Fully Convolutional Adaptation Networks for Semantic Segmentation[*]

Yiheng Zhang [†], Zhaofan Qiu [†], Ting Yao [‡], Dong Liu [†], and Tao Mei [‡]

[†] University of Science and Technology of China, Hefei, China

[‡] Microsoft Research, Beijing, China

{yihengzhang.chn, zhaofanqiu}@gmail.com, {tiyao, tmei}@microsoft.com, dongeliu@ustc.edu.cn

Abstract

The recent advances in deep neural networks have convincingly demonstrated high capability in learning vision models on large datasets. Nevertheless, collecting expert labeled datasets especially with pixel-level annotations is an extremely expensive process. An appealing alternative is to render synthetic data (e.g., computer games) and generate ground truth automatically. However, simply applying the models learnt on synthetic images may lead to high generalization error on real images due to domain shift. In this paper, we facilitate this issue from the perspectives of both visual appearance-level and representation-level domain adaptation. The former adapts source-domain images to appear as if drawn from the "style" in the target domain and the latter attempts to learn domain-invariant representations. Specifically, we present Fully Convolutional Adaptation Networks (FCAN), a novel deep architecture for semantic segmentation which combines Appearance Adaptation Networks (AAN) and Representation Adaptation Networks (RAN). AAN learns a transformation from one domain to the other in the pixel space and RAN is optimized in an adversarial learning manner to maximally fool the domain discriminator with the learnt source and target representations. Extensive experiments are conducted on the transfer from GTA5 (game videos) to Cityscapes (urban street scenes) on semantic segmentation and our proposal achieves superior results when comparing to state-of-the-art unsupervised adaptation techniques. More remarkably, we obtain a new record: mIoU of 47.5% on BDDS (drive-cam videos) in an unsupervised setting.

1. Introduction

Deep Neural Networks have successfully proven highly effective for learning vision models on large-scale datasets. To date in the literature, there are various datasets (e.g., ImageNet [26] and COCO [14]) that include well-annotated

---

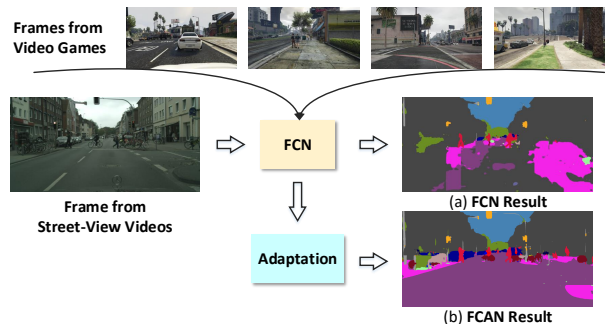[*]This work was performed at Microsoft Research Asia.

Figure 1. Semantic segmentation on one example frame in street-view videos by (a) directly applying FCN trained on images from video games and (b) domain adaptation of FCAN in this work.

images available for developing deep models to a variety of vision tasks, e.g., recognition [8, 27, 29], detection [6, 24], captioning [34] and semantic segmentation [1, 16]. Nevertheless, given a new dataset, the typical solution is still to perform intensive manual labeling despite expensive efforts and time-consuming process. An alternative is to utilize synthetic data which is largely available from computer games [25] and the ground truth could be freely generated automatically. However, many previous experiences have also shown that reapplying a model learnt on synthetic data may hurt the performance in real data due to a phenomenon known as "domain shift" [35]. Take the segmentation results of one frame from real street-view videos in Figure 1 (a) as an example, the model trained on synthetic data from video games fails to properly segment the scene into semantic categories such as road, person and car. As a result, unsupervised domain adaptation would be desirable on addressing this challenge, which aims to utilize labeled examples from the source domain and a large number of unlabeled examples in the target domain to reduce a prediction error on the target data.

A general practice in unsupervised domain adaptation is to build invariance across domains by minimizing the measure of domain shift such as correlation distances [28] or maximum mean discrepancy [32]. We novelly consider the problem from the viewpoint of both appearance-level and representation-level invariance. The objective of

appearance-level invariance is to recombine the image content in one domain with the "style" from the other domain. As such, the images in two domains appear as if they are drawn from the same domain. In other words, the visual appearances tend to be domain-invariant. The inspiration of representation-level invariance is from the advances of adversarial learning for domain adaptation, which is to model domain distribution via an adversarial objective with respect to a domain discriminator. The spirit behind is from generative adversarial learning [7], that trains two models, i.e., a generative model and a discriminative model, by pitting them against each other. In the context of domain adaptation, this adversarial principle is then equivalent to guiding the representation learning in both domains, making the difference between source and target representation distributions indistinguishable through the domain discriminator. We follow this elegant recipe and capitalize on adversarial mechanism to learn image representation that is invariant across domains. In this work, we are particularly investigating the problem of domain adaptation on semantic segmentation task which relies on probably the most accurate pixel-level annotations.

