CyCADA: Cycle-Consistent Adversarial Domain Adaptation

Judy Hoffman [1] **Eric Tzeng** [1] **Taesung Park** [1] **Jun-Yan Zhu** [1] **Phillip Isola** [1 2] **Kate Saenko** [3] **Alexei A. Efros** [1] **Trevor Darrell** [1]

Abstract

Domain adaptation is critical for success in new, unseen environments. Adversarial adaptation models have shown tremendous progress towards adapting to new environments by focusing either on discovering domain invariant representations or by mapping between unpaired image domains. While feature space methods are difficult to interpret and sometimes fail to capture pixel-level and low-level domain shifts, image space methods sometimes fail to incorporate high level semantic knowledge relevant for the end task. We propose a model which adapts between domains using both generative image space alignment and latent representation space alignment. Our approach, Cycle-Consistent Adversarial Domain Adaptation (CyCADA), guides transfer between domains according to a specific discriminatively trained task and avoids divergence by enforcing consistency of the relevant semantics before and after adaptation. We evaluate our method on a variety of visual recognition and prediction settings, including digit classification and semantic segmentation of road scenes, advancing state-of-the-art performance for unsupervised adaptation from synthetic to real world driving 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).

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.

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., 2017b; Shrivastava et al., 2017;

[1]EECS and BAIR, UC Berkeley [2]Openai (work done while at UC Berkeley) [3]CS Department, Boston University. Correspondence to: Judy Hoffman <jhoffman@eecs.berkeley.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

Liu et al., 2017). 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., 2017a) was applied to larger images, but in a controlled environment with visually simple images for robotic applications. Image to Image translation techniques, such as CycleGAN (Zhu et al., 2017), have produced visually appealing results which preserve local content in natural scenes, but are not designed with an end task in mind and so may not always preserve semantics. For example, a model adapting from digits taken from Google Street View to handwritten digits can learn to make a printed 8 look like a hand-written 1.

We propose Cycle-Consistent Adversarial Domain Adaptation (CyCADA), which adapts representations at both the pixel-level and feature-level while enforcing semantic consistency. We enforce both structural and semantic consistency during adaptation using a cycle-consistency loss (ie. the source should match the source mapped to target mapped back to source) and semantics losses based on a particular visual recognition task. The semantics losses both guide the overall representation to be discriminative and enforce semantic consistency before and after mapping between domains. Our approach offers a unified domain adversarial learning model which combines the interpretability and low level structural consistency of prior image-level approaches (Liu & Tuzel, 2016a; Bousmalis et al., 2017b; Shrivastava et al., 2017; Zhu et al., 2017; Liu et al., 2017) together with the regularization and strong empirical performance of prior feature-level approaches (Ganin & Lempitsky, 2015; Tzeng et al., 2017), as illustrated in Table 1.

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 83%, nearly closing the gap to the target-trained model and providing 16% relative improvement over current state-of-the-art.

We demonstrate that enforcing both semantic (task-specific) consistency and cycle consistency between input and stylized images prevents label flipping on the large shift between SVHN and MNIST (example, prevents a SVHN 9 from being mapped into an MNIST 2). On our semantic segmentation tasks (GTA to CityScapes) we did not observe label flipping to be a major source of error, even without the semantic consistency loss, but found cycle consistency to be critical. Because of this, and due to memory constraints, we focus on cycle consistency to preserve structure during transfer for the segmentation tasks. Overall, our experiments confirm that domain adaptation can benefit greatly from a combination of pixel and representation transformations, with the joint adaptation model achieving the highest performance across a range of visual recognition tasks.

	Pixel Loss	Feature Loss	Semantic Consistent	Cycle Consistent
CycleGAN (Zhu et al., 2017)	✓			✓
Feature Adapt†		✓	✓	
Pixel Adapt‡	✓		✓	
CyCADA	✓	✓	✓	✓

Table 1: Our model, CyCADA, may use pixel, feature, and semantic information during adaptation while learning an invertible mapping through cycle consistency. †(Ganin & Lempitsky, 2015; Tzeng et al., 2017), ‡(Taigman et al., 2017a; Bousmalis et al., 2017b; Liu et al., 2017)

2. 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_T that correctly predicts the label for the target data X_T.

