

A FULLY CONVOLUTIONAL TRI-BRANCH NETWORK (FCTN) FOR DOMAIN ADAPTATION

Junting Zhang, Chen Liang and C.-C. Jay Kuo

University of Southern California, Los Angeles, CA, USA

{juntingz, lian455}@usc.edu, cckuo@sipi.usc.edu

ABSTRACT

A domain adaptation method for urban scene segmentation is proposed in this work. We develop a fully convolutional tri-branch network, where two branches assign pseudo labels to images in the unlabeled target domain while the third branch is trained with supervision based on images in the pseudo-labeled target domain. The re-labeling and re-training processes alternate. With this design, the tri-branch network learns target-specific discriminative representations progressively and, as a result, the cross-domain capability of the segmenter improves. We evaluate the proposed network on large-scale domain adaptation experiments using both synthetic (GTA) and real (Cityscapes) images. It is shown that our solution achieves the state-of-the-art performance and it outperforms previous methods by a significant margin.

Index Terms— Domain Adaptation, Semantic Segmentation, Urban Scene, Tri-training

1. INTRODUCTION

Semantic segmentation for urban scenes is an important yet challenging task for a variety of vision-based applications, including autonomous driving cars, smart surveillance systems, etc. With the success of convolutional neural networks (CNNs), numerous successful fully-supervised semantic segmentation solutions have been proposed in recent years [1, 2]. To achieve satisfactory performance, these methods demand a sufficiently large dataset with pixel-level labels for training. However, creating such large datasets is prohibitively expensive as it requires human annotators to accurately trace segment boundaries. Furthermore, it is difficult to collect traffic scene images with sufficient variations in terms of lighting conditions, weather, city and driving routes.

To overcome the above-mentioned limitations, one can utilize the modern urban scene simulators to automatically generate a large amount of synthetic images with pixel-level labels. However, this introduces another problem, *i.e.* distributions mismatch between the source domain (synthesized data) and the target domain (real data). Even if we synthesize images with the state-of-the-art simulators [3, 4], there still exists visible appearance discrepancy between these two domains. The testing performance in the target domain using the network trained solely by the source domain images is severely degraded. The domain adaptation (DA) technique is developed to bridge this gap. It is a special example of transfer learning that leverages labeled data in the source domain to learn a robust classifier for unlabeled data in the target domain. DA methods for object classification have several challenges such as shifts in lighting and variations in object's appearance and pose. There are even more challenges in DA methods for semantic segmentation because

of variations in the scene layout, object scales and class distributions in images. Many successful domain-alignment-based methods work for DA-based classification but not for DA-based segmentation. Since it is not clear what comprises data instances in a deep segmenter [5], DA-based segmentation is still far from its maturity.

In this work, we propose a novel fully convolutional tri-branch network (FCTN) to solve the DA-based segmentation problem. In the FCTN, two labeling branches are used to generate pseudo segmentation ground-truth for unlabeled target samples while the third branch learns from these pseudo-labeled target samples. An alternating re-labeling and re-training mechanism is designed to improve the DA performance in a curriculum learning fashion. We evaluate the proposed method using large-scale synthesized-to-real urban scene datasets and demonstrate substantial improvement over the baseline network and other benchmarking methods.

2. RELATED WORK

The current literatures on visual domain adaptation mainly focus on image classification [6]. Being inspired by shallow DA methods, one common intuition of deep DA methods is that adaptation can be achieved by matching the distribution of features in different domains. Most deep DA methods follow a siamese architecture with two streams, representing the source and target models. They aim to obtain domain-invariant features by minimizing the divergence of features in the two domains and a classification loss [7, 8, 9, 10], where the classification loss is evaluated in the source domain with labeled data only. However, these methods assume the existence of a universal classifier that can perform well on samples drawn from whichever domain. This assumption tends to fail since the class correspondence constraint is rarely imposed in the domain alignment process. Without such an assumption, feature distribution matching may not lead to classification improvement in the target domain. The ATDA method proposed in [11] avoids this assumption by employing the asymmetric tri-training. It can assign pseudo labels to unlabeled target samples progressively and learn from them using a curriculum learning paradigm. This paradigm has been proven effective in the weakly-supervised learning tasks [12] as well.

