CyCADA: Cycle-Consistent Adversarial Domain Adaptation

Anonymous authors
Paper under double-blind review

Abstract

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 networks 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 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.

1 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, see Table 6).

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., 2016; 2017; Liu & Tuzel, 2016b; 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. 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.

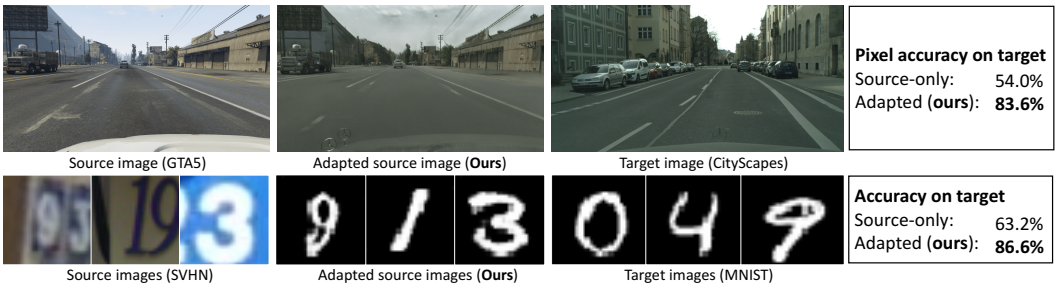


Figure 1: We propose CyCADA, an adversarial unsupervised adaptation algorithm which uses cycle and semantic consistency to perform adaptation at multiple levels in a deep network. Our model provides significant performance improvements over source model baselines.

	Pixel Loss	Feature Loss	Semantic Loss	Cycle Consistent
CycleGAN (Zhu et al., 2017)	✓			✓
Feature Adapt (Ganin & Lempitsky, 2015; Tzeng et al., 2017)		✓	✓	
Pixel Adapt (Taigman et al., 2017a; Bousmalis et al., 2017)	✓		✓	
CyCADA	✓	✓	✓	✓

Table 1: Our model, CyCADA, may use pixel, feature, and semantic information during adaptation while learning an invertible mapping through cycle consistency.

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 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 (Bousmalis et al., 2016; Liu & Tuzel, 2016b) adversarial domain adaptation methods together with cycle-consistent image-to-image translation techniques (Zhu et al., 2017), as illustrated in Table 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.

2 Related Work

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, 2016b) 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. (2017) 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.

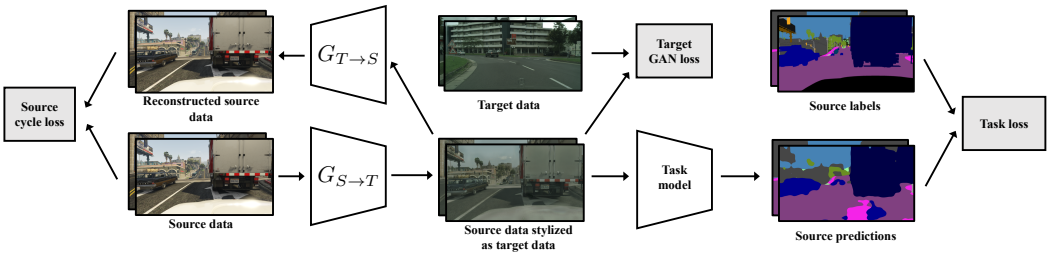


Figure 2: Cycle-consistent adversarial adaptation of pixel-space inputs. By directly remapping source training data into the target domain, we remove the low-level differences between the domains, ensuring that our task model is well-conditioned on target data. Our model also includes a target cycle loss and the source GAN loss which we omit here for clarity.

3 Cycle-Consistent Adversarial Domain Adaption

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. For K-way classification with a cross-entropy loss, this corresponds to

$$\mathcal{L}_{\text{task}}(f_S, X_S, Y_S) = \mathbb{E}_{(x_s,y_s)\sim(X_S,Y_S)} - \sum_{k=1}^{K} \mathbb{1}_{[k=y_s]} \log\left(\sigma(f_S^{(k)}(x_s))\right) \tag{1}$$

where σ denotes the softmax function. 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 corresponds to the loss function

$$\mathcal{L}_{\text{GAN}}(G_{S\to T}, D_T, X_T, X_S) = \mathbb{E}_{x_t\sim X_T}\left[\log D_T(x_t)\right] + \mathbb{E}_{x_s\sim X_S}\left[\log(1 - D_T(G_{S\to T}(x_s)))\right] \tag{2}$$

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 $\mathcal{L}_{\text{task}}(f_T, G_{S\to T}(X_S), Y_S)$.