Pretrain Source Task Model. We 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

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

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.

Pixel-level Adaptation. 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

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

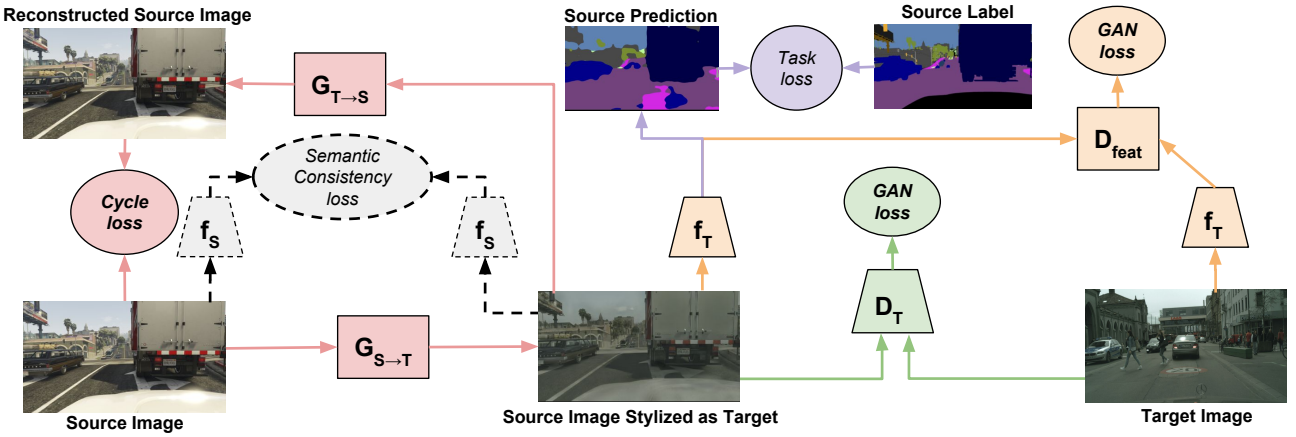


Figure 2: Cycle-consistent adversarial adaptation overview. By directly remapping source training data into the target domain, we remove the low-level differences between the domains, ensuring that our task model is well-conditioned on target data. We depict here the image-level adaptation as composed of the pixel GAN loss (green), the source cycle loss (red), and the source and target semantic consistency losses (black dashed) – used when needed to prevent label flipping. For clarity the target cycle is omitted. The feature-level adaptation is depicted as the feature GAN loss (orange) and the source task loss (purple).

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 $\mathcal{L}_{\text{task}}(f_T, G_{S\to T}(X_S), Y_S)$ (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 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) (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 $\mathcal{L}_{\text{GAN}}(G_{T\to S}, D_S, X_S, X_T)$. 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. This helps to prevent label flipping described above and illustrated in Figure 4(a). We use the pretrained 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:

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

See Figure 2 black portion. This can be viewed as analogous to content losses in style transfer (Gatys et al., 2016) or in pixel adaptation (Taigman et al., 2017a), where the shared content to preserve is dictated by the source task model f_S.

Feature-level Adaptation. 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 (see Figure 2 orange portion):

$$\mathcal{L}_{\text{GAN}}(f_T, D_{\text{feat}}, f_S(G_{S\to T}(X_S)), X_T). \tag{5}$$

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, 2016b)	91.2	89.1	-
ADDA (Tzeng et al., 2017)	89.4 ± 0.2	90.1 ± 0.8	76.0 ± 1.8
PixelDA (Bousmalis et al., 2017b)	**95.9**	-	-
UNIT (Liu et al., 2017)	**95.9**	93.6	**90.5***
CyCADA (Ours)	**95.6 ± 0.2**	**96.5 ± 0.1**	**90.4 ± 0.4**
Target Fully Supervised	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. *UNIT trains with the extended SVHN (>500K images vs ours 72K).

