AFRL-RI-RS-TR-2020-009

DATA-EFFICIENT LEARNING BY ADVERSARIAL DISCRIMINATIVE DOMAIN ADAPTATION

REGENTS OF THE UNIVERSITY OF CALIFORNIA

FEBRUARY 2020

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

AFRL-RI-RS-TR-2020-009 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
AMY WONG
Work Unit Manager

/ S /
JULIE BRICHACEK
Chief, Information Systems Division
Information Directorate

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
FEBRUARY 2020	FINAL TECHNICAL REPORT	AUG 2017 – AUG 2019

4. TITLE AND SUBTITLE

DATA-EFFICIENT LEARNING BY ADVERSARIAL DISCRIMINATIVE DOMAIN ADAPTATION

5a. CONTRACT NUMBER FA8750-17-2-0280

5b. GRANT NUMBER N/A

5c. PROGRAM ELEMENT NUMBER 62788F

6. AUTHOR(S)

Trevor Darrell

5d. PROJECT NUMBER S2DL

5e. TASK NUMBER UO

5f. WORK UNIT NUMBER CB

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

REGENTS OF THE UNIVERSITY OF CALIFORNIA
The University of California, Berkeley,
2150 SHATTUCK AVE RM 313, Berkeley, CA, 94704

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory/RISC
525 Brooks Road
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S) AFRL/RI

11. SPONSOR/MONITOR'S REPORT NUMBER AFRL-RI-RS-TR-2020-009

12. DISTRIBUTION AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The goal of this effort was to streamlined the end-to-end process of using machine learning for AF/DoD tasks. Focusing on domain adaptation, University of California at Berkeley developed a discriminatively-trained Cycle-Consistent Adversarial Domain Adaptation model, a Semantic Pixel-Level Adaptation Transform approach to detector adaptation that efficiently generates cross-domain image pairs, and an adaptation method that exploits the continuity between gradually varying domains by adapting in sequence from the source to the most similar target domain. The models can be applied in a variety of visual recognition and prediction settings. They show new state-of-the-art results across multiple adaptation tasks.

15. SUBJECT TERMS

Domain Adaptation, Machine Learning, Computer Vision

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**AMY WONG**
U	U	U	UU	23	**19b. TELEPHONE NUMBER** (*Include area code*) **N/A**

TABLE OF CONTENTS

Section		Page
1.0	SUMMARY	1
2.0	INTRODUCTION	1
3.0	METHODS, ASSUMPTIONS, AND PROCEDURES	4
4.0	RESULTS AND DISCUSSION	7
5.0	CONCLUSIONS	14
6.0	REFERENCES	15
LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS		19

1.0 Summary

Domain adaptation is critical for success in new, unseen environments. Adversarial adaptation models applied in feature spaces discover domain invariant representations, but are difficult to visualize and sometimes fail to capture pixel-level and low-level domain shifts. Recent work has shown that generative adversarial net- works combined with cycle-consistency constraints are surprisingly effective at mapping images between domains, even without the use of aligned image pairs. We propose a novel discriminatively-trained Cycle-Consistent Adversarial Domain Adaptation model (CyCADA). CyCADA adapts representations at both the pixel-level and feature-level, enforces cycle-consistency while leveraging a task loss, and does not require aligned pairs. Our model can be applied in a variety of visual recognition and prediction settings. We show new state-of-the-art results across multiple adaptation tasks, including digit classification and semantic segmentation of road scenes demonstrating transfer from synthetic to real world domains.

2.0 Introduction

Deep neural networks excel at learning from large amounts of data, but can be poor at generalizing learned knowledge to new datasets or environments. Even a slight departure from a network's training domain can cause it to make spurious predictions and significantly hurt its performance (Tzeng et al., 2017). The visual domain shift from non-photorealistic synthetic data to real images presents an even more significant challenge. While we would like to train models on large amounts of synthetic data such as data collected from graphics game engines, such models fail to generalize to real-world imagery. For example, a state-of-the-art semantic segmentation model trained on synthetic dashcam data fails to segment the road in real images, and its overall per-pixel label accuracy drops from 93% (if trained on real imagery) to 54% (if trained only on synthetic data).

Feature-level unsupervised domain adaptation methods address this problem by aligning the features extracted from the network across the source (e.g. synthetic) and target (e.g. real) domains, without any labeled target samples. Alignment typically involves minimizing some measure of distance between the source and target feature distributions, such as maximum mean discrepancy (Long & Wang, 2015), correlation distance (Sun & Saenko, 2016), or adversarial discriminator accuracy (Ganin & Lempitsky, 2015; Tzeng et al., 2017). This class of techniques suffers from two main limitations. First, aligning marginal distributions does not enforce any semantic consistency, e.g. target features of a car may be mapped to source features of a bicycle. Second, alignment at higher levels of a deep representation can fail to model aspects of low-level appearance variance which are crucial for the end visual task.

Generative pixel-level domain adaptation models perform similar distribution alignment—not in feature space but rather in raw pixel space—translating source data to the "style" of a target domain. Recent methods can learn to translate images given only unsupervised data from both domains (Bousmalis et al., 2017; Liu & Tuzel, 2016; Shrivastava et al., 2017). The results are visually compelling, but such image-space models have only been shown to work for small image sizes and limited domain shifts. A more recent approach (Bousmalis et al., 2017) was applied to larger (but still not high resolution) images, but in a visually controlled image for robotic applications. Furthermore, they also do not necessarily preserve content: while the translated image may "look" like it came from the right domain, crucial semantic information may be lost. For example, a model adapting from line-drawings to photos could learn to make a line-drawing of a cat look like a photo of a dog.

How can we encourage the model to preserve semantic information in the process of distribution alignment? In this paper, we explore a simple yet powerful idea: give an additional objective to the model to reconstruct the original data from the adapted version. Cycle-consistency was recently proposed in a cross-domain image generation Generative Adversarial Network (GAN) model, CycleGAN (Zhu et al., 2017), which showed transformative image-to-image generation results, but was agnostic to any particular task.

