SPIGAN: PRIVILEGED ADVERSARIAL LEARNING FROM SIMULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep Learning for Computer Vision depends mainly on the source of supervision. Photo-realistic simulators can generate large-scale automatically labeled synthetic data, but introduce a domain gap negatively impacting performance. We propose a new unsupervised domain adaptation algorithm, called SPIGAN, relying on Simulator Privileged Information (PI) and Generative Adversarial Network (GAN). We use internal data structures of the simulator as PI only available during training of a target task network. We experimentally evaluate our approach on semantic segmentation. We use the real-world Cityscapes and Vistas datasets, training using only unlabeled real-world images with synthetic labeled data and z-buffer (depth) PI from the SYNTHIA dataset. Our method improves over no adaptation and state-of-the-art unsupervised domain adaptation techniques.

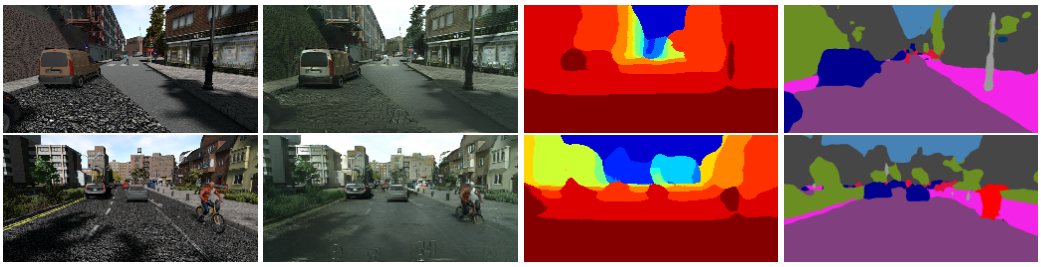


Figure 1: SPIGAN example inputs and outputs. From left to right: input images from a simulator; adapted images from SPIGAN's generator network; predictions from SPIGAN's privileged network (depth layers); semantic segmentation predictions from the target task network.

1 INTRODUCTION

Learning from as little human supervision as possible is a major challenge in Machine Learning. In Computer Vision, labeling images and videos is the main bottleneck towards achieving large scale learning and generalization. Recently, training in simulation has shown continuous improvements in several tasks, such as optical flow (Mayer et al., 2016), object detection (Marín et al., 2010; Vazquez et al., 2014; Xu et al., 2014; Sun & Saenko, 2014; Peng et al., 2015), tracking (Gaidon et al., 2016), pose and viewpoint estimation (Shotton et al., 2011; Papon & Schoeler, 2015; Su et al., 2015), action recognition (de Souza et al., 2017), and semantic segmentation (Handa et al., 2016; Ros et al., 2016; Richter et al., 2016). However, large domain gaps between synthetic and real domains remain as the main handicap of this type of strategies. This is often addressed by manually labeling some amount of real-world target data to train the model on mixed synthetic and real-world labeled data (supervised domain adaptation). In contrast, several recent *unsupervised* domain adaptation algorithms have leveraged the potential of Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) for pixel-level adaptation in this context (Bousmalis et al., 2017; Shrivastava et al., 2016). These methods often use simulators as black-box generators of (x, y) input / output training samples for the desired task.

Our main observation is that simulators internally know a lot more about the world and how the scene is formed. This Privileged Information includes physical properties that might be useful for learning. This additional information z is not available in the real-world and is therefore generally

ignored during learning. In this paper, we propose a novel adversarial learning algorithm, called *SPIGAN*, to leverage Simulator *Privileged Information (PI)* for GAN-based *unsupervised* learning of a target task network from *unpaired unlabeled* real-world data.

We jointly learn four different networks: (i) a generator G (to adapt the pixel-level distribution of synthetic images to be more like real ones), (ii) a discriminator D (to distinguish adapted and real images), (iii) a task network T (to predict the desired label y from image x), and (iv) a *privileged network* P trained on both synthetic images x and adapted ones $G(x)$ to predict their associated privileged information z. Our main contribution is a new method to leverage *PI* from a simulator via the privileged network P, which acts as an auxiliary task and regularizer to the task network T, the main output of our SPIGAN learning algorithm.

We evaluate our approach on semantic segmentation in urban scenes, a challenging real-world task. We use the standard Cityscapes (Cordts et al., 2016) and Vistas (Neuhold et al., 2017) datasets as target real-world data (without using any of the training labels) and SYNTHIA (Ros et al., 2016) as simulator output. Although our method applies to any kind of PI that can be predicted via a deep network (optical flow, instance segmentation, object detection, material properties, forces, ...), we consider one of the most common and simple forms of PI available in any simulator: depth from its z-buffer. We show that SPIGAN can successfully learn a semantic segmentation network T using no real-world labels, partially bridging the sim-to-real gap (see Figure 1). SPIGAN also outperforms related state-of-the-art unsupervised domain adaptation methods.

The rest of the paper is organized as follows. Section 2 presents a brief review of related works. Section 3 presents our SPIGAN unsupervised domain adaptation algorithm using simulator privileged information. We report our quantitative experiments on semantic segmentation in Section 4, and conclude in Section 5.