Previous work on segmentation-based DA is much less. Hoffman *et al.* [13] consider each spatial unit in an activation map of a fully convolutional network (FCN) as an instance, and extend the idea in [9] to achieve two objectives: 1) minimizing the global domain distance between two domains using a fully convolutional adversarial training and 2) enhancing category-wise adaptation capability via multiple instance learning. The adversarial training aims to align intermediate features from two domains. It implies the existence of a single good mapping from the domain-invariant feature space to the correct segmentation mask. To avoid this condition,

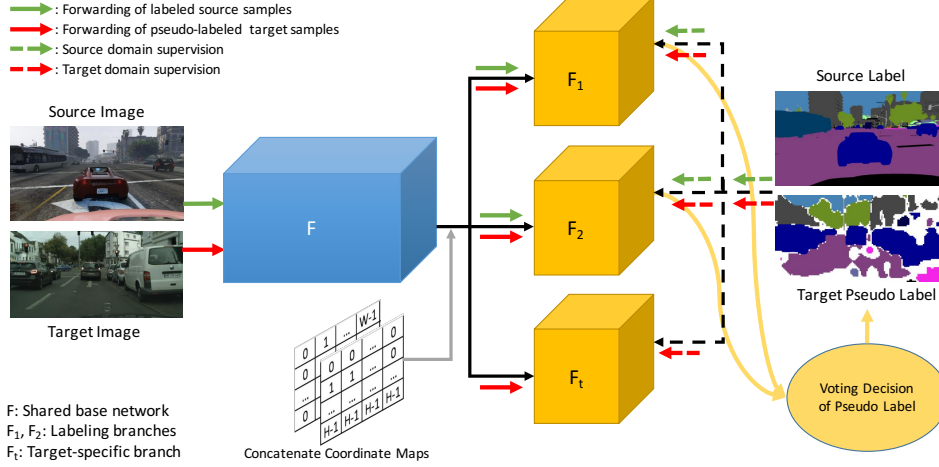


Fig. 1: An overview of the proposed fully convolutional tri-branch network (FCTN). It has one shared base network denoted by F followed by three branches of the same architecture denoted by F_1 , F_2 and F_t . Branches F_1 and F_2 assign pseudo labels to images in the unlabeled target domain, while branch F_t is trained with supervision from images in the pseudo-labeled target domain.

Zhang *et al.* [5] proposed to predict the class distribution over the entire image and some representative super pixels in the target domain first. Then, they use the predicted distribution to regularize network training. In this work, we avoid the single good mapping assumption and rely on the remarkable success of the ATDA method [11]. In particular, we develop a curriculum-style method that improves the cross-domain generalization ability for better performance in DA-based segmentation.

3. PROPOSED DOMAIN ADAPTATION NETWORK

The proposed fully convolutional tri-branch network (FCTN) model for cross-domain semantic segmentation is detailed in this section. The labeled source domain training set is denoted by $\mathcal{S} = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ while the unlabeled target domain training set is denoted by $\mathcal{T} = \{x_i^t\}_{i=1}^{n_t}$, where x is an image, y is the ground truth segmentation mask and n_s and n_t are the sizes of training sets of two domains, respectively.

3.1. Fully Convolutional Tri-branch Network Architecture

An overview of the proposed FCTN architecture is illustrated in Fig. 1. It is a fully convolutional network that consists of a shared base network (F) followed by three branch networks (F_1 , F_2 and F_t). Branches F_1 and F_2 are labeling branches. They accept deep features extracted by the shared base net, F , as the input and predict the semantic label of each pixel in the input image. Although the architecture of the three branches are the same, their roles and functions are not identical. F_1 and F_2 generate pseudo labels for the target images based on prediction. F_1 and F_2 learn from both labeled source images and pseudo-labeled target images. In contrast, F_t is a target-specific branch that learns from pseudo-labeled target images only.