We propose Cycle-Consistent Adversarial Domain Adaptation (CyCADA), which adapts representations at both the pixel-level and feature-level while enforcing local and global structural consistency through pixel cycle-consistency and semantic losses. CyCADA unifies prior feature-level (Ganin & Lempitsky, 2015; Tzeng et al., 2017) and image-level (Liu & Tuzel, 2016; Bousmalis et al., 2017b; Shrivastava et al., 2017) adversarial domain adaptation methods together with cycle-consistent image-to-image translation techniques (Zhu et al., 2017) 1. It is applicable across a range of deep architectures and/or representation levels, and has several advantages over existing unsupervised domain adaptation methods. We use a reconstruction (cycle-consistency) loss to encourage the cross-domain transformation to preserve local structural information and a semantic loss to enforce semantic consistency.

We apply our CyCADA model to the task of digit recognition across domains and the task of semantic segmentation of urban scenes across domains. Experiments show that our model achieves state of the art results on digit adaptation, cross-season adaptation in synthetic data, and on the challenging synthetic-to-real scenario. In the latter case, it improves per-pixel accuracy from 54% to 82%, nearly closing the gap to the target-trained model.

Our experiments confirm that domain adaptation can benefit greatly from cycle-consistent pixel transformations, and that this is especially important for pixel-level semantic segmentation with contemporary FCN architectures. Further, we show that adaptation at both the pixel and representation level can offer complementary improvements with joint pixel-space and feature adaptation leading to the highest performing model for digit classification tasks.

The problem of visual domain adaptation was introduced along with a pairwise metric transform solution by Saenko et al. (2010) and was further popularized by the broad study of visual dataset bias (Torralba & Efros, 2011). Early deep adaptive works focused on feature space alignment through minimizing the distance between first or second order feature space statistics of the source and target (Tzeng et al., 2014; Long & Wang, 2015). These latent distribution alignment approaches were further improved through the use of domain adversarial objectives whereby a domain classifier is trained to distinguish between the source and target representations while the domain representation is learned so as to maximize the error of the domain classifier. The representation is optimized using the standard minimax objective (Ganin & Lempitsky, 2015), the symmetric confusion objective (Tzeng et al., 2015), or the inverted label objective (Tzeng et al., 2017). Each of these objectives is related to the literature on generative adversarial networks (Goodfellow et al., 2014) and follow-up work for improved training procedures for these networks (Salimans et al., 2016b; Arjovsky et al., 2017).

The feature-space adaptation methods described above focus on modifications to the discriminative representation space. In contrast, other recent methods have sought adaptation in the pixel-space using various generative approaches. One advantage of pixel-space adaptation, as we have shown, is that the result may be more human interpretable, since an image from one domain can now be visualized in a new domain. CoGANs (Liu & Tuzel, 2016) jointly learn a source and target representation through explicit weight sharing of certain layers while each source and target has a unique generative adversarial objective. Ghifary et al. (2016) uses an additional reconstruction objective in the target domain to encourage alignment in the unsupervised adaptation setting.

In contrast, another approach is to directly convert the target image into a source style image (or visa versa), largely based on Generative Adversarial Networks (GANs) (Goodfellow et al., 2014). Researchers have successfully applied GANs to various applications such as image generation (Denton et al., 2015; Radford et al., 2015; Zhao et al., 2016), image editing (Zhu et al., 2016) and feature learning (Salimans et al., 2016a; Donahue et al., 2017). Recent work (Isola et al., 2016; Sangkloy et al., 2016; Karacan et al., 2016) adopt conditional GANs (Mirza & Osindero, 2014) for these image-to-image translation problems (Isola et al., 2016), but they require input-output image pairs for training, which is in general not available in domain adaptation problems.

There also exist lines of work where such training pairs are not given. Yoo et al. (2016) learns a source to target encoder-decoder along with a generative adversarial objective on the reconstruction which is is applied for predicting the clothing people are wearing. The Domain Transfer Network (Taigman et al., 2017b) trains a generator to transform a source image into a target image by enforcing consistency in the embedding space. Shrivastava et al. (2017) instead uses an L1 reconstruction loss to force the generated target images to be similar to their original source images. This works well for limited domain shifts where the domains are similar in pixel-space, but can be too limiting for settings with larger domain shifts. Bousmalis et al. (2017b) use a content similarity loss to ensure the generated target image is similar to the original source image; however, this requires prior knowledge about which parts of the image stay the same across domains (e.g. foreground). Our method does not require pre-defining what content is shared between domains and instead simply translates images back to their original domains while ensuring that they remain identical to their original versions. BiGAN (Donahue et al., 2017) and ALI (Dumoulin et al., 2016) take an approach of simultaneously learning the transformations between the pixel and the latent space. More recently, Cycle-consistent Adversarial Networks (CycleGAN) (Zhu et al., 2017) produced compelling image translation results such as generating photorealistic images from impressionism paintings or transforming horses into zebras at high resolution using the cycle-consistency loss. This loss was simultaneously proposed by Yi et al. (2017) and Kim et al. (2017) to great effect as well. Our motivation comes from such findings about the effectiveness of the cycle-consistency loss.

Few works have explicitly studied visual domain adaptation for the semantic segmentation task. Adaptation across weather conditions in simple road scenes was first studied by Levinkov & Fritz (2013). More recently, a convolutional domain adversarial based approached was proposed for more general drive cam scenes and for adaptation from simulated to real environments (Hoffman et al., 2016). Ros et al. (2016b) learns a multi-source model through concatenating all available labeled data and learning a single large model and then transfers to a sparsely labeled target domain through distillation (Hinton et al., 2015). Chen et al. (2017) use an adversarial objective to align both global and class-specific statistics, while mining additional temporal data from street view datasets to learn a static object prior. Zhang et al. (2017) instead perform segmentation adaptation by aligning label distributions both globally and across superpixels in an image.

3.0 Methods, Assumptions, Procedures

We consider the problem of unsupervised adaptation, where we are provided source data X_S, source labels Y_S, and target data X_T, but no target labels. The goal is to learn a model f that can correctly predict the label for the target data X_T.