By consolidating the idea of appearance-level and representation-level invariance into unsupervised domain adaption for enhancing semantic segmentation, we present a novel Fully Convolutional Adaptation Networks (FCAN) architecture, as shown in Figure 2. The whole framework consists of Appearance Adaptation Networks (AAN) and Representation Adaptation Networks (RAN). Ideally, AAN is to construct an image that captures high-level content in a source image and low-level pixel information of the target domain. Specifically, AAN starts with a white noise image and adjusts the output image by using gradient descent to minimize the Euclidean distance between the feature maps of the output image and those of the source image or mean feature maps of the images in target domain. In RAN, a shared Fully Convolutional Networks (FCN) is first employed to produce image representation in each domain, followed by bilinear interpolation to upsample the outputs for pixel-level classification, and meanwhile a domain discriminator to distinguish between source and target domain. An Atrous Spatial Pyramid Pooling (ASPP) strategy is particularly devised to enlarge the field of view of filters in feature map and endow the domain discriminator with more power. RAN is trained by optimizing two losses, i.e., classification loss to measure pixel-level semantics and adversarial loss to maximally fool the domain discriminator with the learnt source and target representations. With both appearance-level and representation-level adaptations, our FCAN could better build invariance across domains and thus obtain encouraging segmentation results in Figure 1 (b).

The main contribution of this work is the proposal of Fully Convolutional Adaptation Networks for addressing the issue of semantic segmentation in the context of domain adaptation. The solution also leads to the elegant views of what kind of invariance should be built across domains for adaptation and how to model the domain invariance in a deep learning framework especially for the task of semantic segmentation, which are problems not yet fully understood in the literature.

2. Related Work

We briefly group the related works into two categories: semantic segmentation and deep domain adaptation.

Semantic segmentation is one of the most challenging tasks in computer vision, which attempts to predict pixel-level semantic labels of the given image or video frame. Inspired by the recent advance of Fully Convolutional Networks (FCN) [16], there have been several techniques, ranging from multi-scale feature ensemble (e.g., Dilated Convolution [36], RefineNet [13], DeepLab [1] and HAZNet [33]) to context information preservation (e.g., ParseNet [15], P-SPNet [37] and DST-FCN [23]), being proposed. The original FCN formulation could also be improved by exploiting some post processing techniques (e.g., conditional random fields [38]). Moreover, as most semantic segmentation methods rely on the pixel-level annotations which require extremely expensive labeling efforts, researchers have also strived to leverage weak supervision instead (e.g., instance-level bounding boxes [3], image-level tags [22]) for semantic segmentation task. To achieve this target, the techniques such as multiple instance learning [20], EM algorithm [18] and constrained CNN [19] are exploited in the literature. An alternative in [10] utilizes the pixel-level annotations from auxiliary categories to generalize semantic segmentation to categories where only image-level labels are available. The goal of this work is to study the exploration of freely accessible synthetic data with annotations and largely unlabeled real data for annotating real images on the pixel level, which is an emerging research area.

Deep Domain adaptation aims to transfer model learnt in a labeled source domain to a target domain in a deep learning framework. The research of this topic has proceeded along three different dimensions: unsupervised adaptation, supervised adaptation and semi-supervised adaptation. Unsupervised domain adaptation refers to the setting when the labeled target data is not available. Deep Correlation Alignment (CORAL) [28] exploits Maximum Mean Discrepancy (MMD) to match the mean and covariance of source and target distributions. Adversarial Discriminative Domain Adaptation (ADDA) [31] optimizes the adaptation model with adversarial training. In contrast, when the labeled target data is available, we refer to the problem as supervised domain adaptation. Tzeng *et al.* [30] utilizes a binary domain classifier and devises the domain confusion loss to encourage the predicted domain labels to be u-
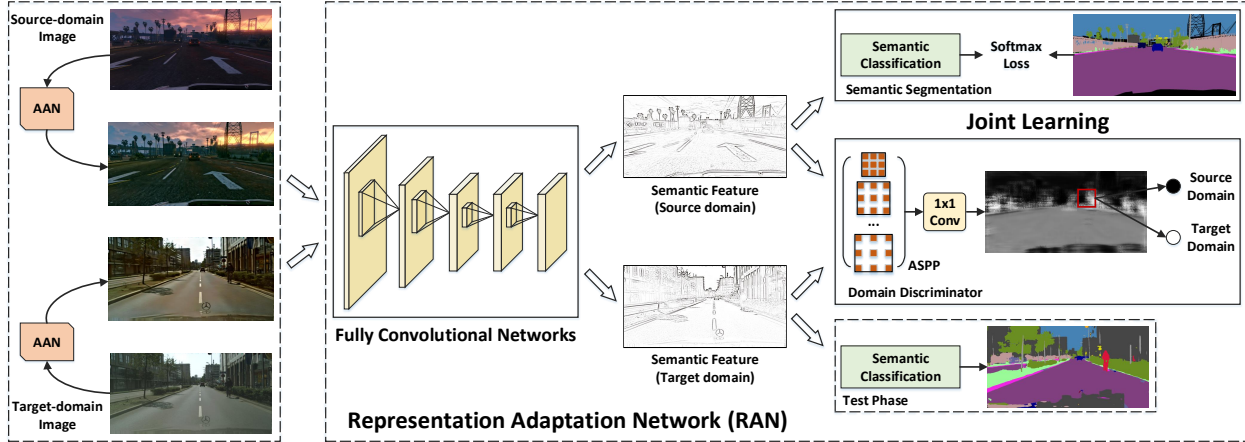
Figure 2. An overview of our Fully Convolutional Adaptation Networks (FCAN) architecture. It consists of two main components: the Appearance Adaptation Networks (AAN) on the left and the Representation Adaptation Networks (RAN) on the right. AAN transfers images from one domain to the other one and thus the visual appearance tends to be domain-invariant. RAN learns domain-invariant representations in an adversarial manner by maximally fooling the domain discriminator with the learnt source and target representations. An extended Atrous Spatial Pyramid Pooling (ASPP) layer is particularly devised to leverage the regions across different scales for enhancing the discriminative capability. RAN is jointly optimized with supervised segmentation loss on source images plus adversarial loss.

niformly distributed. Deep Domain Confusion (DDC) [32] applies MMD as well as the regular classification loss on the source to learn representations that are both discriminative and domain invariant. In addition, semi-supervised domain adaptation methods have also been proposed, which exploit both labeled and unlabeled target data. Deep Adaptation Network (DAN) [17] embeds all task specific layers in a reproducing kernel Hilbert space. Both semi-supervised and unsupervised settings are considered.