We use the DeepLab-LargeFOV (also known as the DeepLab v1) [14] as the reference model due to its simplicity and superior performance in the semantic segmentation task. The DeepLab-LargeFOV is a re-purposed VGG-16 [15] network with dilated convolutional kernels. The shared base network F contains 13 convolutional layers while the three branch networks are formed by three convolutional layers that are converted from fully connected layers in the original VGG-16 network. Although the DeepLab-LargeFOV

is adopted here, any effective FCN-based semantic segmentation framework can be used in the proposed FCTN architecture as well.

3.2. Encoding Explicit Spatial Information

Being inspired by PFN [16], we attach the pixel coordinates as the additional feature map to the last layer of F . The intuition is that the urban traffic scene images have structured layout and certain classes usually appear in a similar location in images. However, a CNN is translation-invariant by nature. That is, it makes prediction based on patch-based feature regardless of the patch location in the original image. Assume that the last layer in F has a feature map of size $H \times W \times D$, where H , W and D are the height, width and depth of the feature map, respectively. We generate two spatial coordinate maps X and Y of size $H \times W$, where values of $X(p_x, p_y)$ and $Y(p_x, p_y)$ are set to be p_x/W and p_y/H for pixel p at location (p_x, p_y) , respectively. We concatenate spatial coordinate maps X and Y to the original feature maps along the depth dimension. Thus, the output feature maps are of dimension $H \times W \times (D + 2)$. By incorporating the spatial coordinate maps, the FCTN can learn more location-aware representations.

3.3. Assigning Pseudo Labels to Target Images

Being inspired by the ATDA method [11], we generate pseudo labels by feeding images in the target domain training set to the FCTN and collect predictions from both labeling branches. For each input image, we assign the pseudo-label to a pixel if the following two conditions are satisfied: 1) the classifiers associated with labeling branches, F_1 and F_2 , agree in their predicted labels on this pixel; 2) the higher confidence score of these two predictions exceeds a certain threshold. In practice, the confidence threshold is set very high (say, 0.95 in our implementation) because the use of many inaccurate pseudo labels tends to mislead the subsequent network training. In this way, high-quality pseudo labels for target images are used to guide the network to learn target-specific discriminative features. The pseudo-labeled target domain training set is denoted by $\mathcal{T}_l = \{(x_i^t, \hat{y}_i^t)\}_{i=1}^{n_t}$, where \hat{y} is the partially pseudo-labeled segmentation mask. Some sample pseudo-labeled segmentation masks are shown in Fig. 2. In the subsequent training, the not-yet-labeled pixels are simply ignored in the loss computation.

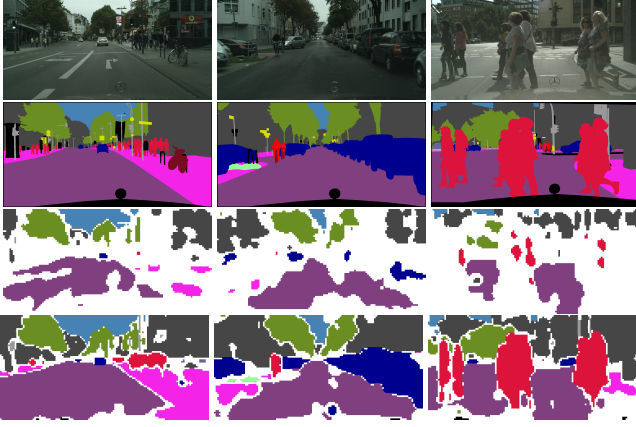


Fig. 2: Illustration of pseudo labels used in the 2-round curriculum learning in the GTA-to-Cityscapes DA experiments. The first row shows the input images. The second row shows the ground truth segmentation masks. The third and fourth row shows the pseudo labels used in the first and second round of curriculum learning, respectively. Note in the visualization of pseudo labels, white pixels indicate the unlabeled pixels. Best viewed in color.

3.4. Loss Function

Weight-Constrained Loss. As suggested in the standard tri-training algorithm [17], the three classifiers in F_1 , F_2 and F_t must be diverse. Otherwise, the training degenerates to self-training. In our case, one crucial requirement to obtain high-quality pseudo-labels from two labeling branches F_1 and F_2 is that they should have different views on one sample and make decisions on their own.