We can begin by simply learning a source model f_S that can perform the task on the source data. However, while the learned model f_S will perform well on the source data, typically domain shift between the source and target domain leads to reduced performance when evaluating on target data. To mitigate the effects of domain shift, we follow previous adversarial adaptation approaches and learn to map samples across domains such that an adversarial discriminator is unable to distinguish the domains. By mapping samples into a common space, we enable our model to learn on source data while still generalizing to target data.

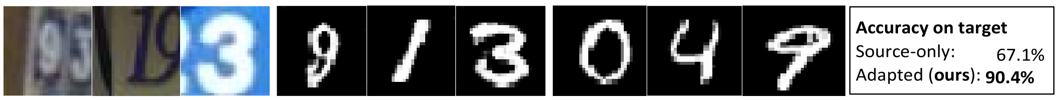
To this end, we introduce a mapping from source to target $G_{S\to T}$ and train it to produce target samples that fool an adversarial discriminator D_T. Conversely

, the adversarial discriminator attempts to classify the real target data from the source target data.

This objective ensures that $G_{S\to T}$, given source samples, produces convincing target samples. In turn, this ability to directly map samples between domains allows us to learn a target model f_T by minimizing $L_{\text{task}}(f_T , G_{S\to T} (X_S), Y_S)$ (see Figure 2 green portion).

However, while previous approaches that optimized similar objectives have shown effective results, in practice they can often be unstable and prone to failure. Although the GAN loss portion ensures that $G_{S\to T} (x_s)$ for some x_s will resemble data drawn from X_T, there is no way to guarantee that $G_{S\to T} (x_s)$ preserves the structure or content of the original sample x_s.

Source image (GTA5) Adapted source image (**Ours**) Target image (CityScapes)



Source images (SVHN) Adapted source images (**Ours**) Target images (MNIST)

Figure 1: We propose CyCADA, an adversarial unsupervised adaptation algorithm which uses cycle and semantic consistency to perform adaptation at multiple levels in a deep network. Our model provides significant performance improvements over source model baselines.

In order to encourage the source content to be preserved during the conversion process, we impose a cycle-consistency constraint on our adaptation method (Zhu et al., 2017; Yi et al., 2017; Kim et al., 2017) (see Figure 2 red portion). To this end, we introduce another mapping from target to source $G_{T\to S}$ and train it according to the same GAN loss $\mathrm{L}_{\mathrm{GAN}}(G_{T\to S}, D_S, X_S, X_T)$. We then require that mapping a source sample from source to target and back to the source reproduces the original sample, thereby enforcing cycle-consistency. In other words, we want $G_{T\to S}(G_{S\to T}(x_s))\ x_s$ and $G_{S\to T}(G_{T\to S}(x_t))\ x_t$. This is done by imposing an L1 penalty on the reconstruction error, which is referred to as the *cycle-consistency loss.*

Additionally, as we have access to source labeled data, we explicitly encourage high semantic consistency before and after image translation. We pretrain a source task model f_S, fixing the weights, we use this model as a noisy labeler by which we encourage an image to be classified in the same way after translation as it was before translation according to this classifier. Let us define the predicted label from a fixed classifier, f, for a given input X as $p(f, X) = \arg\max(f(X))$. Then we can define the semantic consistency before and after image translation.

We have thus far described an adaptation method which combines cycle consistency, semantic consistency, and adversarial objectives to produce a final target model. As a pixel-level method, the adversarial objective consists of a discriminator which distinguishes between two image sets, e.g. transformed source and real target image. Note that we could also consider a feature-level method which discriminates between the features or semantics from two image sets as viewed under a task network. This would amount to an additional feature level GAN loss.

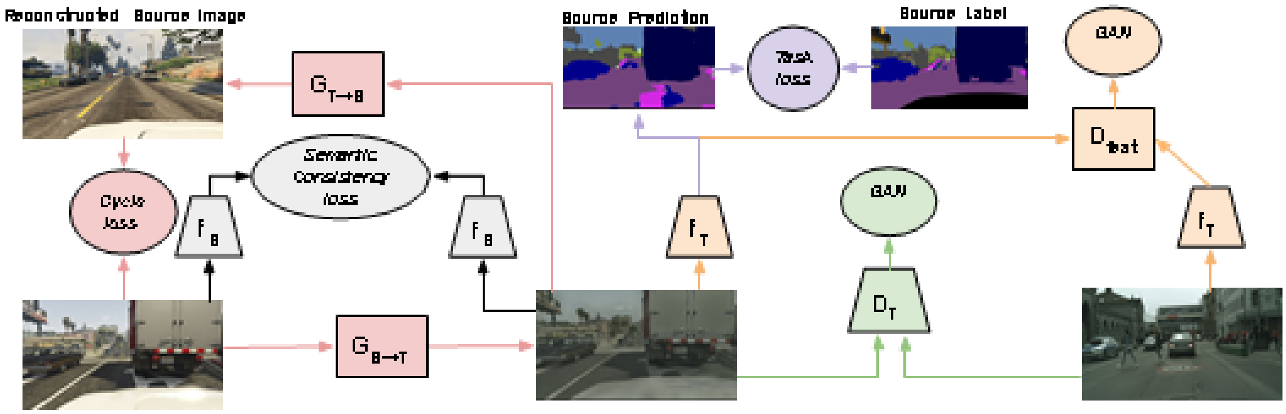


Figure 2: Cycle-consistent adversarial adaptation of pixel-space inputs. By directly remapping source training data into the target domain, we remove the low-level differences between the domains, ensuring that our task model is well-conditioned on target data. We depict here the image-level GAN loss (green), the feature level GAN loss (orange), the source and target semantic consistency losses (black), the source cycle loss (red), and the source task loss (purple). For clarity the target cycle is omitted.

Taken together, these loss functions form our complete objective. We have introduced a method for unsupervised adaptation which generalizes adversarial objectives to be viewed as operating at the pixel or feature level. In addition, we introduce the use of cycle- consistency together with semantic transformation constraints to guide the mapping from one domain to another. In this work, we apply CyCADA to both digit adaptation and to semantic segmentation. We implement G as a pixel-to-pixel convnet, f as a convnet classifier or a Fully-Convolutional Net (FCN) and D as a convnet with binary outputs.