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 in Equation 2 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.

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). To this end, we introduce another mapping from target to source $G_{T\to S}$ and train it according to the same GAN loss $\mathcal{L}_{\text{GAN}}(G_{T\to S}, D_S, X_S, X_T)$. 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 $G_{T\to S}(G_{S\to T}(x_s)) \approx x_s$ and $G_{S\to T}(G_{T\to S}(x_t)) \approx x_t$. This is done by imposing an L1 penalty on the reconstruction error, which is referred to as the *cycle-consistency loss*:

$$\begin{aligned}\mathcal{L}_{\text{cyc}}(G_{S\to T}, G_{T\to S}, X_S, X_T) = {} & \mathbb{E}_{x_s\sim X_S}\left[||G_{T\to S}(G_{S\to T}(x_s)) - x_s||_1\right] \\ & + \mathbb{E}_{x_t\sim X_T}\left[||G_{S\to T}(G_{T\to S}(x_t)) - x_t||_1\right].\end{aligned} \tag{3}$$

Additionally, as we have access to source labeled data, we explicitly encourage high semantic consistency before and after image translation. We use the trained source task model f_S 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 as follows:

$$\begin{aligned}\mathcal{L}_{\text{sem}}(G_{S\to T}, G_{T\to S}, X_S, X_T, f_S) = \mathcal{L}_{\text{task}}(f_S, X_S, p(f_S, X_S)) \\ + \mathcal{L}_{\text{task}}(f_S, G_{S\to T}(X_S), p(f_S, X_S))\end{aligned} \tag{4}$$

This can be viewed as analogously to content losses in style transfer (Gatys et al., 2016), where the shared content to preserve is dictated by the source task model f_S.

Taken together, these loss functions form our complete objective:

$$\begin{aligned}&\mathcal{L}_{\text{CyCADA}}(f_T, X_S, X_T, Y_S, G_{S\to T}, G_{T\to S}, D_S, D_T) \\ &\quad = \mathcal{L}_{\text{task}}(f_T, G_{S\to T}(X_S), Y_S) \\ &\quad + \mathcal{L}_{\text{GAN}}(G_{S\to T}, D_T, X_T, X_S) + \mathcal{L}_{\text{GAN}}(G_{T\to S}, D_S, X_S, X_T) \\ &\quad + \mathcal{L}_{\text{cyc}}(G_{S\to T}, G_{T\to S}, X_S, X_T) + \mathcal{L}_{\text{sem}}(G_{S\to T}, G_{T\to S}, X_S, X_T, f_S).\end{aligned} \tag{5}$$

This ultimately corresponds to solving for a target model f_T according to the optimization problem

$$f_T^* = \arg\min_{f_T} \min_{\substack{G_{S\to T} \\ G_{T\to S}}} \max_{D_S, D_T} \mathcal{L}_{\text{CyCADA}}. \tag{6}$$

We have presented this method as the addition of cycle-consistency and semantic constraints to standard adversarial adaptation approaches. In this setting, typically adaptation is done in feature space, and X_S and X_T are feature representations such as intermediate layers of convolutional networks. However, we can also consider the case where X_S and X_T are simply input images, such as in (Bousmalis et al., 2016; Liu & Tuzel, 2016b). In this case, adaptation is being performed in image space, and together Equations 2 and 3 reproduce the CycleGAN objective. Thus, as mentioned above, we can equivalently view CyCADA as CycleGAN augmented with an additional task loss, for the end goal of ultimately performing a discriminative task. In this work, we apply CyCADA to the semantic segmentation task and implement G as a pixel-to-pixel convnet, f as an Fully-Convolutional Net (FCN) and D as a convnet with binary outputs.

4 Experiments

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 SYNTHIA (Ros et al., 2016a), GTA (Richter et al., 2016) and CityScapes (Cordts et al., 2016) datasets.