Taken together, these loss functions form our complete objective:

$$\begin{aligned}\mathcal{L}_{\text{CyCADA}}(f_T, X_S, X_T, Y_S, G_{S\to T}, G_{T\to S}, D_S, D_T) \quad (6)\\ = \mathcal{L}_{\text{task}}(f_T, G_{S\to T}(X_S), Y_S)\\ + \mathcal{L}_{\text{GAN}}(G_{S\to T}, D_T, X_T, X_S)\\ + \mathcal{L}_{\text{GAN}}(G_{T\to S}, D_S, X_S, X_T)\\ + \mathcal{L}_{\text{GAN}}(f_T, D_{\text{feat}}, f_S(G_{S\to T}(X_S)), X_T)\\ + \mathcal{L}_{\text{cyc}}(G_{S\to T}, G_{T\to S}, X_S, X_T)\\ + \mathcal{L}_{\text{sem}}(G_{S\to T}, G_{T\to S}, X_S, X_T, f_S).\end{aligned}$$

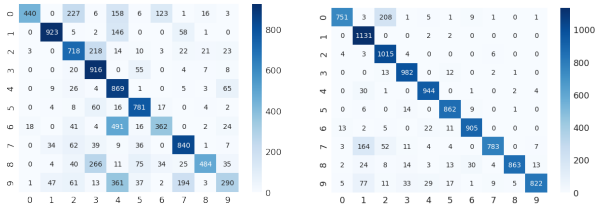
This ultimately corresponds to solving for a target model f_T according to the optimization problem

$$f_T^* = \arg\min_{f_T} \min_{\substack{G_{S\to T}\\ G_{T\to S}}} \max_{D_S, D_T} \mathcal{L}_{\text{CyCADA}} \quad (7)$$

We have proposed a method for unsupervised adaptation which views prior adversarial objectives as operating at the pixel or feature level and generalizes to a method which may benefit from both approaches. In addition, we combine cycle-consistency together with semantic transformation constraints to regularize the mapping from one domain to another. In the next section, we apply CyCADA to both digit classification and to semantic segmentation, implementing G_S and G_T as a pixel-to-pixel convnet, f_S and f_T as a convnet classifier or a Fully-Convolutional Net (FCN), and D_S, D_T, and D_{feat} as a convnet with binary outputs.

3. 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 GTA (Richter et al., 2016) and CityScapes (Cordts et al., 2016) datasets, see Appendix A.1.2 for an additional experiment with the SYNTHIA (Ros et al., 2016a) dataset.



(a) Source only Model (b) CyCADA model

Figure 3: Confusion matrices for the SVHN → MNIST experiment before and after adaptation.

Model	Accuracy (%)
Source only	67.1
CyCADA - no feat adapt, no semantic loss	70.3
CyCADA - no feat adapt	71.2
CyCADA - no cycle consistency	75.7
CyCADA - no pixel adapt	83.8
CyCADA (Full)	**90.4**
Target Fully Supervised	99.2

Table 3: **Ablation of CyCADA on SVHN→MNIST Domain Shift.** We show that each component of our method, joint feature and pixel space adaptation, with semantic and cycle losses during pixel adaptation, contributes to the overall performance.

3.1. Digit Adaptation

We evaluate our method across the adaptation shifts of USPS to MNIST, MNIST to USPS, and SVHN to MNIST. We train our model using the training sets, MNIST - 60,000 images, USPS - 7,291 images, standard SVHN train - 73,257 images. Evaluation is reported on the standard test sets: MNIST - 10,000 images, USPS - 2,007 images. We report classification accuracy for each shift compared to prior work and relevant baselines in Table 2 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 (see Supplemental A.1.1 for full implementation details). Note, for the relatively easy shift of MNIST to USPS our method performs comparably with state-of-the-art approaches. In the reverse direction, when adapting from USPS images to MNIST, which involves a fraction of the supervised digit labeled data, our method outperforms competing approaches. For SVHN to MNIST our method outperforms all other deep distribution alignment approaches except for UNIT (Liu et al., 2017), but the reported performance in UNIT uses the extended training set of >500,000 images from SVHN whereas we report performance using the standard set of only 73,257 images.

(a) Without Semantic Loss (b) Without Cycle Loss

Figure 4: **Ablation: Effect of Semantic or Cycle Consistency.** Each triple contains the SVHN image (*left*), the image translated into MNIST style (*middle*), and the image reconstructed back into SVHN (*right*). (a) Without semantic loss, both the GAN and cycle constraints are satisfied (translated image matches MNIST style and reconstructed image matches original), but the image translated to the target domain lacks the proper semantics. (b) Without cycle loss, the reconstruction is not satisfied and though the semantic consistency leads to some successful semantic translations (*top*) there are still cases of label flipping (*bottom*).