We further proposed a method called Semantic Pixel-Level Adaptation Transforms (SPLAT) that performs pixel transformations from source to target to tackle the object detection adaptation problem. SPLAT is flexible and can optionally make use of additional label information such as segmentation labels when such data is available. In turn, the ability to make use of this extra information leads to a more efficient cycle-free pixel adaptation method that both runs faster than existing approaches and allows us to learn detection models that are more accurate in the target domain.

Existing pixel adaptation methods have demonstrated that it is possible to directly transform source images into target images in order to learn task models that are effective in the target domain. However, their applications were limited primarily to the realm of classification and segmentation. We show for the first time that these methods can be applied to detection domain adaptation, producing a cross-domain aligned set of images. We demonstrate that pixel-adaptation methods can be used to learn target domain object detectors that are robust to the negative effects of domain shift.

Our detection adaptation method is agnostic to the specific pixel transformation model that is used. In practice, performing detection with existing pixel-level adaptation methods such as CyCADA yields surprisingly strong performance. However, these models can be quite difficult to train and do not make use of the semantic labels present in adaptation tasks. Thus, as an additional contribution, we introduce a novel, lightweight pixel-level adaptation method that eliminates the cumbersome cycle loss in favor of a loss that incorporates label information.

In continuous adaptation, we are presented with a source domain *S*, and multiple target domains T; that represent continuous shifts of *S* at time *i*. We assume access to source images A, and labels V, drawn from a source domain distribution *p* (z, p), as well as target images At, drawn from target distributions (p), with no labeled observations. We additionally assume that the source domain is similar to the target domain at time tb, that the target domain is smoothly varying, and that *p* is more similar to p, than *pz* is to p. Since direct supervised learning on the target domains is not possible, continuous adaptation instead learns a source representation mapping, *M*, and a source classifier, C, and then adapts that model for use in the stream of target domains. We present a general framework for continuous adaptation with replay, where we evolve the model to the new distribution while simultaneously guiding it to not deviate too far from how it previously performed on prior distributions.

We introduce an adaptation model, Continuous Unsupervised Adaptation (CUA), which progressively evolves to correctly classify multiple shifted domains. Standard unsupervised adaptation effectively adapts between a single source distribution *p*(z, p) and a single target distribution *pz(:s, y)* by aligning features from both domains. In other words, they learn the source and target mappings, *M* and *M*, so as to minimize the distance between the empirical source and target mapping distributions

4.0 Results and Discussion

Without dense labels, as is the case when only detection labels are available in the source, transformations are learned using CycleGAN alignment. Otherwise, when dense labels are available we introduce a more efficient cycle-free method, which exploits pixel-level semantic labels to condition the training of the transformation network. The end task is then trained using detection box labels from the source, potentially including labels inferred on unlabeled source data.

We show new state-of-the-art results across multiple adaptation tasks, including digit classification and semantic segmentation of road scenes demonstrating transfer from synthetic to real world domains.

We show both that pixel-level transforms outperform prior approaches to detector domain adaptation, and that our cycle-free method outperforms prior models for unconstrained cycle-based learning of generic transformations while running 3.8 times faster. Our combined model improves on prior detection baselines by 12.5 mean average precision (mAP) adapting from Sim

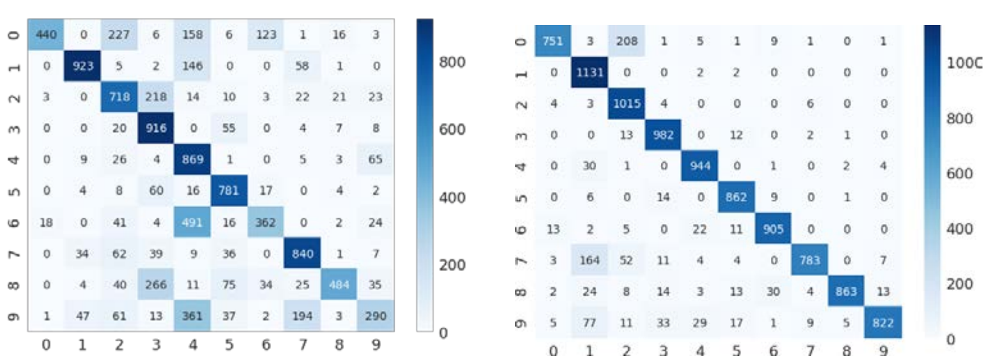
10K to Cityscapes, recovering over 50% of the missing performance between the unadapted baseline and the labeled-target upper bound.

Table 1: Unsupervise d domain adaptation across digit datasets. Our model is competitive with or outperforms state-of-the-art models for each shift. For the difficult shift of SVHN to MNIST we also note that feature space adaptation provides additional benefit beyond the pixel-only adaptation

Model	MNIST → USPS	USPS → MNIST	SVHN → MNIST
Source only	82.2 ± 0.8	69.6 ± 3.8	67.1 ± 0.6
DANN (Ganin et al., 2016)	-	-	73.6
DTN (Taigman et al., 2017a)	-	-	84.4
CoGAN (Liu & Tuzel, 2016a)	91.2	89.1	-
ADDA (Tzeng et al., 2017)	89.4 ± 0.2	90.1 ± 0.8	76.0 ± 1.8
CyCADA pixel only	**95.6 ± 0.2**	96.4 ± 0.1	70.3 ± 0.2
CyCADA pixel+feat	**95.6 ± 0.2**	**96.5 ± 0.1**	**90.4 ± 0.4**
Target only	96.3 ± 0.1	99.2 ± 0.1	99.2 ± 0.1

Our detailed evaluation is as follows. We evaluate CyCADA on several unsupervised adaptation scenarios. We first focus on adaptation for digit classification using the MNIST (LeCun et al., 1998), USPS, and Street View House Numbers (SVHN) (Netzer et al., 2011) datasets. After which we present results for the task of semantic image segmentation, using the SYNTHetic collection of Imagery and Annotations (SYNTHIA) (Ros et al., 2016a), GTA (Richter et al., 2016) and CityScapes (Cordts et al., 2016) datasets.