In short, our work in this paper mainly focuses on unsupervised adaptation for semantic segmentation task, which is seldom investigated. The most closely related work is the FCNWild [9], which addresses the cross-domain segmentation problem by only exploiting fully convolutional adversarial training for domain adaptation. Our method is different from [9] in that we solve the domain shift from the perspectives of both visual appearance-level and representation-level domain adaptation, which bridges the domain gap in a more principled way.

3. Fully Convolutional Adaptation Networks (FCAN) for Semantic Segmentation

In this section we present our proposed Fully Convolutional Adaptation Networks (FCAN) for semantic segmentation. Figure 2 illustrates the overview of our framework. It consists of two main components: the Appearance Adaptation Networks (AAN) and the Representation Adaptation Networks (RAN). Given the input images from two domains, AAN is first utilized to transfer images from one domain to the other from the perspective of visual appearance. By recombining the image content in one domain with the "style" from the other one, the visual appearance tends to be

domain-invariant. We take the transformation from source to target as an example in this section, and the other options will be elaborated in our experiments. On the other hand, RAN learns domain-invariant representations in an adversarial manner and a domain discriminator is devised to classify which domain the image region corresponding to the receptive field of each spatial unit in the feature map comes from. The objective of RAN is to guide the representation learning in both domains, making the source and target representations indistinguishable through the domain discriminator. As a result, our FCAN addresses domain adaptation problem from the viewpoint of both visual appearance-level and representation-level domain invariance and is potentially more effective at undoing the effects of domain shift.

3.1. Appearance Adaptation Networks (AAN)

The goal of AAN is to make the images from different domains visually similar. In other words, AAN tries to adapt the source images to appear as if drawn from the target domain. To achieve this, the low-level features over all the images in target domain should be separated and regarded as the "style" of target domain, as these features encode the low-level forms of the images, e.g., texture, lighting and shading. In contrast, the high-level content in terms of objects and their relations in the source image should be extracted and recombined with the "style" of target domain to produce an adaptive image.

Figure 3 illustrates the architecture of AAN. Given a set of images $\mathcal{X}_t = \{x_t^i | i = 1, \dots, m\}$ in target domain and one image from source domain x_s, we begin with a white noise image and iteratively render this image with the semantic content in x_s plus the "style" of $\mathcal{X}_t$ to produce an adaptive image x_o. Specifically, a pre-trained CNN is u-
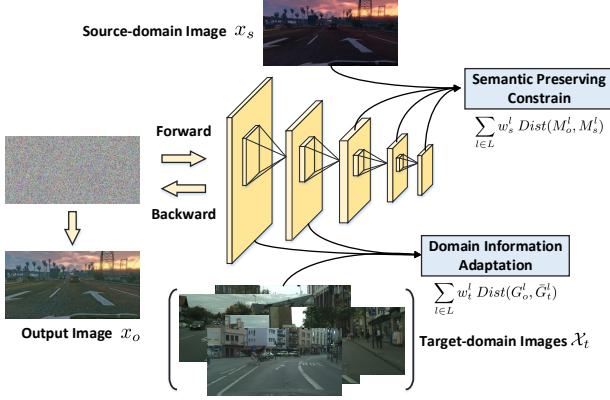
Figure 3. The architecture of Appearance Adaptation Networks (AAN). Given the target image set $\mathcal{X}_t$ and one source image x_s, we begin with a white noise image and adjust it towards an adaptive image x_o, which appears as if it is drawn from target domain but contains semantic content in the source image. A pre-trained CNN is utilized to extract feature maps. The high-level image content of x_s is preserved by minimizing the distance between feature maps of x_s and x_o, while the style of target domain is kept by minimizing the distance between feature correlations of x_o and $\mathcal{X}_t$.

tilized to extract feature maps for each image. Suppose every convolutional layer l in the CNN has N_l response maps, where N_l is the number of channels, and the size of each response map is $H_l \times W_l$, where H_l and W_l denotes the height and width of the map, respectively. As such, the feature maps in the layer l could be represented as $M^l \in \mathbb{R}^{N_l \times H_l \times W_l}$. Basically the responses in different convolutional layers characterize image content on different semantic level, where deeper layer responds to higher semantics. To better govern the semantic content in source image x_s, different weights are assigned to different layers to reflect the contribution of each layer. The objective function is then formulated as

$$\min_{x_o} \sum_{l \in L} w_s^l \, Dist(M_o^l, M_s^l) \,, \tag{1}$$

where L is the set of layers to be considered for measurement. w_s^l is the weight of layer l, M_o^l and M_s^l is the feature map of layer l on x_o and x_s, respectively. By minimizing the Euclidean distance in Eq.(1), the image content in x_s is expected to be preserved in the adaptive image x_o.