To further 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 (Figure 3). 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.

Next, we perform a sequence of ablation studies on the three parts of our model. Table 3 reports the quantitative performance gain on the SVHN→MNIST domain shift for removing each piece of the model, demonstrating the importance of including each component. We also discuss and show qualitative comparisons below.

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. Figure 4(a) demonstrates two examples where cycle constraints alone fail to produce the desired behavior for our end task. An SVHN image is mapped to a convincing MNIST style image and back to a SVHN image with correct semantics. However, the MNIST-like image has mismatched semantics. Our proposed approach uses the source labels to train a weak classification model which can be used to enforce semantic consistency before and after translation, resolving this issue.

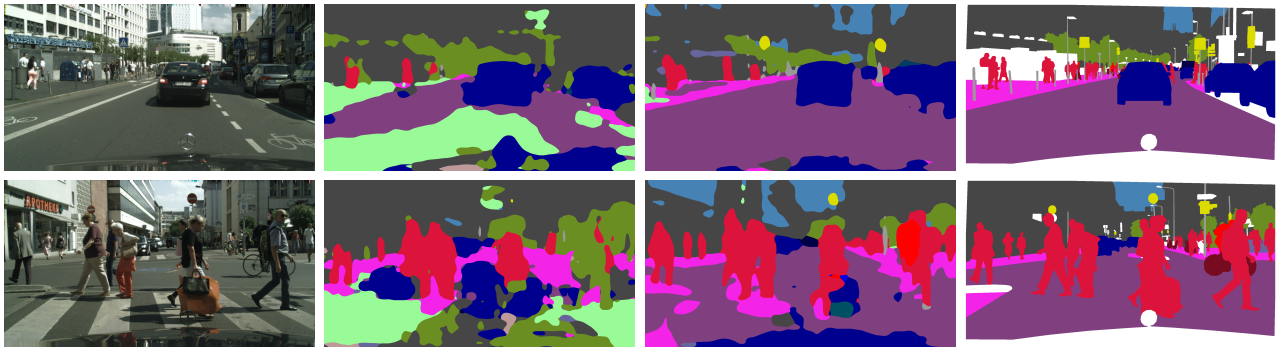
Ablation: No Cycle Consistency. We study the importance of the cycle consistency loss. First note that without this loss there is no reconstruction guarantee, thus in Figure 4(b) we see that the translation back to SVHN fails. In addition, we find that while the semantic loss does encourage correct semantics it relies on the weak source labeler and thus label flipping still occurs (see right image triple).

3.2. Semantic Segmentation Adaptation

Next, we evaluate CyCADA on semantic segmentation. 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 report three metrics of overall 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 6 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. We did not have the required memory for



(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).

GTA5 → Cityscapes																							
	Architecture	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	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
FCN-wld (Hoffman et al., 2016)	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	-	-
CDA (Zhang et al., 2017b)	A	26.4	22.0	74.7	6.0	11.9	8.4	16.3	11.1	75.7	13.3	**66.5**	38.0	**9.3**	55.2	**18.8**	18.9	0.0	**16.8**	**14.6**	27.8	-	-
FCTN (Zhang et al., 2017a)	A	72.2	28.4	74.9	18.3	10.8	**24.0**	**25.3**	17.9	80.1	**36.7**	61.1	44.7	0.0	74.5	8.9	1.5	0.0	0.0	0.0	30.5	-	-
CyCADA (Ours)	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 Supervised	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 (Ours)	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 Supervised	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

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. 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).

(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.

this at the time of submission, but leave it to future work to deploy model parallelism or experiment with larger GPU memory.

To 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. For an additional experiment evaluating cross-season adaptation in synthetic environments see the Appendix A.2.

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 implementation details please see Appendix A.1.2.

3.2.1. 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, ablation in Table 5, with qualitative results shown in Figure 5. We experiment with two different base architectures: the commonly used VGG16-FCN8s (Long et al., 2015) architecture as well as the state-of-the-art DRN-26 (Yu et al., 2017) architecture. 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 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 (see Appendix for full details).