We evaluate our method across the adaptation shifts of USPS to MNIST, MNIST to USPS, and SVHN to MNIST, using the full training sets during learning phases and evaluating on the standard test sets. We report classification accuracy for each shift in Table 1 and find that our method outperforms competing approaches on average. The classifier for our method for all digit shifts uses a variant of the LeNet architecture. Note that the recent pixel-da method by Bousmalis et al. (2017b) presents results for only the MNIST to USPS shift and reports 95.9% accuracy, while our method achieves 95.6% accuracy. However, the pixel-da approach cross validates with some labeled data which is not an equivalent evaluation setting.



(a) Source only Model (b) CyCADA model

Figure 3: Confusion matricies for SVHN → MNIST experiment.

Ablation: Pixel vs Feature Level Transfer. We begin by evaluating the contribution of the pixel space and feature space transfer. We find that in the case of the small domain shifts between USPS and MNIST, the pixel space adaptation by which we train a classifier using images translated using CycleGAN (Zhu et al., 2017), performs very well, outperforming or comparable to prior adaptation approaches. Feature level adaptation offers a small benefit in this case of a small pixel shift. However, for the more difficult shift of SVHN to MNIST, we find that feature level adaptation outperforms the pixel level adaptation, and importantly, both may be combined to produce an overall model which outperforms all competing methods.

Ablation: No Semantic Consistency. We experiment without the addition of our semantic consistency loss and find that the standard unsupervised CycleGAN approach diverged when training SVHN to MNIST often suffering from random label flipping. Cycle constraints alone fail to have the desired behavior for our end task. An SVHN image is mapped to a convincing MNIST type image and back to a SVHN image with correct semantics. However, the MNIST-like image has mismatched semantics. Our modified version, which uses the source labels to train a weak classification model which can be used to enforce semantic consistency before and after translation, resolves this issue and produces strong performance.

Ablation: No Cycle Consistency. We study the result when learning without the cycle consistency loss. First note that there is no reconstruction guarantee in this case, thus we see that the translation back to SVHN fails. In addition, we find that while the semantic loss does encourage correct semantics it relies on the weak source labeler and thus label flipping still occurs.

SEMANTIC SEGMENTATION ADAPTATION: The task is to assign a semantic label to each pixel in the input image, e.g. *road*, *building*, etc. We limit our evaluation to the unsupervised adaptation setting, where labels are only available in the source domain, but we are evaluated solely on our performance in the target domain.

For each experiment, we use three metrics to evaluate performance. Let n_{ij} be the number of pixels of class i predicted as class j, let $t_i = \sum_j n_{ij}$ be the total number of pixels of class i, and let N be the number of classes. Our three evaluation metrics are, mean intersection-over-union (mIoU), frequency weighted intersection-over-union (fwIoU), and pixel accuracy.

Cycle-consistent adversarial adaptation is general and can be applied at any layer of a network. Since optimizing the full CyCADA objective end-to-end is memory-intensive in practice, we train our model in stages. First, we perform image-space adaptation and map our source data into the target domain. Next, using the adapted source data with the original source labels, we learn a task model that is suited to operating on target data. Finally, we perform another round of adaptation between the adapted source data and the target data in feature-space, using one of the intermediate layers of the task model. Additionally, we do not use the semantic loss for the segmentation experiments as it would require loading generators, discriminators, and an additional semantic segmenter into memory all at once for two images.

GTA5 → Cityscapes

Table 2: Adaptation between GTA5 and Cityscapes, showing IoU for each class and mean IoU, freq- weighted IoU and pixel accuracy. CyCADA significantly outperforms baselines, nearly closing the gap to the target-trained oracle on pixel accuracy. FCNs in the wild is by Hoffman et al. (2016). We compare our model using two base semantic segmentation architectures (A) VGG16-FCN8s (Long et al., 2015) base network and (B) DRN-26 (Yu et al., 2017).

Source only	A	26.0	14.9	65.1	5.5	12.9	8.9	6.0	2.5	70.0	2.9	47.0	24.5	0.0	40.0	12.1	1.5	0.0	0.0	0.0	17.9	41.9	54.0
FCNs in the wild*	A	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	—	—
CyCADA feat-only	A	**85.6**	30.7	74.7	14.4	13.0	17.6	13.7	5.8	74.6	15.8	**69.9**	38.2	3.5	72.3	16.0	5.0	0.1	3.6	0.0	29.2	71.5	82.5
CyCADA pixel-only	A	83.5	**38.3**	**76.4**	20.6	**16.5**	22.2	**26.2**	**21.9**	**80.4**	28.7	65.7	49.4	4.2	74.6	16.0	26.6	2.0	8.0	0.0	34.8	73.1	82.8
CyCADA pixel+feat	A	85.2	37.2	**76.5**	**21.8**	15.0	**23.8**	22.9	21.5	**80.5**	**31.3**	60.7	**50.5**	**9.0**	**76.9**	**17.1**	**28.2**	**4.5**	**9.8**	0.0	**35.4**	**73.8**	**83.6**
Oracle - Target Super	A	96.4	74.5	87.1	35.3	37.8	36.4	46.9	60.1	89.0	54.3	89.8	65.6	35.9	89.4	38.6	64.1	38.6	40.5	65.1	60.3	87.6	93.1
Source only	B	42.7	26.3	51.7	5.5	6.8	13.8	23.6	6.9	75.5	11.5	36.8	49.3	0.9	46.7	3.4	5.0	0.0	5.0	1.4	21.7	47.4	62.5
CyCADA feat-only	B	78.1	31.1	71.2	10.3	14.1	29.8	28.1	20.9	74.0	16.8	51.9	53.6	6.1	65.4	8.2	20.9	1.8	13.9	5.9	31.7	67.4	78.4
CyCADA pixel-only	B	63.7	24.7	69.3	21.2	17.0	30.3	33.0	**32.0**	80.5	25.3	62.3	62.0	**15.1**	73.1	19.8	23.6	5.5	16.2	**28.7**	37.0	63.8	75.4
CyCADA pixel+feat	B	**79.1**	**33.1**	**77.9**	**23.4**	**17.3**	**32.1**	**33.3**	31.8	**81.5**	**26.7**	**69.0**	**62.8**	14.7	**74.5**	**20.9**	**25.6**	**6.9**	**18.8**	20.4	**39.5**	**72.4**	**82.3**
Oracle - Target Super	B	97.3	79.8	88.6	32.5	48.2	56.3	63.6	73.3	89.0	58.9	93.0	78.2	55.2	92.2	45.0	67.3	39.6	49.9	73.6	67.4	89.6	94.3