Next, the "style" of one image is in general treated as a kind of statistical measurement or *pattern*, which is agnostic to spatial information [4]. In CNN, one of such statistical measurements is the correlations between different response maps. Hence, the "style" of an image G^l on layer l could be computed by

$$G^{l,ij} = M^{l,i} \odot M^{l,j} \,. \tag{2}$$

$G^{l,ij}$ is the inner product between the vectorized i-th and j-th response map of M^l. In our case, we extend the "style" of

one image to that of one domain ($\bar{G}_t^l$ of the target domain) by averaging G^l over all the images in target domain. In order to synthesize the "style" of target domain into x_o, we formulate the objective in each layer as

$$\min_{x_o} \sum_{l \in L} w_t^l \, Dist(G_o^l, \bar{G}_t^l) \,, \tag{3}$$

where w_t^l is the weight for layer l. Finally, the overall loss function $\mathcal{L}_{AAN}$ to be minimized is

$$\mathcal{L}_{AAN}(x_o) = \sum_{l \in L} w_s^l \, Dist(M_o^l, M_s^l) + \alpha \sum_{l \in L} w_t^l \, Dist(G_o^l, \bar{G}_t^l) \,, \tag{4}$$

where α is the weight to balance semantic content in the source image and the style of target domain. In the training, similar to [5], AAN adjusts the output image by backpropagating the gradients derived from Eq. (4) to x_o, resulting in the domain-invariant appearance.

3.2. Representation Adaptation Networks (RAN)

With the Appearance Adaptation Networks, the images from different domains appear to be from the same domain. To further reduce the impact of domain shift, we attempt to learn domain-invariant representations. Consequently, Representation Adaptation Networks (RAN) is designed to adapt representations across domains, which is derived from the idea of adversarial learning [7]. The adversarial principle in our RAN is equivalent to guiding the learning of feature representations in both domains by fooling a domain discriminator D with the learnt source and target representations. Specifically, RAN first utilizes a shared Fully Convolutional Network (FCN) to extract the representations of images or adaptive images through AAN from both domains. This FCN model F here aims to learn indistinguishable image representations across two domains. Furthermore, the discriminator D attempts to differentiate between source and target representations, whose outputs are the domain prediction of each image region that corresponds to the spatial unit in the final feature map. Formally, given the training set $\mathcal{X}_s = \{x_s^i | i = 1, \ldots, n\}$ in source domain and $\mathcal{X}_t = \{x_t^i | i = 1, \ldots, m\}$ in target domain, the adversarial loss $\mathcal{L}_{adv}$ is the average classification loss over all spatial units, which is formulated as

$$\mathcal{L}_{adv}(\mathcal{X}_s, \mathcal{X}_t) = -E_{x_t \sim \mathcal{X}_t} [\frac{1}{Z} \sum_{i=1}^{Z} log(D_i(F(x_t)))]$$
$$- E_{x_s \sim \mathcal{X}_s} [\frac{1}{Z} \sum_{i=1}^{Z} log(1 - D_i(F(x_s)))] \,, \tag{5}$$

where Z is the number of spatial units in the output of D. Similar to the standard GANs, the adversarial training of our RAN is to optimize the following minimax function

$$\max_F \min_D \mathcal{L}_{adv}(\mathcal{X}_s, \mathcal{X}_t) \,. \tag{6}$$

Given the fact that there are many different objects of various size in real data, we further take the utilization of multi-scale representations into account to enhance the adversarial learning. One traditional multi-scale strategy is to resize the images with multiple resolutions, which indeed improves the performance but at the cost of large computation. In this work, we extend Atrous Spatial Pyramid Pooling (ASPP) [1] to implement this, as shown in Figure 2. Specifically, k dilated convolutional layers with different sampling rates are exploited in parallel to produce k feature representations on the output of FCN independently, each with c feature channels. All the feature channels are then stacked up to form a new feature map with ck channels, followed by a 1×1 convolutional layer plus a sigmoid layer to generate the final score map. Each spatial unit in the score map presents the probability of the corresponding image region belonging to the target domain. In addition, we simultaneously optimize the standard pixel-level classification loss $\mathcal{L}_{seg}$ for supervised segmentation on the images from source domain, where the labels are available. Hence, the overall objective of RAN integrates $\mathcal{L}_{seg}$ and $\mathcal{L}_{adv}$ as

$$\max_{F} \min_{D} \left\{ \mathcal{L}_{adv}(\mathcal{X}_s, \mathcal{X}_t) - \lambda \mathcal{L}_{seg}(\mathcal{X}_s) \right\} , \quad (7)$$

where λ is the tradeoff parameter. Through fooling the domain discriminator with the source and target representations, our RAN is able to produce domain-invariant representations. In test stage, the images in target domain are fed into the learnt FCN to produce representations for pixel-level classification.