2 Related Work

Domain adaptation (cf. Csurka (2017) for a recent review) is generally approached either as domain-invariant learning (Hoffman et al., 2013; Herath et al., 2017; Yan et al., 2017; Ganin & Lempitsky, 2015) or as a statistical alignment problem (Tzeng et al., 2014; Long et al., 2015). Our work focuses on *unsupervised* adaptation methods in the context of deep learning. The Domain Adversarial Neural Network (DANN) (Tzeng et al., 2014; Ganin & Lempitsky, 2015; Ganin et al., 2016) is a popular approach that learns domain invariant features by maximizing domain confusion. This approach has been successfully adopted and extended by many other researchers, e.g., Purushotham et al. (2017); Chen et al. (2017); Zhang et al. (2017). Curriculum Domain Adaptation (Zhang et al., 2017) is a recent evolution for semantic segmentation that reduces the domain gap via a curriculum learning approach (solving simple tasks first, such as global label distribution in the target domain).

Recently, adversarial domain adaptation based on GANs (Goodfellow et al., 2014) have shown encouraging results for unsupervised domain adaptation directly at the pixel level. These techniques learn a generative model for source-to-target image translation, including from and to multiple domains (Taigman et al., 2016; Shrivastava et al., 2016; Zhu et al., 2017; Isola et al., 2017; Kim et al., 2017). In particular, CycleGAN (Zhu et al., 2017) leverages cycle consistency using a forward GAN and a backward GAN to improve the training stability and performance of image-to-image translation. An alternative to GAN is Variational Auto-Encoders (VAEs), which have also been used for image translation (Liu et al., 2017).

Several related works propose GAN-based unsupervised domain adaptation methods to address the specific domain gap between synthetic and real-world images. SimGAN (Shrivastava et al., 2016) leverages simulation for the automatic generation of large annotated datasets with the goal of refining synthetic images to make them look more realistic. Sadat Saleh et al. (2018) effectively leverages synthetic data by treating foreground and background in different manners. Similar to our approach, recent methods consider the final recognition task during the image translation process. Closely related to our work, PixelDA (Bousmalis et al., 2017) is a pixel-level domain adaptation method that jointly trains a task classifier along with a GAN using simulation as its source domain but no privileged information. These approaches focus on simple tasks and visual conditions that are easy to simulate, hence having a low domain gap to begin with. On the other hand, Hoffman et al. (2016) pioneers the research line of using semantic segmentation as the task network in adversarial

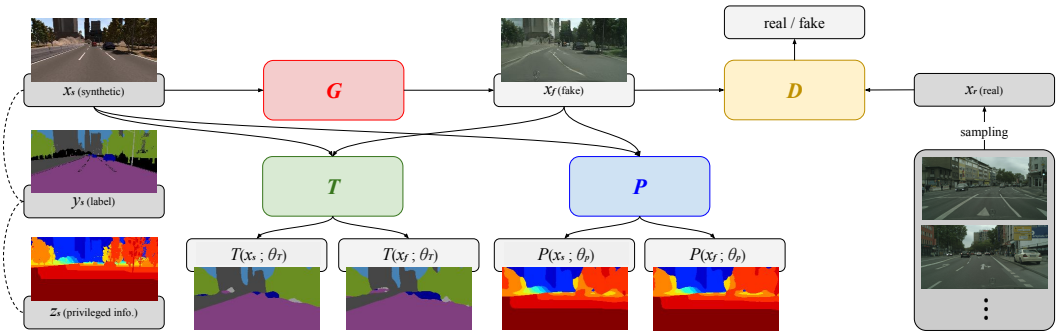


Figure 2: SPIGAN learning algorithm from unlabeled real-world images x_r and the unpaired output of a simulator (synthetic images x_s, their labels y_s, e.g. semantic segmentation ground truth, and Privileged Information PI z_s, e.g., depth from the z-buffer) modeled as random variables. Four networks are learned jointly: (i) a generator $G(x_s) \sim x_r$, (ii) a discriminator D between $G(x_s) = x_f$ and x_r, (iii) a perception task network $T(x_r) \sim y_r$, the main target output of SPIGAN (e.g., semantic segmentation), and (iv) a privileged network P to support the learning of T by predicting the simulator's PI z_s.

training. Zhang et al. (2017) uses a curriculum learning style approach to reduce domain gap. Saito et al. (2017) conducts domain adaptation by utilizing the task-specific decision boundaries with classifiers. Sankaranarayanan et al. (2018) leverages the GAN framework by learning general representation shared between the generator and segmentation networks. Chen et al. (2018) use a target guided distillation to encourage the task network to imitate the pretrained model. Zhang et al. (2018) propose to combine appearance and representation adaptation.

Our main novelty is the use of Privileged Information from a simulator in a generic way by considering a privileged network in our architecture (see Figure 2). We show that for the challenging task of semantic segmentation of urban scenes, our approach significantly improves by augmenting the learning objective with our auxiliary privileged task, especially in the presence of a large sim-to-real domain gap, the main problem in challenging real-world conditions.

Our work is inspired by Learning Using Privileged Information (LUPI) (Vapnik & Vashist, 2009), which is linked to distillation (Hinton et al., 2015) as shown by Lopez-Paz et al. (2015). LUPI's goal is to leverage additional data only available at training time. For unsupervised domain adaptation from a simulator, there is a lot of potentially useful information about the generation process that could inform the adaptation. However, that information is only available at training time, as we do not have access to the internals of the real-world data generator. Several works have used the LUPI formalism for domain adaptation (Chen et al., 2014; Li et al., 2014; Sarafianos et al., 2017). In particular, Sarafianos et al. (2017) propose an extension of the SVM+ framework (Lapin et al., 2014), which is linked to importance sampling. None of these works, however, exploit the rich privileged information from simulators for sim-to-real unsupervised adaptation.