For our first evaluation, we consider the SYNTHIA dataset (Ros et al., 2016a), which contains synthetic renderings of urban scenes. We use the SYNTHIA video sequences, which are rendered across a variety of environments, weather conditions, and lighting conditions. This provides a synthetic testbed for evaluating adaptation techniques. For comparison with previous work, in this work we focus on adaptation between seasons. We use only the front-facing views in the sequences so as to mimic dashcam imagery, and adapt from fall to winter. The subset of the dataset we use contains 13 classes and consists of 10,852 fall images and 7,654 winter images.

To further demonstrate our method's applicability to real-world adaptation scenarios, we also evaluate our model in a challenging synthetic-to-real adaptation setting. For our synthetic source domain, we use the GTA5 dataset (Richter et al., 2016) extracted from the game Grand Theft Auto V, which contains 24966 images. We consider adaptation from GTA5 to the real-world Cityscapes dataset (Cordts et al., 2016), from which we used 19998 images without annotation for training and 500 images for validation. Both of these datasets are evaluated on the same set of 19 classes, allowing for straightforward adaptation between the two domains.

Image-space adaptation also affords us the ability to visually inspect the results of the adaptation method. This is a distinct advantage over opaque feature-space adaptation methods, especially in truly unsupervised settings—without labels, there is no way to empirically evaluate the adapted model, and thus no way to verify that adaptation is improving task performance. Visually confirming that the conversions between source and target images are reasonable, while not a *guarantee* of improved task performance, can serve as a sanity check to ensure that adaptation is not completely diverging.

We start by exploring the abilities of pixel space adaptation alone (using FCN8s architecture) for the setting of adapting across seasons in synthetic data. For this we use the SYNTHIA dataset and adapt from fall to winter weather conditions. Typically in unsupervised adaptation settings it is difficult to interpret what causes the performance improvement after adaptation. Therefore, we use this setting as an example where we may directly visualize the shift from fall to winter and inspect the intermediate pixel level adaptation result from our algorithm. This visually interpretable result matches our expectation of the true shift between these domains and indeed results in favorable final semantic segmentation performance from fall to winter. We find that CyCADA achieves state-of-the-art performance on this task with image space adaptation alone, however does not recover full supervised learning performance (train on target). Some example errors includes adding snow to the sidewalks, but not to the road, while in the true winter domain snow appears in both locations. However, even this mistake is interesting as it implies that the model is learning to distinguish road from sidewalk during pixel adaptation, despite the lack of pixel annotations.

Cycle-consistent adversarial adaptation achieves state-of-the-art adaptation performance. We see that under the fwIoU and pixel accuracy metrics, CyCADA approaches oracle performance, falling short by only a few points, despite being entirely unsupervised. This indicates that CyCADA is extremely effective at correcting the most common classes in the dataset.

To evaluate our method's applicability to real-world adaptation settings, we investigate adaptation from synthetic to real-world imagery. Once again, CyCADA achieves state-of-the-art results, recovering approximately 40% of the performance lost to domain shift. CyCADA also improves or maintains performance on all 19 classes. Examination of fwIoU and pixel accuracy as well as individual class intersection-over-union (IoUs) reveals that our method performs well on most of the common classes. Although some classes such as *train* and *bicycle* see little or no improvement, we note that those classes are poorly represented in the GTA5 data, making recognition very difficult. We compare our model against Shrivastava et al. (2017) for this setting, but found this approach did not converge and resulted in worse performance than the source only model.

We visualized the results of image-space adaptation between GTA5 and Cityscapes. The most obvious difference between the original images and the adapted images is the saturation levels—the GTA5 imagery is much more vivid than the Cityscapes imagery, so adaptation adjusts the colors to compensate. We also observe texture changes, which are perhaps most apparent in the road: in-game, the roads appear rough with many blemishes, but Cityscapes roads tend to be fairly uniform in appearance, so in converting from GTA5 to Cityscapes, our model removes most of the texture. Somewhat amusingly, our model has a tendency to add a hood ornament to the bottom of the image, which, while likely irrelevant to the segmentation task, serves as a further indication that image-space adaptation is producing reasonable results.

Details of our implementation are as follows. We begin by pretraining the source task model, f_S , using the task loss on the labeled source data. Next, we perform pixel-level adaptation using our image space GAN losses together with semantic consistency and cycle consistency losses. This yeilds learned parameters for the image transformations, $G_{S\to T}$ and $G_{T\to S}$, image discriminators, D_S and D_T, as well as an initial setting of the task model, f_T, which is trained using pixel transformed source images and the corresponding source pixel labels. Finally, we perform feature space adaptation in order to update the target semantic model, f_T, to have features which are aligned between the source images mapped into target style and the real target images. During this phase, we learn the feature discriminator, D_{feat} and use this to guide the representation update to f_T. In general, our method could also perform phases 2 and 3 simultaneously, but this would require more GPU memory then available at the time of these experiments.

For all feature space adaptation we equally weight the generator and discriminator losses. We only update the generator when the discriminator accuracy is above 60% over the last batch (digits) or last 100 iterations (semantic segmentation) – this reduces the potential for volatile training. If after an epoch (entire pass over dataset) no suitable discriminator is found, the feature adaptation stops, otherwise it continues until max iterations are reached.