4. Implementation

4.1. Appearance Adaptation

We adopt the pre-trained ResNet-50 [8] architecture as the basic CNN. In particular, we only include the five convolutional layers in the set, i.e., $L = \{conv1, res2c, res3d, res4f, res5c\}$, as the representations of these layers in general have the highest capability in each scale. The weights w_s^l and w_t^l of layers for the images in source and target domain are generally determined on the visual appearances of adaptive images. In addition, when optimizing Eq. (4), a common problem is the need to set the tradeoff parameter α to balance content and "style." As the ultimate goal is to semantically segment each pixel in the images, it is required to preserve the semantic content precisely. As a result, the impact of "style" is regarded as only a "delta" function to adjust the appearance and we empirically set a small weight of $\alpha = 10^{-14}$ for this purpose. The number of maximum iteration I is fixed to $1k$. In each iteration i, the image x_o is updated by $x_o^i = x_o^{i-1} - w^{i-1} \frac{g^{i-1}}{\|g^{i-1}\|_1}$, where $g^{i-1} = \frac{\partial \mathcal{L}_{app}(x_o^{i-1})}{\partial x_o^{i-1}}$, $w^{i-1} = \beta \frac{I-i}{I}$ and $\beta = 10$.

4.2. Representation Adaptation

In our implementations, we employ dilated fully convolutional network [1] originated from ResNet-101 [8] as our FCN, which has proven to be effective on generating powerful representations for semantic segmentation. The feature maps of the last convolutional layer (i.e., $res5c$) are fed into both segmentation and adversarial branches. In supervised segmentation branch, we also augment the outputs of FCN with Pyramid Pooling [37] to integrate contextual prior into representation. In adversarial branch, we use $k = 4$ dilated convolutional layers in parallel to produce multiple feature maps, each with $c = 128$ channels. The sampling rate of different dilated convolution kernel is $1, 2, 3$ and 4, respectively. Finally, a sigmoid layer is utilized next to the ASPP to output the predictions, which are in the range of $[0, 1]$.

4.3. Training Strategy

Our proposal is implemented on Caffe [12] framework and mini-batch stochastic gradient descent algorithm is exploited to optimize the model. We pre-train RAN on source domain with only segmentation loss. The initial learning rate is 0.0025. Similar to [1], we use the "poly" learning rate policy with power fixed to 0.9. Momentum and weight decay is set to 0.9 and 0.0005, respectively. The batch size is 6. The maximum iteration number is $30k$. Then, we fine-tune RAN jointly with segmentation loss and adversarial loss. The tradeoff parameter λ is set to 5. The initial learning rate is 0.0001. The batch size is 8 and the maximum iteration number is $10k$. The rest hyper-parameters are the same with those in pre-training.

5. Experiments

5.1. Datasets

We conduct a thorough evaluation of our FCAN on the domain adaptation from GTA5 [25] (game videos) dataset to Cityscapes (urban street scenes) dataset [2].

The GTA5 dataset contains 24,966 images (video frames) from the game Grand Theft Auto V (GTA5) and the pixel-level ground truth for each image is also created. In the game, the images are captured on the virtual city of Los Santos, which is originated from the city of Los Angeles. The resolution of each image is 1914×1052. There are 19 classes which are compatible with other segmentation datasets for outdoor scenes (e.g., Cityscapes) and utilized in the evaluation. The Cityscapes dataset is one popular benchmark for semantic understanding of urban street scenes, which contains high quality pixel-level annotations of 5,000 images (frames) collected in street scenes from 50 different cities. The image resolution is 2048×1024. Following the standard protocol in segmentation task (e.g., [2]), 19 semantic labels (car, road, person, building, etc.) are used for evaluation. In between, the training, validation,

Table 1. The mIoU performance comparisons between different ways of utilizing AAN.

Train	Validation	FCN	RAN
Src	Tar	29.15	44.81
Src	Tar_Ada	34.68	45.03
Src_Ada	Tar	31.71	**46.21**
Src_Ada	Tar_Ada	36.25	45.59
Late Fusion		37.61	**46.60**

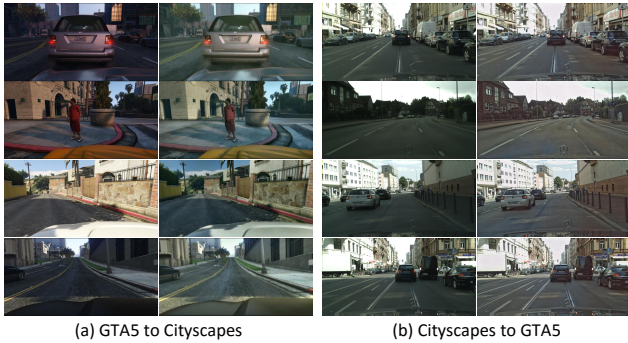(a) GTA5 to Cityscapes   (b) Cityscapes to GTA5

Figure 4. Examples of appearance-level adaptation through AAN.

and test sets contains 2,975, 500, and 1,525 frames, respectively. Following the settings in [9, 21], only the validation set (500 frames) are exploited for validating the unsupervised semantic segmentation in our experiments.

In addition, we also take the Berkeley Deep Driving Segmentation (BDDS) dataset [9] as another target domain for verifying the merit of our FCAN. The BDDS dataset consists of thousands of dashcam video frames with pixel-level annotations, which share compatible label space with Cityscapes. The image resolution is 1280×720. Following the settings in [9, 21], 1,500 frames are used for evaluation.

In all experiments, we adopt the Intersection over Union (IoU) per category and mean IoU over all the categories as the performance metrics.