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.

GTA5 → Cityscapes				
	Architecture	**mIoU**	**fwIoU**	**Pixel acc.**
Source only	A	17.9	41.9	54.0
CyCADA feat-only	A	29.2	71.5	82.5
CyCADA pixel-only no cycle	A	19.8	55.7	70.5
CyCADA pixel-only	A	34.8	73.1	82.8
CyCADA (Full)	A	**35.4**	**73.8**	**83.6**
Oracle - Target Supervised	A	60.3	87.6	93.1
Source only	B	21.7	47.4	62.5
CyCADA feat-only	B	31.7	67.4	78.4
CyCADA pixel-only no cycle	B	19.7	54.5	69.9
CyCADA pixel-only	B	37.0	63.8	75.4
CyCADA (Full)	B	**39.5**	**72.4**	**82.3**
Oracle - Target Supervised	B	67.4	89.6	94.3

Table 5: Ablation of our method, CyCADA on the GTA5 to Cityscapes adaptation. 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).

4. 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 (Goodfel-

low 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, 2016a) 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. Liu et al. (2017) considers learning unique encoders which reach a shared latent space and can be reconstructed into the same domnain or translated into the other domain. Manually defined sharing of certain layers are used to encourage consistency between the two domain models. 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/ALI (Donahue et al., 2017; Dumoulin et al., 2016) take an approach of simultaneously learning the transformations between the pixel and the latent space. 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.

An approach to adaptation which is complementary to this work involves seeking to produce approximate labels for the target domain and incorporate those into the training set for supervised learning (Haeusser et al., 2017).

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. (2017b) instead perform segmentation adaptation by aligning label distributions both globally and across superpixels in an image.

5. Conclusion

We proposed an unsupervised domain adversarial learning method that unifies cycle-consistent image translation adversarial models with adversarial adaptation methods. CyCADA offers the interpretability of image-space adaptation, by visualizing the intermediate output of our method, while producing a discriminative and task relevant model through semantic consistency and representation space adaptation. We experimentally validated our model on a variety of adaptation tasks including digit adaptation and synthetic to real adaptation for semantic segmentation of driving scenes. We presented extensive ablations of our method demonstrating the importance of each component of our method, where the combination results in a state-of-the-art approach.

Acknowledgements

The authors would like to thank the following funding agencies for their support: Prof. Darrell was supported in part by DARPA; NSF awards IIS-1212798, IIS-1427425, and IIS-1536003, Berkeley DeepDrive, and the Berkeley Artificial Intelligence Research Center. Prof. Efros was supported in part by NSF IIS-1633310 and Berkeley DeepDrive.

References

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein gan. *CoRR*, abs/1701.07875, 2017. URL http://arxiv.org/abs/1701.07875.

Bousmalis, K., Irpan, A., Wohlhart, P., Bai, Y., Kelcey, M., Kalakrishnan, M., Downs, L., Ibarz, J., Pastor, P., Konolige, K., Levine, S., and Vanhoucke, V. Using simulation and domain adaptation to improve efficiency of deep robotic grasping. *CoRR*, abs/1709.07857, 2017a. URL http://arxiv.org/abs/1709.07857.

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

Chen, Y.-H., Chen, W.-Y., Chen, Y.-T., Tsai, B.-C., Frank Wang, Y.-C., and Sun, M. No more discrimination: Cross city adaptation of road scene segmenters. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. The Cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

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

Donahue, J., Krähenbühl, P., and Darrell, T. Adversarial feature learning. In *ICLR*, 2017.

Dumoulin, V., Belghazi, I., Poole, B., Lamb, A., Arjovsky, M., Mastropietro, O., and Courville, A. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.

Ganin, Y. and Lempitsky, V. Unsupervised domain adaptation by backpropagation. In Blei, D. and Bach, F. (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.

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

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

Ghifary, M., Kleijn, W. B., Zhang, M., Balduzzi, D., and Li, W. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European Conference on Computer Vision (ECCV)*, pp. 597–613. Springer, 2016.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*. 2014. URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

Haeusser, P., Frerix, T., Mordvintsev, A., and Cremers, D. Associative domain adaptation. In *International Conference on Computer Vision (ICCV)*, 2017.

