoRR*, abs/1510.02192, 2015.

[23] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell. Deep domain confusion: Maximizing for domain invariance. *CoRR*, abs/1412.3474, 2014.

[24] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? *CoRR*, abs/1411.1792, 2014.

[25] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *CoRR*, abs/1511.07122, 2015.

[26] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 487–495. Curran Associates, Inc., 2014.