3 Simulator Privileged Information GAN

3.1 Unsupervised learning with a simulator

Our goal is to design a procedure to learn a model (neural network) that solves a perception task (e.g., semantic segmentation) using raw sensory data coming from a target domain (e.g., videos of a car driving in urban environments) *without using any ground truth data from the target domain*. We formalize this problem as *unsupervised domain adaptation* from a synthetic domain (source domain) to a real domain (target domain). The source domain consists of labeled synthetic images together with Privileged Information *PI*, obtained from the internal data structures of a simulator. The target domain consists of unlabeled images.

The simulated source domain serves as an idealized representation of the world, offering full control of the environment (weather conditions, type of scene, sensor configurations, etc.) with automatic generation of raw sensory data and labels for the task of interest. The main challenge we address in

this work is how to overcome the gap between this synthetic source domain and the target domain to ensure generalization of the task network in the real-world *without target supervision*.

Our main hypothesis is that the PI provided by the simulator is a rich source of information to guide and constrain the training of the target task network. The PI can be defined as any information internal to the simulator, such as depth, optical flow, or physical properties about scene components used during simulation (e.g., materials, forces, etc.). We leverage the simulator's PI within a GAN framework, called SPIGAN. Our approach is described in the next section.

3.2 SPIGAN

Let $X_r = \{x_r^{(j)},\ j = 1 \dots N^r\}$ be a set of N^r unlabeled real-world images x_r. Let $X_s = \{(x_s^{(i)}, y_s^{(i)}, z_s^{(i)}),\ i = 1 \dots N^s\}$ be a set of N^s simulated images x_s with their labels y_s and PI z_s. We describe our approach assuming a unified treatment of the PI, but our method trivially extends to multiple separate types of PI.

SPIGAN (cf. Fig. 2) jointly learns a model $(\theta_G, \theta_D, \theta_T, \theta_P)$, consisting of: (i) a generator $G(x; \theta_G)$; (ii) a discriminator $D(x; \theta_D)$; (iii) a task predictor $T(x; \theta_T)$ and (iv) a privileged network $P(x; \theta_P)$. The generator G is a mapping function, transforming an image x_s in X_s (source domain) to x_f in X_f (adapted or fake domain). SPIGAN has to make the adapted domain to be statistically close to the target domain in order to maximize the accuracy of the task predictor $T(x; \theta_T)$ during testing. The discriminator D is expected to tell the difference between x_f and x_r, playing an adversarial game with the generator until a termination criteria is met (refer to section 4.1) . The target task network T is learned on the synthetic x_s and adapted $G(x_s; \theta_G)$ images to predict the synthetic label y_s, assuming the generator presents a reasonable degree of label (content) preservation. This assumption is met for the regime of our experiments. Similarly, the privileged network P is trained on the same input but to predict the PI z, which in turn assumes the generator G is also PI-preserving. During testing only $T(x; \theta_T)$ is needed to do inference for the selected perception task.

The main learning goal is to train a model θ_T that can correctly perform a perception task T in a real domain (target), showing a performance in the range to their fully-supervised counterparts. All models are trained jointly in order to exploit all available information to constraint the solution space. In this way, the PI provided by the privileged network P is used to constrain the learning of T, and to encourage the generator to model the target domain while being label- and PI-preserving. Our joint learning objective is described in the following section.

3.3 Learning objective

We design a consistent set of loss functions and domain-specific constraints related to the main prediction task T. We optimize the following minimax objective:

$$\min_{\boldsymbol{\theta}_G, \boldsymbol{\theta}_T, \boldsymbol{\theta}_P} \max_{\boldsymbol{\theta}_D} \ \alpha \mathcal{L}_{\text{GAN}} + \beta \mathcal{L}_T + \gamma \mathcal{L}_P + \delta \mathcal{L}_{\text{perc}} \tag{1}$$

where α, β, γ, δ are the weights for adversarial loss, task prediction loss, PI regularization, and perceptual regularization respectively, further described below.

Adversarial loss $\mathcal{L}_{\text{GAN}}$. Instead of using a standard adversarial loss, we use a least-squares based adversarial loss Mao et al. (2016); Zhu et al. (2017), which stabilizes the training process and generates better image results in our experiments:

$$\begin{aligned} \mathcal{L}_{\text{GAN}}(D, G) =& \mathbb{E}_{x_r \sim \mathcal{P}_r}[(D(x_r; \theta_D) - 1)^2] \\ &+ \mathbb{E}_{x_s \sim \mathcal{P}_s}[D(G(x_s; \theta_G); \theta_D)^2] \end{aligned} \tag{2}$$

where $\mathcal{P}_r$ (resp. $\mathcal{P}_s$) denotes the real-world (resp. synthetic) data distribution.

Task prediction loss $\mathcal{L}_T$. We learn the task network by optimizing its loss over both synthetic images x_s and their adapted version $G(x_s, \theta_G)$. This assumes the generator is label-preserving, i.e., that y_s can be used as a label for both images. Thanks to our joint objective, this assumption is directly encouraged during the learning of the generator through the estimation of the scene geometry (depth) carried out by θ_P. Naturally, different tasks require different loss functions. In our

experiments, we consider the task of semantic segmentation and use the standard cross-entropy loss (Eq. 4) over images of size $W \times H$ and a probability distribution over C semantic categories. The total combined loss in the special case of semantic segmentation is therefore defined as:

$$\mathcal{L}_T(T, G) = \mathcal{L}_{\text{CE}}(x_s, y_s) + \mathcal{L}_{\text{CE}}(G(x_s; \theta_G), y_s) \tag{3}$$

$$\mathcal{L}_{\text{CE}}(x, y) = \frac{-1}{WH} \sum_{u,v}^{W,H} \sum_{c=1}^{C} \mathbb{1}_{[c=y_{u,v}]} \log(T(x; \theta_T)_{u,v}) \tag{4}$$

where $\mathbb{1}_{[a=b]}$ is the indicator function.