For all digit experiments we use a variant of the LeNet architecture as the task net. Our feature discriminator network consists of 3 fully connected layers. The image discriminator network consists of 6 convolutional layers culminating in a single value per pixel. Finally, to generate one image domain from another we use a multilayer network which consists of convolution layers followed by two residual blocks and then deconvolution layers. All stages are trained using the Adam optimizer.

We experiment with both the VGG16-FCN8s Long et al. (2015) architecture as well as the DRN-26 Yu et al. (2017) architecture. For FCN8s, we train our source semantic segmentation model for 100k iterations using stochastic gradient descent (SGD) with learning rate $1e^{-3}$ and momentum 0.9. For the DRN-26 architecture, we train our source semantic segmentation model for 115K iterations using SGD with learning rate $1e^{-3}$ and momentum 0.9. We use a crop size of 600x600 and a batch size of 8 for this training. For cycle-consistent image level adaptation, we followed the network architecture and hyperparameters of CycleGAN(Zhu et al., 2017). All images were resized to have width of 1024 pixels while keeping the aspect ratio, and the training was performed with randomly cropped patches of size 400 by 400. Also, due to large size of the dataset, we trained only 20 epochs. For feature level adaptation, we train using SGD with momentum, 0.99, and learning rate 1e-5. We weight the representation loss ten times less than the discriminator loss as a convenience since otherwise the discriminator did not learn a suitable model within a single epoch. Then the segmentation model was trained separately using the adapted source images and the ground truth labels of the source data. Due to memory limitations we can only include a single source and single target image at a time (crops of size 768x768), this small batch is one of the main reasons for using a high momentum parameter.

We illustrate the performance of a recent pixel level adaptation approach proposed by Shrivastava et al. (2017) on our semantic segmentation data – GTA to Cityscapes. These images are significantly larger and more complex than those shown in the experiments in the original paper. We show image to image translation results under three different settings of the model hyperparameter, λ, which controls the tradeoff between the reconstruction loss and the visual style loss. When $\lambda = 10$, the resulting image converges to a near replica of the original image, thus preserving content but lacking the correct target style. When $\lambda = 1$ or $\lambda = 2.5$, the results lack any consistent semantics making it difficult to perceive the style of the transformed image. Thus, the resulting performance for this model is 11.6 mIoU for FCN8s with VGG, well below the performance of the corresponding source model of 17.9 mIoU.

To understand the types of mistakes which are improved upon and those which still persist after adaptation, we present the confusion matrices before and after our approach for the digit experiment of SVHN to MNIST. Before adaptation we see common confusions are 0s with 2s, 4s, and 7s. 6 with 4, 8 with 3, and 9 with 4. After adaptation all errors are reduced, but we still find that 7s are confused with 1s and 0s with 2s. These errors make some sense as with hand written digits, these digits sometimes resemble one another. It remains an open question to produce a model which may overcome these types of errors between highly similar classes.

We test detection adaptation performance from synthetic to real driving scenarios. Our source domain consists of Grand Theft Auto V imagery and our target domain is real-world driving data captured from dashboard mounted cameras. Consistent with prior work, we train our source detection model on *Sim10k*, a synthetic detection dataset generated from Grand Theft Auto V, and test on *Cityscapes*. *Sim10k* contains 10,000 images and 58,701 bounding boxes, all of which are used to train our adaptation model. There are no bounding box annotations *Cityscapes*, so we construct tight bounding boxes around instance-level segmentations. We treat the 2975 training images as our target dataset and evaluate on the 500 validation images. We only test detection results for car.

Our model is amenable to training with additional source data that is either unlabeled or has labels for a different task. In our experiments, we take advantage of an additional in-domain source dataset with segmentation labels to improve the speed of our pixel-adaptation model. We use a distinct set of images generated from Grand Theft Auto V and annotated with semantic segmentation labels. This is the *GTA5* dataset. It contains 24966 images that we use to train the pixel adaptation component of SPLAT-Lite. In our pixel adaptation analysis, we describe the training de- tails of SPLAT-lite and show that, by compromising training speed, we can achieve comparable accuracy without additional source data.

We separately investigate the effects of the pseudo-label and aligned pair losses. Our results show that, for the detection adaptation task, training on the pseudo-labeled dataset of source adapted images strictly outperforms training with a feature loss on aligned pairs. We first examine these losses independently, then explore combining both. All of our experiments are evaluated the validation set of *Cityscapes*.

When using these losses independently, we find that the pseudo-label loss outperforms the aligned pair loss by a significant margin: pseudo-labeling achieves 51.5 mAP versus the aligned pair loss, which achieves 43.3. Somewhat surprisingly, combining the two losses appears to underperform the pseudo-label loss alone, achieving a mAP of only 47.5. We hypothesize that directly matching the features between source and target is too restrictive and prevents the model from properly learning features that work well in the target domain, and report final results using only the pseudo-label loss.

We evaluate CUA for unsupervised classification adaptation to continuously shifting domains. For our continuous shifts, we consider the setting of MNIST digits being gradually rotated. This setting causes traditional unsupervised adaptation methods to fail when attempting to adapt to all variations together. We compare our model CUA against multiple state-of-the-art unsupervised adaptation methods that perform adaptation to a batch of target domains. Our method significantly outperforms the competing approaches and nearly reaches fully supervised performance.

5.0 Conclusion

We presented a cycle-consistent adversarial domain adaptation method that unifies cycle-consistent adversarial models with adversarial adaptation methods. CyCADA is able to adapt even in the absence of target labels and is broadly applicable at both the pixel-level and in feature space. An image-space adaptation instantiation of CyCADA also provides additional interpretability and serves as a useful way to verify successful adaptation. Finally, we experimentally validated our model on a variety of adaptation tasks: state-of-the-art results in multiple evaluation settings indicate its effectiveness, even on challenging synthetic-to-real tasks.