4.1 Digit Adaptation

We evaluate our method across the adaptation shifts of USPS to MNIST, MNIST to USPS, and SVHN to MNIST. We report classification accuracy for each shift in Table 2. 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. However, for the more difficult shift of SVHN to MNIST, we find that feature level adaptation outperforms the pixel level adaptation, and both may be combined to produce an overall model which outperforms all competing methods. Additionally, the standard unsupervised CycleGAN approach diverged when training SVHN to MNIST often suffering from random label flipping. Note that the recent pixel-da method by Bousmalis et al. (2017) presents results for only the MNIST to USPS shift and reports 95.9% accuracy, while our method achieves 94.8% accuracy. However, the pixel-da approach cross validates with some labeled data which is not an equivalent evaluation setting. Figure 3 demonstrates an example where cycle constraints alone fail to have the

Model	MNIST → USPS	USPS → MNIST	SVHN → MNIST
Source only	69.8 ± 4.0	59.2 ± 2.9	64.4 ± 1.1
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	94.3 ± 0.6	95.5 ± 0.2	71.4 ± 0.7
CyCADA pixel+feat	**94.8 ± 0.2**	**95.7 ± 0.2**	**88.3 ± 0.2**
Target only	96.3 ± 0.1	99.2 ± 0.1	99.2 ±0.1

Table 2: **Unsupervised 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.

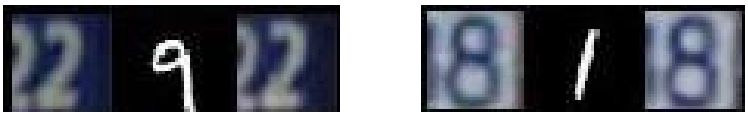


Figure 3: **Baseline SVHN to MNIST translation failures.** Examples translation failures without the semantic consistency loss. Each triple contains the original SVHN image (*left*), the image translated into MNIST style (*middle*), and the image reconstructed back into SVHN (*right*). While both the GAN and cycle constraints are satisfied (translated image matches MNIST style and reconstructed image matches original), the image translated to the target domain lacks the proper semantics.

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.

4.2 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.

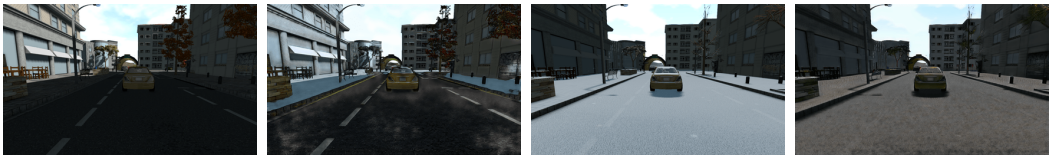
For each experiment, we use three metrics to evaluate performance. Let n_{ij} be the number of pixels of class i predicted as class j, let $t_i = \sum_j n_{ij}$ be the total number of pixels of class i, and let N be the number of classes. Our three evaluation metrics are, mean intersection-over-union (mIoU), frequency weighted intersection-over-union (fwIoU), and pixel accuracy, which are defined as follows:
$\text{mIoU} = \frac{1}{N} \cdot \frac{\sum_i n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}}$, $\text{fwIoU} = \frac{1}{\sum_k t_k} \cdot \frac{\sum_i n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}}$, $\text{pixel acc.} = \frac{\sum_i n_{ii}}{\sum_i t_i}$.

Cycle-consistent adversarial adaptation is general and can be applied at any layer of a network. Since optimizing the full CyCADA objective in Equation 5 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 segmentater into memory all at once for two images. We did not have the required memory for this at the time of submission, but leave it to future work to deploy model parallelism or experiment with larger GPU memory.

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

SYNTHIA Fall → Winter																
	sky	building	road	sidewalk	fence	vegetation	pole	car	traffic sign	pedestrian	bicycle	lanemarking	traffic light	**mIoU**	**fwIoU**	**Pixel acc.**
Source only	91.7	80.6	79.7	12.1	71.8	44.2	26.1	42.8	49.0	38.7	45.1	41.3	24.5	49.8	71.7	82.3
FCNs in the wild	92.1	86.7	91.3	20.8	72.7	**52.9**	**46.5**	64.3	**50.0**	**59.5**	**54.6**	**57.5**	**26.1**	59.6	—	—
CyCADA pixel-only	**92.5**	**90.1**	**91.9**	**79.9**	**85.7**	47.1	36.9	**82.6**	45.0	49.1	46.2	54.6	21.5	**63.3**	**85.7**	**92.1**
Oracle (Train on target)	93.8	92.2	94.7	90.7	90.2	64.4	38.1	88.5	55.4	51.0	52.0	68.9	37.3	70.5	89.9	94.5

Table 3: Adaptation between seasons in the SYNTHIA dataset. We report IoU for each class and mean IoU, freq-weighted IoU and pixel accuracy. Our CyCADA method achieves state-of-the-art performance on average across all categories. *FCNs in the wild is by Hoffman et al. (2016).