PI regularization $\mathcal{L}_P$. Similarly, the auxiliary task of predicting PI also requires different losses depending on the type of PI. In our experiments, we use depth from the z-buffer and an ℓ_1-norm:

$$\begin{aligned}\mathcal{L}_P(P, G) = &||P(x_s; \theta_P) - z_s||_1 \\ &+ ||P(G(x_s; \theta_G); \theta_P) - z_s||_1\end{aligned} \tag{5}$$

Perceptual regularization $\mathcal{L}_{\text{perc}}$. To maintain the semantics of the source images in the generated images, we additionally use the perceptual loss Johnson et al. (2016); Chen & Koltun (2017):

$$\mathcal{L}_{\text{perc}}(G) = ||\phi(x_s) - \phi(G(x_s; \theta_G))||_1 \tag{6}$$

where ϕ is a mapping from image space to a pre-determined feature space Chen & Koltun (2017) (see 4.1 for more details).

Optimization. In practice, we follow the standard adversarial training strategy to optimize our joint learning objective (Eq. 1). We alternate between updates to the parameters of the discriminator θ_D, keeping all other parameters fixed, then fix θ_D and optimize the parameters of the generator θ_G, the privileged network θ_P, and most importantly the task network θ_T. We discuss the details of our implementation, including hyper-parameters, in section 4.1.

4 EXPERIMENTS

We evaluate our unsupervised domain adaptation method on the task of semantic segmentation in real-world domain for which training labels are not available.

As our source synthetic domain, we select the public SYNTHIA dataset (Ros et al., 2016) as synthetic source domain given the availability of automatic annotations and PI. SYNTHIA is a dataset generated from an autonomous driving simulator of urban scenes. These images were generated under different weathers and illumination conditions to maximize visual variability. Pixel-wise segmentation and depth labels are provided for each image. In our experiment, we use the sequence of SYNTHIA-RAND-CITYSCAPES, which contains semantic segmentation labels that are more compatible to Cityscapes. For target real-world domains, we use the Cityscapes (Cordts et al., 2016) and Mapillary Vistas (Neuhold et al., 2017) datasets. Cityscapes is one of most widely used real-world urban scene image segmentation datasets with images collected around urban streets in Europe. For this dataset, We use the standard split for training and validation with $2,975$ and 500 images respectively. Mapillary Vistas dataset is a more challenging dataset with a wider variety of images the aspect of camera, geometry distribution, weather and illumination condition, and etc. We use $16,000$ images for training and $2,000$ images for evaluation. During training, none of the labels from real-world domain are used.

In our experiment, We first evaluate adaptation from SYNTHIA to Cityscapes on 16 classes, following the standard evaluation protocol used in Hoffman et al. (2016); Zhang et al. (2017); Saito et al. (2017); Sankaranarayanan et al. (2018); Chen et al. (2018). Then we show the positive impact of using PI by conducting ablation study with and without PI (depth) during adaptation from SYNTHIA to both Cityscapes and Vistas, on a common 7 categories ontology. To be consistent with the semantic segmentation best practices, we use standard intersection-over-union (IoU) per category and mean intersection-over-union (mIoU) as our main validation metric.

4.1 IMPLEMENTATION DETAILS

We adapt the generator and discriminator model architectures from CycleGAN (Zhu et al., 2017) and (Johnson et al., 2016). For simplicity, we use a single sim-to-real generator (no cycle consistency) consisting of two down-sampling convolution layers, nine ResNet blocks (He et al., 2016) and two fractionally-strided convolution layers. Our discriminator is a PatchGAN (Isola et al., 2017) network with 3 layers. We use the standard FCN8s architecture Long et al. (2015) for both the task predictor T and the privileged network P, given its ease of training and its acceptance in domain adaptation works Hoffman et al. (2016). For the perceptual loss $\mathcal{L}_{\text{perc}}$, we follow the implementation in Chen & Koltun (2017). The feature is constructed by the concatenation of the activations of a pre-trained VGG19 network Witten et al. (2016) of layers "conv1_2", "conv2_2", "conv3_2", "conv4_2", "conv5_2". The weights in our joint adversarial loss (Eq. 1) are set to $\alpha = 1$, $\beta = 0.5$, $\gamma = 0.1$, $\delta = 0.33$, for the GAN, task, privileged, and perceptual objectives respectively. These hyper-parameters were found through grid search using a small validation set.

Another critical hyper-parameter for unsupervised learning is the stop criteria. We experimentally observed that the stabilizing effects of the task and privileged losses (Eqs. 3,5) on the GAN objective (Eq. 2) made a simple rule effective for early stopping decision. The iteration when the discriminator loss is significantly and consistently better than the generator loss. This supports our intuition that effective adaptation of the task network might not always be linked to the best image generator. Images are resized to 360×640 during training and evaluation with the objective of reducing training times. A crop size of 320×320 is used. In all adversarial learning cases, we do five steps of the generator for every step of the other networks. The Adam optimizer (Kingma & Ba, 2014) is used to adjust all parameters in our PyTorch implementation (Paszke et al., 2017).