Our proposed method SPLAT is a novel approach to detector adaptation that utilizes pixel-level transforms to adapt from source to target domains. SPLAT is a flexible model; it works on unlabeled target data, and when target labels are present, can additionally condition on them with a cycle- free loss to operate more accurately and efficiently. Our model is also able to make use of unlabeled source data by inferring additional label information, thereby increasing the amount of training data available to learned task networks. By incorporating these improvements, our final model adapts images 3.8 times faster than previous cycle-based methods while improving 12.5 mAP over previous methods on a challenging detection adaptation setting. We further developed a novel scheme for continuous domain adaptation.

6.0 References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *CoRR*, abs/1701.07875, 2017. URL http://arxiv.org/abs/1701.07875.

Konstantinos Bousmalis, Alex Irpan, Paul Wohlhart, Yunfei Bai, Matthew Kelcey, Mrinal Kalakr- ishnan, Laura Downs, Julian Ibarz, Peter Pastor, Kurt Konolige, Sergey Levine, and Vincent Vanhoucke. Using simulation and domain adaptation to improve efficiency of deep robotic grasping. *CoRR*, abs/1709.07857, 2017a. URL http://arxiv.org/abs/1709.07857.

Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017b.

Yi-Hsin Chen, Wei-Yu Chen, Yu-Ting Chen, Bo-Cheng Tsai, Yu-Chiang Frank Wang, and Min Sun. No more discrimination: Cross city adaptation of road scene segmenters. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

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. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Emily L Denton, Soumith Chintala, Rob Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *Neural Information Processing Systems (NIPS)*, pp. 1486–1494, 2015.

Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. In *ICLR*, 2017.

Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Alex Lamb, Martin Arjovsky, Olivier Mastropietro, and Aaron Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.

Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 2016.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In David Blei and Francis Bach (eds.), *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pp. 1180–1189. JMLR Workshop and Conference Proceedings, 2015. URL http://jmlr.org/proceedings/papers/v37/ganin15.pdf.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 2414–2423, 2016.

Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In

European Confer- ence on Computer Vision (ECCV), pp. 597–613. Springer, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neu- ral Information Processing Systems 27*. 2014. URL http://papers.nips.cc/paper/ 5423-generative-adversarial-nets.pdf.

G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. 2015.

Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. FCNs in the wild: Pixel-level adversarial and constraint-based adaptation. *CoRR*, abs/1612.02649, 2016. URL http://arxiv.org/ abs/1612.02649.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016.

Levent Karacan, Zeynep Akata, Aykut Erdem, and Erkut Erdem. Learning to generate images of outdoor scenes from attributes and semantic layouts. *arXiv preprint arXiv:1612.00215*, 2016.

Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, 2017.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998.

Evgeny Levinkov and Mario Fritz. Sequential bayesian model update under structured scene prior for semantic road scenes labeling. In *IEEE International Conference on Computer Vision (ICCV)*, 2013. URL http://scalable.mpi-inf.mpg.de/files/2013/10/levinkov13iccv. pdfhttp://www.d2.mpi-inf.mpg.de/sequential-bayesian-update.

M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. In *Advances in Neural Information Processing Systems*, 2016a.

Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *Neural Information Processing Systems (NIPS)*, 2016b.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *CVPR (to appear)*, November 2015.

Mingsheng Long and Jianmin Wang. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning (ICML)*, 2015.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014. URL http://arxiv.org/abs/1411.1784.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. URL http://ufldl.stanford.edu/ housenumbers/nips2011_housenumbers.pdf.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint* *arXiv:1511.06434*, 2015.

Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *European Conference on Computer Vision (ECCV)*, volume 9906 of *LNCS*, pp. 102–118. Springer International Publishing, 2016.

German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio Lopez. The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016a.

Germán Ros, Simon Stent, Pablo F. Alcantarilla, and Tomoki Watanabe. Training constrained deconvolutional networks for road scene semantic segmentation. *CoRR*, abs/1604.01545, 2016b.

Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pp. 213–226. Springer, 2010.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *arXiv preprint arXiv:1606.03498*, 2016a.

Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *CoRR*, abs/1606.03498, 2016b. URL http://arxiv. org/abs/1606.03498.

Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Scribbler: Controlling deep image synthesis with sketch and color. *arXiv preprint arXiv:1612.00835*, 2016.

Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Baochen Sun and Kate Saenko. Deep CORAL: correlation alignment for deep domain adaptation. In *ICCV workshop on Transferring and Adapting Source Knowledge in Computer Vision (TASK-CV)*, 2016.

Y. Taigman, A. Polyak, and L. Wolf. Unsupervised cross-domain image generation. In *International Conference on Learning Representations*, 2017a.

Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. In *International Conference on Learning Representations (ICLR)*, 2017b.

Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR'11*, June 2011.

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *CoRR*, abs/1412.3474, 2014. URL http://arxiv.org/ abs/1412.3474.

Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *International Conference in Computer Vision (ICCV)*, 2015.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. URL http://arxiv. org/abs/1702.05464.

Zili Yi, Hao Zhang, Ping Tan Gong, et al. Dualgan: Unsupervised dual learning for image-to-image translation. *arXiv preprint arXiv:1704.02510*, 2017.

Donggeun Yoo, Namil Kim, Sunggyun Park, Anthony S. Paek, and In-So Kweon. Pixel-level domain transfer. In *European Conference on Computer Vision (ECCV)*, 2016. URL http:
//arxiv.org/abs/1603.07442.

Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmen- tation of urban scenes. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016.

Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Generative visual manip- ulation on the natural image manifold. In *European Conference on Computer Vision (ECCV)*, 2016.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *International Conference on Computer Vision (ICCV)*, 2017

LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS

CUA	Continuous Unsupervised Adaptation
CyCADA	Cycle-Consistent Adversarial Domain Adaptation
FCN	Fully-Convolutional Net
fwIoU	Frequency weighted intersection-over-union
GAN	Generative Adversarial Network
IoU	Intersection-over-union
mAP	Mean average precision
mIoU	Mean intersection-over-union
SGD	Stochastic gradient descent
SPLAT	Semantic Pixel-Level Adaptation Transforms
SVHN	Street View House Numbers
SYNTHIA	SYNTHetic collection of Imagery and Annotations