(a) Fall (b) Fall → Winter (c) Winter (d) Winter → Fall

Figure 4: **Cross Season Image Translation.** Example image-space conversions for the SYNTHIA seasons adaptation setting. We show real samples from each domain (Fall and Winter) alongside conversions to the opposite domain.

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.

4.2.1 Implementation details

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. This process is diagrammed in Figure 2.

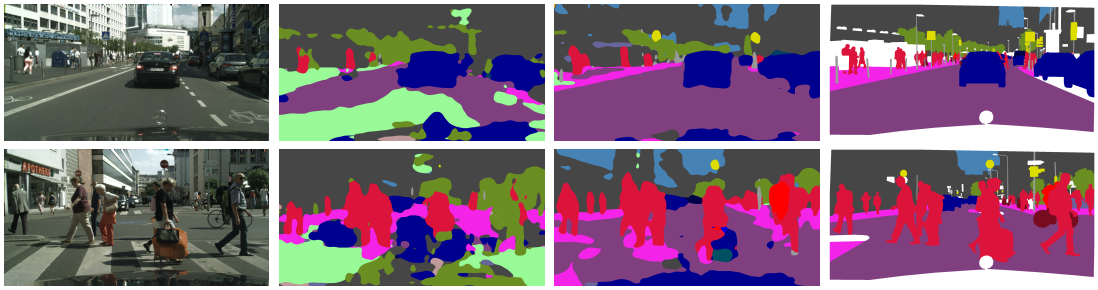
For cycle-consistent image translation, we followed the network architecture and hyperparameters of CycleGAN(Zhu et al., 2017). All images were resized to have width of 1024 pixels while keeping the aspect ratio, and the training was performed with randomly cropped patches of size 400 by 400. Also, due to large size of the dataset, we trained only 20 epochs. Then the segmentation model was trained separately using the adapted source images and the ground truth labels of the source data. We used VGG16-FCN8 architecture (Long et al., 2015), and trained for 60k iterations using SGD with learning rate 0.001.

4.2.2 Cross-season adaptation

We start by exploring the abilities of pixel space adaptation alone 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

GTA5 → Cityscapes	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorbike	bicycle	**mIoU**	**fwIoU**	**Pixel acc.**
Source only	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*	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	**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	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	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 (Train on target)	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

Table 4: 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).



(a) Test Image (b) Source Prediction (c) CyCADA Prediction (d) Ground Truth

Figure 5: **GTA5 to CityScapes Semantic Segmentation.** Each test CityScapes image (a) along with the corresponding predictions from the source only model (b) and our CyCADA model (c) are shown and may be compared against the ground truth annotation (d).

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. In Figure 4 we show the result of pixel only adaptation as we generate a winter domain image (b) from a fall domain image (a), and visa versa (c-d). We may clearly see the changes of adding or removing snow. 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 as shown in Table 3. 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. This conclusion is supported by inspection of the individual classes in Table 3, where we see the largest improvement on common classes such as *road* and *sidewalk*.

4.2.3 Synthetic to Real Adaptation

To evaluate our method's applicability to real-world adaptation settings, we investigate adaptation from synthetic to real-world imagery. The results of this evaluation are presented in Table 4 with qualitative results shown in Figure 5. 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

(a) GTA5 (b) GTA5 → Cityscapes (c) CityScapes (d) CityScapes → GTA5

Figure 6: **GTA5 to CityScapes Image Translation.** Example images from the GTA5 (a) and Cityscapes (c) datasets, alongside their image-space conversions to the opposite domain, (b) and (d), respectively. Our model achieves highly realistic domain conversions.

class 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 visualize the results of image-space adaptation between GTA5 and Cityscapes in Figure 6. 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.

5 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.

References

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

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in Neural Information Processing Systems*, pp. 343–351, 2016.

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 2017.

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 Conference 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 Neural 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. URL `http://arxiv.org/abs/1604.01545`.

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.

Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation 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 manipulation 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.