4.2 RESULTS AND DISCUSSION

In this section we present our evaluation of the SPIGAN algorithm in the context of adapting a semantic segmentation network from SYNTHIA to Cityscapes. Depth maps from SYNTHIA are used as PI in the proposed algorithm.

We compare our results to several state-of-art domain adaptation algorithms, including FCNs in the wild (Hoffman et al., 2016), Curriculum DA (Zhang et al., 2017), and Learning from synthetic data (LSD) (Sankaranarayanan et al., 2018).As the task of semantic segmentation tends to be sensitive to training resolution, we report results from all methods at a resolution of 320×640 to make a fair comparison with SPIGAN.

Quantitative results for these methods are shown in Table 1 for the semantic segmentation task on the target domain of Cityscapes (validation set). As reference baselines we include results training only on source images and non-adapted labels. We also provide our algorithm performance without the PI for comparison (i.e., $\gamma = 0$ in Eq. 1, named "SPIGAN-no-PI").

Results show that on Cityscapes SPIGAN achieves state-of-the-art semantic segmentation adaptation in terms of mean IoU. A finer analysis of the results attending to individual classes suggests that the use of PI helps to estimate layout-related classes such as road and sidewalk and object-related classes such as person, rider, car, bus and motorcycle. SPIGAN achieves an improvement of 3% in mean IoU with respect to the non-PI method. This improvement is because of the regularization provided by $P(x; \theta_P)$ during training, which decreases the number of artifacts as shown in Figure 3. This comparison also justify the main contribution of this algorithm, which provides a general approach to leverage synthetic data and PI from the simulator to improve the adaptation process.

4.3 ABLATION STUDY

To better understand the proposed algorithm, and the impact of PI, we conduct further experiments comparing SPIGAN (with PI) and the SPIGAN-no-PI (without PI). We also include results on the Vistas dataset, which presents a more challenging adaptation problem due to the higher diversity of its images. For these experiments, we use a 7 semantic classes ontology to produce a balanced ontology common to the three datasets (SYNTHIA, Cityscapes and Vistas). Adaptation results for both target domains are given in Table 2.

Method	road	sidewalk	building	wall	fence	pole	T-light	T-sign	vegetation	sky	person	rider	car	bus	motorcycle	bicycle	mIoU
FCNs in the wild source only	6.4	17.7	29.7	1.2	0.0	15.1	0.0	7.2	30.3	66.8	51.1	1.5	47.3	3.9	0.1	0.0	17.4
FCNs in the wild	11.5	19.6	30.8	**4.4**	0.0	20.3	0.1	11.7	42.3	68.7	51.2	3.8	54.0	3.2	0.2	0.6	20.2
Curriculum DA source only	5.6	11.2	59.6	0.8	**0.5**	21.5	8.0	5.3	72.4	75.6	35.1	9.0	23.6	4.5	0.5	**18.0**	22.0
Curriculum DA(I+SP)	65.2	26.1	**74.9**	0.1	**0.5**	10.7	3.7	3.0	76.1	70.6	**47.1**	8.2	43.2	20.7	0.7	13.1	29.0
LSD source only	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	23.2
LSD	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	34.5
SPIGAN source only	14.2	12.4	73.3	0.5	0.1	**23.1**	5.8	**12.2**	**79.4**	**78.6**	45.1	7.8	32.8	5.7	5.6	1.6	24.9
SPIGAN-no-PI	68.8	24.5	73.4	3.6	0.1	22.0	5.8	8.9	74.6	77.3	41.5	8.8	58.2	15.4	6.7	8.4	31.1
SPIGAN	**80.5**	**38.9**	73.4	1.8	0.3	20.3	**7.9**	11.3	77.1	77.6	46.6	**13.2**	**63.8**	**22.8**	**8.8**	11.2	**34.7**

Table 1: Semantic Segmentation Task Adapting from SYNTHIA to Cityscapes. We present semantic segmentation results with per-class IoU and mean IoU (mIoU). The highest IoU for each class within all the compared algorithms are highlighted with bold font.

Method	flat	const.	object	nature	sky	human	vehicle	mIoU	Neg. Rate
FCN Source (Cityscapes)	79.6	51.0	8.7	29.0	50.9	3.0	31.6	36.3	–
SPIGAN-no-PI (Cityscapes)	90.3	58.2	6.8	35.8	69.0	9.5	52.1	46.0	0.16
SPIGAN (Cityscapes)	**91.2**	**66.4**	**9.6**	**56.8**	**71.5**	**17.7**	**60.3**	**53.4**	**0.09**
FCN Source (Vistas)	61.5	40.8	10.4	53.3	65.7	16.6	30.4	39.8	–
SPIGAN-no-PI (Vistas)	53.0	30.8	3.6	14.6	53.0	5.8	26.9	26.8	0.80
SPIGAN (Vistas)	**74.1**	**47.1**	**6.8**	**43.3**	**83.7**	**11.2**	**42.2**	**44.1**	**0.42**

Table 2: Semantic Segmentation results (per category and mean IoUs, higher is better) for SYNTHIA adapting to Cityscapes and Vistas. The last column is the ratio of images in the validation set for which we observe negative transfer (lower is better).

In addition to the conventional segmentation performance metrics, we also carried out a study to measure the amount of negative transfer, summarized in Table 2. A negative transfer case is defined as a real-world testing sample that has a mIoU lower than the FCN source prediction (no adaptation).