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

Hoffman, J., Wang, D., Yu, F., and Darrell, T. FCNs in the wild: Pixel-level adversarial and constraint-based adaptation. *CoRR*, abs/1612.02649, 2016. URL http://arxiv.org/abs/1612.02649.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016.

Karacan, L., Akata, Z., Erdem, A., and Erdem, E. Learning to generate images of outdoor scenes from attributes and semantic layouts. *arXiv preprint arXiv:1612.00215*, 2016.

Kim, T., Cha, M., Kim, H., Lee, J. K., and Kim, J. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, 2017.

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

Levinkov, E. and Fritz, M. 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.

Liu, M. and Tuzel, O. Coupled generative adversarial networks. In *Neural Information Processing Systems (NIPS)*, 2016a.

Liu, M., Breuel, T., and Kautz, J. Unsupervised image-to-image translation networks. In *Neural Information Processing Systems (NIPS)*, 2017.

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

Long, J., Shelhamer, E., and Darrell, T. Fully convolutional networks for semantic segmentation. *CVPR (to appear)*, November 2015.

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

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

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. 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.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Richter, S. R., Vineet, V., Roth, S., and Koltun, V. Playing for data: Ground truth from computer games. In Leibe, B., Matas, J., Sebe, N., and Welling, M. (eds.), *European Conference on Computer Vision (ECCV)*, volume 9906 of *LNCS*, pp. 102–118. Springer International Publishing, 2016.

Ros, G., Sellart, L., Materzynska, J., Vazquez, D., and Lopez, A. 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.

Ros, G., Stent, S., Alcantarilla, P. F., and Watanabe, T. Training constrained deconvolutional networks for road scene semantic segmentation. *CoRR*, abs/1604.01545, 2016b. URL http://arxiv.org/abs/1604.01545.

Saenko, K., Kulis, B., Fritz, M., and Darrell, T. Adapting visual category models to new domains. In *European conference on computer vision*, pp. 213–226. Springer, 2010.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. *arXiv preprint arXiv:1606.03498*, 2016a.

Salimans, T., Goodfellow, I. J., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. *CoRR*, abs/1606.03498, 2016b. URL http://arxiv.org/abs/1606.03498.

Sangkloy, P., Lu, J., Fang, C., Yu, F., and Hays, J. Scribbler: Controlling deep image synthesis with sketch and color. *arXiv preprint arXiv:1612.00835*, 2016.

Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., and Webb, R. Learning from simulated and unsupervised images through adversarial training. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

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

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

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

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

Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., and Darrell, T. Deep domain confusion: Maximizing for domain invariance. *CoRR*, abs/1412.3474, 2014. URL http://arxiv.org/abs/1412.3474.

Tzeng, E., Hoffman, J., Darrell, T., and Saenko, K. Simultaneous deep transfer across domains and tasks. In *International Conference in Computer Vision (ICCV)*, 2015.

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. Adversarial discriminative domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. URL http://arxiv.org/abs/1702.05464.

Yi, Z., Zhang, H., Gong, P. T., et al. Dualgan: Unsupervised dual learning for image-to-image translation. *arXiv preprint arXiv:1704.02510*, 2017.

Yoo, D., Kim, N., Park, S., Paek, A. S., and Kweon, I. Pixel-level domain transfer. In *European Conference on Computer Vision (ECCV)*, 2016. URL http://arxiv.org/abs/1603.07442.

Yu, F., Koltun, V., and Funkhouser, T. Dilated residual networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

Zhang, J., Chen, L., and Kuo, C. J. A fully convolutional tri-branch network (FCTN) for domain adaptation. *CoRR*, abs/1711.03694, 2017a. URL http://arxiv.org/abs/1711.03694.

Zhang, Y., David, P., and Gong, B. Curriculum domain adaptation for semantic segmentation of urban scenes. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017b.

Zhao, J., Mathieu, M., and LeCun, Y. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016.

Zhu, J.-Y., Krähenbühl, P., Shechtman, E., and Efros, A. A. Generative visual manipulation on the natural image manifold. In *European Conference on Computer Vision (ECCV)*, 2016.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *International Conference on Computer Vision (ICCV)*, 2017.