5.2. Evaluation of AAN

We first examine the effectiveness of AAN on semantic segmentation from two aspects: 1) images from which domain are adapted by AAN, and 2) adaptation by only performing AAN or plus RAN. Source Adaptation (Src_Ada) here is to render source images with the "style" of the target domain, and vice versa for Target Adaptation (Tar_Ada). **FCN** refers to the setting of semantic segmentation by directly exploiting the FCN learnt on source domain to do prediction on target images. In contrast, **RAN** further performs representation-level adaptation by our RAN.

The mIoU performances between different ways of utilizing AAN are summarized in Table 1. Overall, adapting images in source domain through AAN plus RAN achieves the highest mIoU of 46.21%. The results by applying AAN to images in source or target or both domains consistently exhibits better performance than the setting without the use of AAN (the first row) when directly employing FCN

Table 2. Performance contribution of each design in FCAN.

Method	ABN	ADA	Conv	ASPP	AAN	mIoU
FCN						29.15
+ABN	√					35.51
+ADA	√	√				41.29
+Conv	√	√	√			43.17
+ASPP	√	√	√	√		44.81
FCAN	√	√	√	√	√	**46.60**

in segmentation. The results basically indicate the advantage of exploring appearance-level domain adaptation. The performance in each setting is further improved by RAN, indicating that visual appearance-level and representation-level adaptation are complementary to each other. Another observation is that the performance gain of RAN tends to be large when performing AAN on source images. The gain is however decreased when adapting target images by AAN. We speculate that this may be the result of synthesizing some noise into the adapted target images by AAN especially at the boundary of objects and that in turn affects the segmentation stability. Furthermore, when late fusing the score maps of segmentation predicted by the four settings, the mIoU performance could be boosted up to 46.6%. We refer to this fusion version as AAN in the following evaluations unless otherwise stated.

Figure 4 shows four examples of appearance-level transfer for images in source and target domain, respectively. As illustrated in the figure, the semantic content in original images are all well-preserved in the adaptive images. When rendering the images in GTA5 with "style" of Cityscapes, the overall color of the images becomes bleak and the color saturation tends to be low. In contrast, when reversing the transfer direction, the color of images in Cityscapes gets much brighter and with high saturation. The results demonstrate a good appearance-level transfer in between.

5.3. An Ablation Study of FCAN

Next, we study how each design in FCAN influences the overall performance. Adaptive Batch Normalization (**ABN**) simply replaces the mean and variance of BN layer in FCN learnt in source domain with those computed on the images in target domain. Adversarial Domain Adaptation (**ADA**) leverages the idea of adversarial training to learn domain-invariant representations and the domain discriminator judges the domain on image level. When the domain discriminator is extended to classify each image region, this design is named as **Conv**. **ASPP** further enlarges the field of view of filters to enhance the adversarial learning. **AAN** is our appearance-level adaptation.

Table 2 details the mIoU improvement by considering one more factor for domain adaptation at each stage in FCAN. ABN is a general way to alleviate domain shift irrespective of any domain adaptation frameworks. In our case, ABN successfully brings up the mIoU performance from
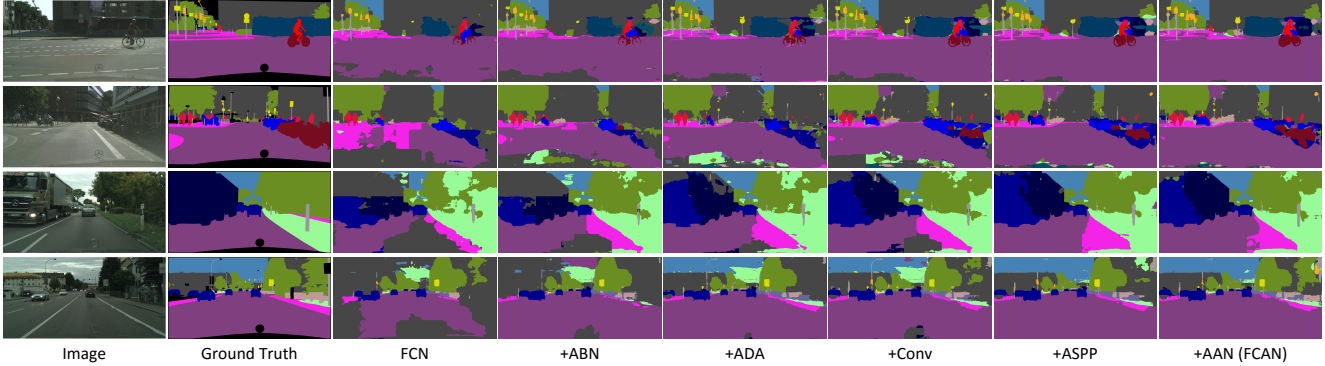
Figure 5. Examples of semantic segmentation results in Cityscapes. The original images, their ground truth and comparative segmentation results at different stages of FCAN are given.

Table 3. Performance comparisons with the state-of-the-art unsupervised domain adaptation methods on Cityscapes.

Method	mIoU
DC [30]	37.64
ADDA [31]	38.30
FCNWild [9]	42.04
FCAN	**46.60**
FCAN(MS)	**47.75**

Figure 6. Per-category IoU performance of different approaches and mIoU performance averaged over all the 19 categories.