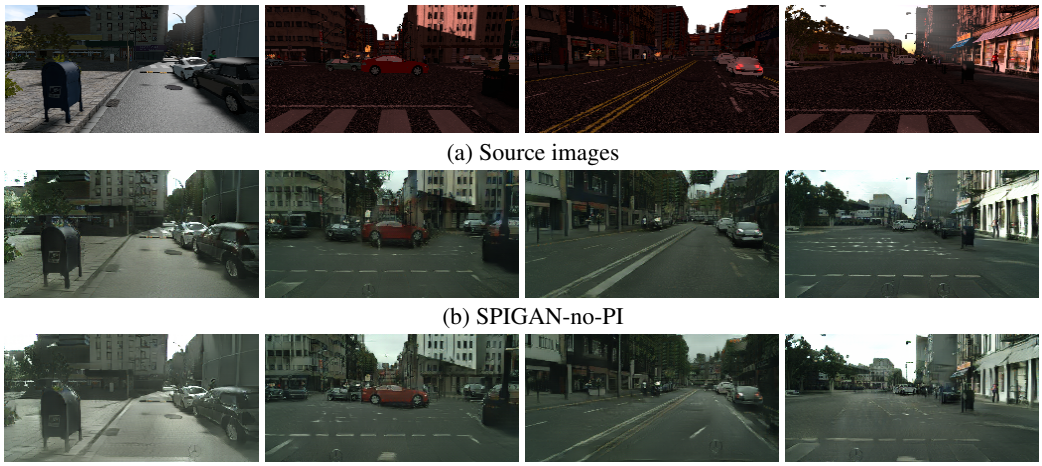
For Cityscapes, the quantitative results in Table 2 show that SPIGAN is able to provide dramatic adaptation as hypothesized. The mean IoU improves by 17.1% with the PI providing an improvement of 7.4%. This is consistent with our observation in the previous experiment (Table 1). We also notice that SPIGAN gets significant improvements on "nature", "construction", and "vehicle" categories. In addition, SPIGAN is able to improve the IoU by $+15\%$ on the "human" category, a difficult class in semantic segmentation. We provide examples of qualitative results for the adaptation from SYNTHIA to Cityscapes in Figure 3 and Figure 5.

Vistas dataset presents a larger domain gap from SYNTHIA due to a greater difference in viewpoint and a larger variety of scenarios. The domain adaptation becomes more challenging as we can tell from both quantitative results in Table 2 and qualitative visualizations in Figure 4 and Figure 6. Apparently, SPIGAN-no-PI suffers from negative transfer and under performs FCN-source for the semantic segmentation task. However, SPIGAN is still able to decrease the domain gap and improves the segmentation mean IoU by 4.3%. The consistent improvement brought by PI in both of the experiments not only shows that PI imposes useful constraints that promote better task-oriented training, but also implies that PI guides the training to reducing domain shift.

By comparing the results on the two different datasets, we also found that all the unsupervised adaptation methods share some similarity in the performance of certain categories. For instance, the "vehicle" category has seen the largest improvement for both Cityscapes and Vistas. This trend is consistent with the well-known fact that "objects" categories are easier to adapt than "stuff" Vazquez et al. (2014). However, the same improvement did not appear in the "human" category mainly due to the fact that the SYNTHIA subset we used in our experiment contains very few humans. This phenomenon has been recently studied in Sadat Saleh et al. (2018).

5 CONCLUSION

We present SPIGAN, a novel method to leverage synthetic data and Privileged Information (PI) available in simulated environments to perform unsupervised domain adaptation of deep networks. Our approach jointly learns a generative pixel-level adaptation network together with a target task network and privileged information models. We showed that our approach is able to partially close large generalization gaps between models trained on the source domain only and models trained in a fully-supervised fashion on the target domain, including for challenging real-world tasks like semantic segmentation of urban scenes. For future work, we plan to investigate SPIGAN applied to additional tasks, with different types of PI that can be obtained from simulation.

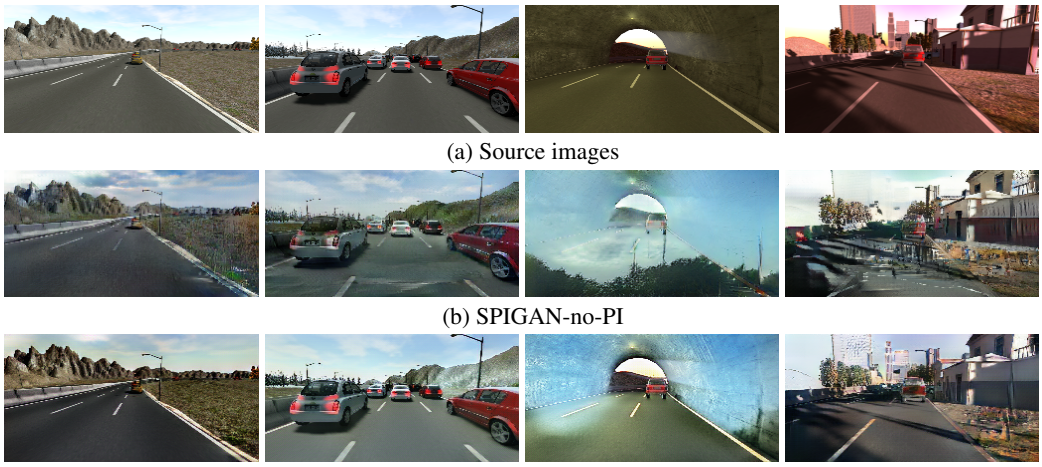


(a) Source images

(b) SPIGAN-no-PI

(c) SPIGAN

Figure 3: Adaptation from SYNTHIA to Cityscapes. Examples of source domain images (a) are shown. (b) Source images after the adaptation process w/o Privileged Information. (c) Source images after the adaptation process using SPIGAN.

