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 images available for developing deep models to a variety of vision tasks, e.g., recognition [8,27,29], detection [6,24], captioning [44] and semantic segmentation [11,16]. 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 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 adaptation 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 [58]). 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) [25] 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-
In this section we present our proposed Fully Convolutional Adaptation Networks (FCAN) for semantic segmentation. Our method is different from [9] in that we solve the domain shift 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. 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) 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.

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.

3.1. 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 \( \mathcal{X}_t = \{ x_i^t \} \) 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 \( x_t \) to produce an adaptive image \( x_{as} \). Specifically, a pre-trained CNN is u-
Figure 3. The architecture of Appearance Adaptation Networks (AAN). Given the target image set \( X_t \) 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 \( X_t \).

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 \( X_s = \{ x_s^i | i = 1, \ldots, n \} \) in source domain and \( X_t = \{ x_t^i | i = 1, \ldots, m \} \) in target domain, the adversarial loss \( \mathcal{L}_{adv} \) is the average classification loss over all spatial units, which is formulated as

\[
\mathcal{L}_{adv}(X_s, X_t) = -E_{x_s \sim X_s} \left[ \frac{1}{Z} \sum_{i=1}^{Z} \log(D_i(F(x_s^i))) \right] - E_{x_t \sim X_t} \left[ \frac{1}{Z} \sum_{i=1}^{Z} \log(1 - D_i(F(x_t^i))) \right],
\]

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

\[
\max_F \min_D \mathcal{L}_{adv}(X_s, X_t).
\]
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) \cite{chen2017rethinking} to implement this, as shown in Figure \ref{fig:aspp}. 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 $1 \times 1$ 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 $L_{seg}$ for supervised segmentation on the images from source domain, where the labels are available. Hence, the overall objective of RAN integrates $L_{seg}$ and $L_{adv}$ as

$$
\max \min_{\Phi} \{ L_{adv}(X_s, X_t) - \lambda L_{seg}(X_s) \}, \tag{7}
$$

where $\lambda$ 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 \cite{he2016deep} 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 $w^l_1$ and $w^l_2$ 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 $\alpha$ 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 $x_o^i = x_o^{i-1} - \beta w^{i-1} \frac{\partial L}{\partial x_o^{i-1}}$, where $g^{i-1} = \frac{\partial L}{\partial x_o^{i-1}}$, and $w^{i-1} = \beta L^{i-1}$ and $\beta = 10$.

4.2. Representation Adaptation

In our implementations, we employ dilated fully convolutional network \cite{chen2017rethinking} originated from ResNet-101 \cite{he2016deep} 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 \cite{chen2017rethinking} 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 \cite{jia2014caffe} 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 \cite{chen2017rethinking}, 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 $\lambda$ 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 \cite{richter2016playing} (game videos) dataset to Cityscapes (urban street scenes) dataset \cite{cordts2016cityscapes}.

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 $1914 \times 1052$. 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 $2048 \times 1024$. Following the standard protocol in segmentation task (e.g., \cite{cordts2016cityscapes}), 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.

<table>
<thead>
<tr>
<th>Train</th>
<th>Validation</th>
<th>FCN</th>
<th>RAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Src</td>
<td>Tar</td>
<td>29.15</td>
<td>44.81</td>
</tr>
<tr>
<td>Src</td>
<td>Tar_Ada</td>
<td>34.68</td>
<td>45.03</td>
</tr>
<tr>
<td>Src_Ada</td>
<td>Tar</td>
<td>31.71</td>
<td><strong>46.21</strong></td>
</tr>
<tr>
<td>Src_Ada</td>
<td>Tar_Ada</td>
<td>36.25</td>
<td>45.59</td>
</tr>
<tr>
<td>Late Fusion</td>
<td></td>
<td>37.61</td>
<td><strong>46.60</strong></td>
</tr>
</tbody>
</table>

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 $1280 \times 720$. 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>ABN</th>
<th>ADA</th>
<th>Conv</th>
<th>ASPP</th>
<th>AAN</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>29.15</td>
</tr>
<tr>
<td>+ABN</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>35.51</td>
</tr>
<tr>
<td>+ADA</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td>41.29</td>
</tr>
<tr>
<td>+Conv</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
<td>43.17</td>
</tr>
<tr>
<td>+ASPP</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td></td>
<td>44.81</td>
</tr>
<tr>
<td>FCAN</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td><strong>46.60</strong></td>
</tr>
</tbody>
</table>

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.

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...
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 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
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 $X_l$. 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

$$\max_F \min_D \left\{ L_{adv}(X_s, X_t) - \lambda_s L_{seg}(X_s) - \lambda_t L_{seg}(X_t) \right\}. $$

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 FCAN. The performance segmentation results are not that good. 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.

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