29.15% to 35.51%. This demonstrates that ABN is a very effective and practical choice. ADA, Conv and ASPP are three specific designs in our RAN and the performance gain of each is 5.78%, 1.88% and 1.64%, respectively. In other words, our RAN leads to a large performance boost of 9.3% in total. The results verify the idea of representation-level adaptation. AAN further contributes an mIoU increase of 1.79% and the mIoU performance of FCAN finally reaches 46.6%. Figure 5 showcases four examples of semantic segmentation results at different stages of our FCAN. As illustrated in the figure, the segmentation results are becoming increasingly accurate as more adaptation designs are included. For instance, at the early stages, the majority categories such as road and sky cannot be well segmented. Instead, even the minority classes such as bicycle and truck are segmented nicely during the latter steps.

5.4. Comparisons with State-of-the-Art

We compare with several state-of-the-art techniques. Domain Confusion [30] (DC) aligns domains via domain confusion loss, which is optimized to learn a uniform distribution across different domains. Adversarial Discriminative Domain Adaptation [31] (ADDA) combines untied weight sharing and adversarial learning for discriminative feature learning. FCNWild [9] adopts fully convolutional adversarial training for domain adaptation on semantic segmentation. For fair comparison, the basic FCN utilized in all the methods are originated from ResNet-101. The performance comparisons are summarized in Table 3. Compared to DC and ADDA in which domain discriminator are both devised on image level, FCNWild and FCAN performing domain-adversarial learning on region level exhibit better
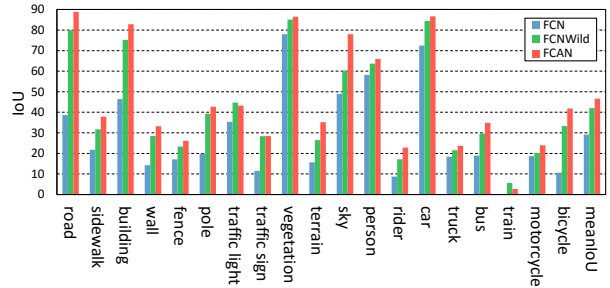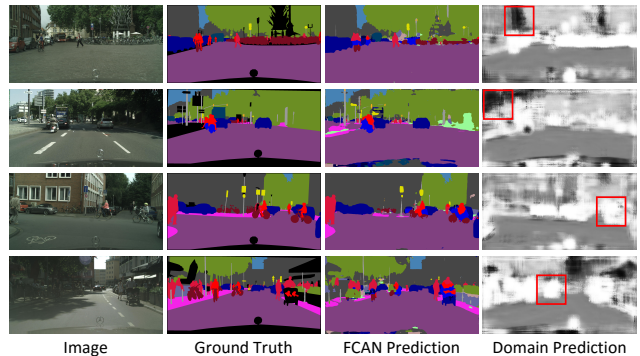

Figure 7. Examples of semantic segmentation results and the prediction maps by domain discriminator where brightness indicates the high probability of the region belonging to target domain.

performance. Furthermore, FCAN by additionally incorporating ASPP strategy and reinforcing by AAN, leads to an apparent improvement over FCNWild. The multi-scale (MS) scheme boosts up the mIoU performance to 47.75%. Figure 6 details the performance across different categories. Our FCAN achieves the best performance in 17 out of 19 categories, which empirically validate the effectiveness of our model on category level.

To examine domain discriminator learnt in FCAN, Figure 7 illustrates four image examples, including the original images, their ground truth, segmentation results by FCAN and prediction maps by domain discriminator. The brightness indicates that the region belongs to target domain with

Table 4. Results of Semi-supervised adaptation for Cityscapes.

# of images	FCN (On Cityscapes)	FCAN (Semi-supervised)
0	-	46.60
50	47.57	56.50
100	54.41	59.95
200	59.53	63.82
400	62.53	66.80
600	65.39	67.58
800	67.01	68.42
1000	68.05	69.17

Table 5. Comparisons of different unsupervised domain adaptation methods on BDDS.

Method	mIoU
FCNWild [9]	39.37
FCAN	43.35
FCAN(MS)	45.47
FCAN(MS+EN)	47.53

high probability. Let's recall that adversarial learning is to maximally fool the domain discriminator. That means ideally the prediction map of the images in target domain should be dark. For example, the domain discriminator predicts wrongly on the regions in the red bounding box in the first two images, which indicates that the representations on these regions tend to be indistinguishable. Hence, these regions (sky) are precisely segmented by FCAN. In contrast, domain discriminator predicts correctly on the regions in the last two images, indicating that the region representations are still domain-dependent. As such, the segmentation results on those regions (bicycle) are not that good.

5.5. Semi-Supervised Adaptation

Another common scenario in practice is that there is a small number of labeled training examples in target domain. Hence, we extend our FCAN to a semi-supervised version, which takes the training set of Cityscapes as labeled data $\mathcal{X}_t^l$. Technically, the pixel-level classification loss on images in target domain is further taken into account and the overall objective in Eq.(7) then changes to $\max_F \min_D \left\{ \mathcal{L}_{adv}(\mathcal{X}_s, \mathcal{X}_t) - \lambda_s \mathcal{L}_{seg}(\mathcal{X}_s) - \lambda_t \mathcal{L}_{seg}(\mathcal{X}_t^l) \right\}$. Table 4 shows the mIoU performances with the increase of labeled training data from target domain. It is also worth noting that here FCN is directly learnt on the labeled data in target domain and FCAN refers to our semi-supervised version. As expected, the performance gain of FCAN tends to be large if only a few hundred images in target domain are included in training. The gain is gradually decreased when increasing the number of images from Cityscapes. Even when the number reaches $1k$, our semi-supervised FCAN is still slightly better than supervised FCN.