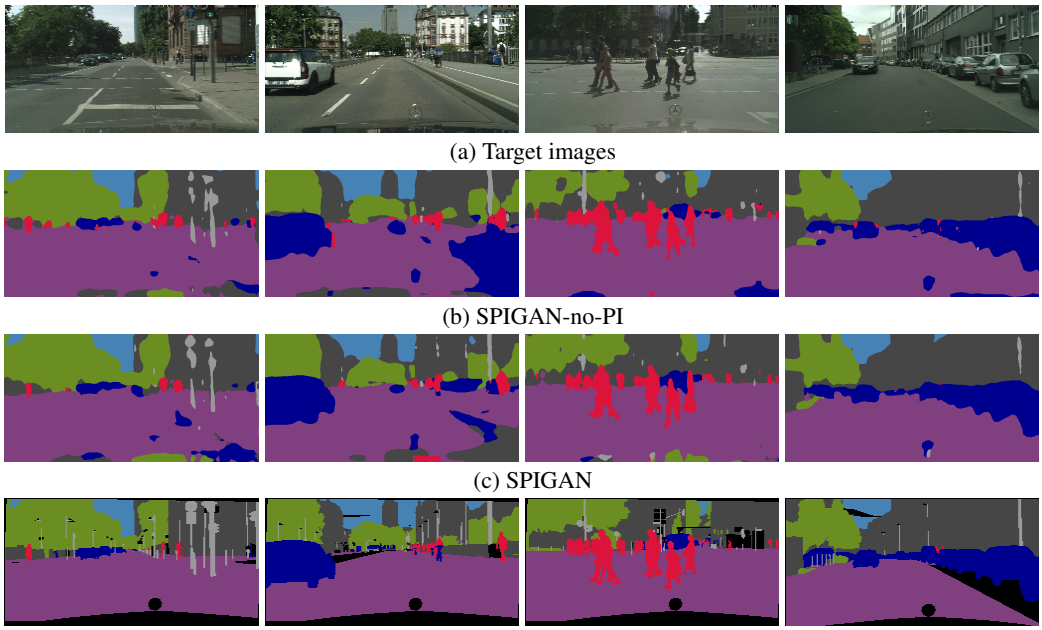


(a) Source images

(b) SPIGAN-no-PI

(c) SPIGAN

Figure 4: Adaptation from SYNTHIA to Vistas. Examples of source domain images (a) are shown. (b) Source images after the adaptation process w/o Privileged Information. (c) Source images after the adaptation process using SPIGAN. Image adaptation is more challenging between these two datasets due to the larger domain gap brought by camera view points. Qualitative results indicate that SPIGAN is encoding more regularization in the image generation.



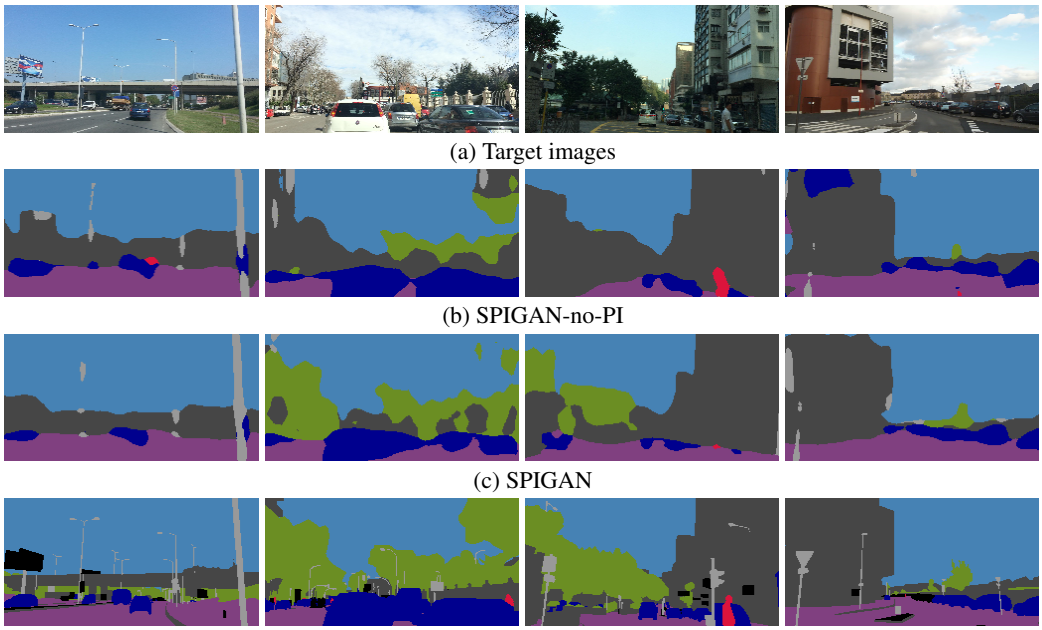
(a) Target images

(b) SPIGAN-no-PI

(c) SPIGAN

(d) Ground truth

Figure 5: Semantic segmentation results on Cityscapes. For a set of real images (a) we show examples of predicted semantic segmentation masks. SPIGAN predictions (c) are more accurate (i.e., closer to the ground truth (d)) than those produced without PI during training (b).



(a) Target images

(b) SPIGAN-no-PI

(c) SPIGAN

(d) Ground truth labels

Figure 6: Semantic segmentation results on Vistas. For a set of real images (a) we show examples of predicted semantic segmentation masks. SPIGAN predictions (c) are more accurate (i.e., closer to the ground truth (d)) than those produced without PI during training (b).

REFERENCES

Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *CVPR*, 2017.

Lin Chen, Wen Li, and Dong Xu. Recognizing rgb images by learning from rgb-d data. In *CVPR*, 2014.

Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *ICCV*, Oct 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 *ICCV*, 2017.

Yuhua Chen, Wen Li, and Luc Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *CVPR*, 2018.

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 *CVPR*, 2016.

Gabriela Csurka. *Domain Adaptation in Computer Vision Applications*. Springer, 2017.

César Roberto de Souza, Adrien Gaidon, Yohann Cabon, and Antonio Manuel López Peña. Procedural generation of videos to train deep action recognition networks. In *CVPR*, 2017.

Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *CVPR*, 2016.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 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*, 2014.

Ankur Handa, Viorica Patraucean, Vijay Badrinarayanan, Simon Stent, and Roberto Cipolla. Understanding real world indoor scenes with synthetic data. In *CVPR*, 2016.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Samitha Herath, Mehrtash Harandi, and Fatih Porikli. Learning an invariant hilbert space for domain adaptation. In *CVPR*, 2017.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Judy Hoffman, Erik Rodner, Jeff Donahue, Kate Saenko, and Trevor Darrell. Efficient learning of domain-invariant image representations. In *ICLR*, 2013.

Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*. Springer, 2016.

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

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Maksim Lapin, Matthias Hein, and Bernt Schiele. Learning using privileged information: Svm+ and weighted svm. *Neural Networks*, 2014.

Wen Li, Li Niu, and Dong Xu. Exploiting privileged information from web data for image categorization. In *ECCV*. Springer, 2014.

Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *arXiv preprint arXiv:1703.00848*, 2017.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015.

David Lopez-Paz, Lon Bottou, Bernhard Schlkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, and Zhen Wang. Multi-class generative adversarial networks with the l2 loss function. *arXiv preprint arXiv:1611.04076*, 2016.

Javier Marín, David Vázquez, David Gerónimo, and Antonio M. López. Learning appearance in virtual scenarios for pedestrian detection. In *CVPR*, 2010.

Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, 2016.

Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017.

Jeremie Papon and Markus Schoeler. Semantic pose using deep networks trained on synthetic RGB-D. In *ICCV*, 2015.

Adam Paszke, Sam Gross, Soumith Chintala, et al. Pytorch, 2017.

Xingchao Peng, Baochen Sun, Karim Ali, and Kate Saenko. Learning deep object detectors from 3D models. In *ICCV*, 2015.

Sanjay Purushotham, Wilka Carvalho, Tanachat Nilanon, and Yan Liu. Variational recurrent adversarial deep domain adaptation. In *ICLR*, 2017.

Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Koltun Vladlen. Playing for data: Ground truth from computer games. In *ECCV*, 2016.

German Ros, Laura Sellart, Joanna Materzyska, David Vázquez, and Antonio M. López. The SYNTHIA dataset: a large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016.

Fatemeh Sadat Saleh, Mohammad Sadegh Aliakbarian, Mathieu Salzmann, Lars Petersson, and Jose M. Alvarez. Effective use of synthetic data for urban scene semantic segmentation. In *ECCV*, 2018.

Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. *arXiv preprint arXiv:1712.02560*, 3, 2017.

Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *CVPR*, 2018.

Nikolaos Sarafianos, Michalis Vrigkas, and Ioannis A Kakadiaris. Adaptive svm+: Learning with privileged information for domain adaptation. *arXiv preprint arXiv:1708.09083*, 2017.

Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipmanand, and Andrew Blake. Real-time human pose recognition in parts from a single depth image. In *CVPR*, 2011.

Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Josh Susskind, Wenda Wang, and Russ Webb. Learning from simulated and unsupervised images through adversarial training. *arXiv preprint arXiv:1612.07828*, 2016.

Hao Su, Charles R. Qi, Yangyan Yi, and Leonidas Guibas. Render for CNN: viewpoint estimation in images using CNNs trained with rendered 3D model views. In *ICCV*, 2015.

Baochen Sun and Kate Saenko. From virtual to reality: Fast adaptation of virtual object detectors to real domains. In *BMVC*, 2014.

Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016.

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 2009.

David Vazquez, Antonio M. López, Javier Marín, Daniel Ponsa, and David Gerónimo. Virtual and real world adaptation for pedestrian detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(4):797 – 809, 2014.

Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

Jiaolong Xu, David Vázquez, Antonio M. López, Javier Marín, and Daniel Ponsa. Learning a part-based pedestrian detector in a virtual world. *T-ITS*, 15(5):2121–2131, 2014.

Ke Yan, Lu Kou, and David Zhang. Learning domain-invariant subspace using domain features and independence maximization. In *IEEE Transactions on Cybernetics*, 2017.

Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *ICCV*, 2017.

Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, and Tao Mei. Fully convolutional adaptation networks for semantic segmentation. In *CVPR*, 2018.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.