5.6. Results on BDDS

In addition to Cityscapes dataset, we also take BDDS as target domain to evaluate the unsupervised setting of our
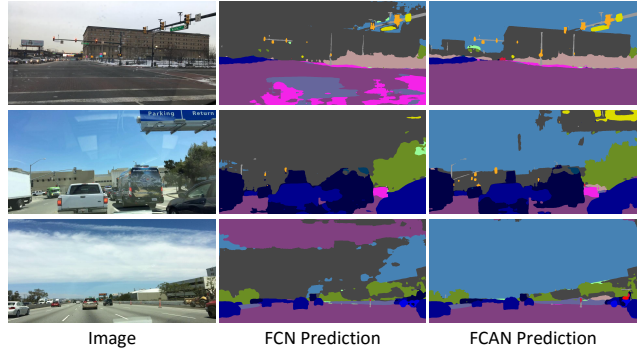

Figure 8. Examples of semantic segmentation results in BDDS.

FCAN. The performance comparisons are summarized in Table 5. In particular, the mIoU performance of FCAN achieves 43.35%, making the improvement over FCNWild by 3.98%. The multi-scale setting, i.e., FCAN(MS), increases the performance to 45.47%. Finally, the ensemble version FCAN(MS+EN) by fusing the models derived from ResNet-101, ResNet-152 and SENet [11], could boost up the mIoU to 47.53%. Figure 8 shows three semantic segmentation examples in BDDS, which are output by FCN and FCAN, respectively. Clearly, FCAN obtains much more promising segmentation results. Even in the case of a reflection (second row) or patches of cloud (third row) in the sky, our FCAN can segment the sky well.

6. Conclusion

We have presented Fully Convolutional Adaptation Networks (FCAN) architecture, which explores domain adaptation for semantic segmentation. Particularly, we study the problem from the viewpoint of both visual appearance-level and representation-level adaptation. To verify our claim, we have devised Appearance Adaptation Networks (AAN) and Representation Adaptation Networks (RAN) respectively in our FCAN for each purpose. AAN is to render an image in one domain with the domain "style" from the other one, resulting in invariant appearance across two domains. RAN aims to guide the representation learning in a domain-adversarial manner, which ideally outputs domain-invariant representations. Experiments conducted on the transfer from game videos (GTA5) to urban street-view scenes (Cityscapes) validate our proposal and analysis. More remarkably, we achieve new state-of-the-art performances when transferring game videos to drive-cam videos (BDDS). Our possible future works include two directions. First, more advanced techniques of rendering the semantic content of an image with another statistical pattern will be investigated in AAN. Second, we will further extend our FCAN to other specific segmentation scenarios, e.g., indoor scenes segmentation or portrait segmentation, where the synthetic data could be easily produced.

References

[1] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *Trans. on PAMI*, 40(4):834–848, 2018.

[2] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

[3] J. Dai, K. He, and J. Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, 2015.

[4] L. A. Gatys, A. S. Ecker, and M. Bethge. Texture synthesis using convolutional neural networks. In *NIPS*, 2015.

[5] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016.

[6] R. Girshick. Fast r-cnn. In *ICCV*, 2015.

[7] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014.

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

[9] J. Hoffman, D. Wang, F. Yu, and T. Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016.

[10] S. Hong, J. Oh, H. Lee, and B. Han. Learning transferrable knowledge for semantic segmentation with deep convolutional neural network. In *CVPR*, 2016.

[11] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. *arXiv preprint arXiv:1709.01507*, 2017.

[12] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*, 2014.

[13] G. Lin, A. Milan, C. Shen, and I. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, 2017.

[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[15] W. Liu, A. Rabinovich, and A. C. Berg. Parsenet: Looking wider to see better. In *ICLR Workshop*, 2016.

[16] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[17] M. Long, Y. Cao, J. Wang, and M. I. Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015.

[18] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille. Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *ICCV*, 2015.

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

[20] D. Pathak, E. Shelhamer, J. Long, and T. Darrell. Fully convolutional multi-class multiple instance learning. In *ICLR Workshop*, 2015.

[21] X. Peng, B. Usman, N. Kaushik, J. Hoffman, D. Wang, and K. Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

[22] P. O. Pinheiro and R. Collobert. From image-level to pixel-level labeling with convolutional networks. In *CVPR*, 2015.

[23] Z. Qiu, T. Yao, and T. Mei. Learning deep spatio-temporal dependence for semantic video segmentation. *Trans. on Multimedia*, 20:939–949, 2017.

[24] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[25] S. R. Richter, V. Vineet, S. Roth, and V. Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016.

[26] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

[27] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[28] B. Sun, J. Feng, and K. Saenko. Return of frustratingly easy domain adaptation. In *AAAI*, 2016.

[29] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[30] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In *ICCV*, 2015.

[31] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017.

[32] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[33] F. Xia, P. Wang, L.-C. Chen, and A. L. Yuille. Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net. In *ECCV*, 2016.

[34] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei. Boosting image captioning with attributes. In *ICCV*, 2017.

[35] T. Yao, Y. Pan, C.-W. Ngo, H. Li, and T. Mei. Semi-supervised domain adaptation with subspace learning for visual recognition. In *CVPR*, 2015.

[36] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016.

[37] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *CVPR*, 2017.

[38] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr. Conditional random fields as recurrent neural networks. In *ICCV*, 2015.