Unlike the case in the co-training algorithm [18], where one can explicitly partition features into different sufficient and redundant views, it is not clear how to partition deep features effectively in our case. Here, we enforce divergence of the weights of the convolutional layers of two labeling branches by minimizing their cosine similarity. Then, we have the following filter weight-constrained loss term:

$$L_w = \frac{\vec{w}_1 \cdot \vec{w}_2}{\|\vec{w}_1\| \|\vec{w}_2\|} \quad (1)$$

where \vec{w}_1 and \vec{w}_2 are obtained by the flattening and concatenating the weights of convolutional filters in convolutional layers of F_1 and F_2 , respectively.

Weighted Pixel-wise Cross-entropy Loss. In the curriculum learning stage, we take a minibatch of samples with one half from \mathcal{S} and the other half from \mathcal{T}_l at each step. We calculate the segmentation losses separately for each half of samples. For the source domain images samples, we use the vanilla pixel-wise softmax cross-entropy loss, denoted by L_S , as the segmentation loss function.

Furthermore, as mentioned in Sec. 3.3, we assign pseudo labels to target domain pixels based on predictions of two labeling branches. This mechanism tends to assign pseudo labels to the prevalent and easy-to-predict classes, such as the road, building, etc., especially in the early stage (this can be seen in Fig. 2). Thus, the pseudo labels can be highly imbalanced in classes. If we treat all classes equally, the gradients from challenging and relatively rare classes will be insignificant and the training will be biased toward prevalent classes. To remedy this, we use a weighted cross-entropy loss for target domain samples, denoted by $L_{\mathcal{T}_l}$. We calculate weights using the median frequency balancing scheme [19], where

the weight assigned to class c in the loss function becomes

$$\alpha_c = \frac{\text{median_freq}}{\text{freq}(c)}, \quad (2)$$

where $\text{freq}(c)$ is the number of pixels of class c divided by the total number of pixels in the source domain images whenever c is present, and median_freq is the median of these frequencies $\{\text{freq}(c)\}_{c=1}^C$, and where C is the total number of classes. This scheme works well under the assumption that the global class distributions of the source domain and the target domain are similar.

Total Loss Function. There are two stages in our training procedure. We first pre-train the entire network using minibatches from \mathcal{S} so as to minimize the following objective function:

$$L = \alpha L_w + L_S \quad (3)$$

Once the curriculum learning starts, the overall objective function becomes

$$L = \alpha L_w + L_S + \beta L_{\mathcal{T}_l} \quad (4)$$

where L_S is evaluated on \mathcal{S} and averaged over predictions of F_1 and F_2 branches, $L_{\mathcal{T}_l}$ is evaluated on \mathcal{T}_l and averaged over predictions of all three top branches, and α and β are hyper-parameters determined by the validation split.

3.5. Training Procedure

The training process is illustrated in Algorithm 1. We first pretrain the entire FCTN on the labeled source domain training set \mathcal{S} for *iters* iterations, optimizing the loss function in Eq. (3). We then use the pre-trained model to generate the initial pseudo labels for the target domain training set \mathcal{T} , using the method described in Sec. 3.3. We re-train the network using \mathcal{S} and \mathcal{T}_l for several steps. At each step, we take a minibatch of samples with half from \mathcal{S} and half from \mathcal{T}_l , optimizing the terms in Eq. (4) jointly. We repeat the re-labeling of \mathcal{T} and the re-training of the network for several rounds until the model converges.

Algorithm 1 Training procedure for our fully convolutional tri-branch network (FCTN).

Input: labeled source domain training set $\mathcal{S} = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ and unlabeled target domain training set $\mathcal{T} = \{x_i^t\}_{i=1}^{n_t}$

Pretraining on \mathcal{S} :

for $i = 1$ to *iters* **do**

 train F, F_1, F_2, F_t with minibatches from \mathcal{S}

end for

Curriculum Learning with \mathcal{S} and \mathcal{T} :

for $i = 1$ to *rounds* **do**

$\mathcal{T}_l \leftarrow \text{LABELING}(F, F_1, F_2, \mathcal{T})$

▷ See Sec. 3.3

for $k = 1$ to *steps* **do**

 train F, F_1, F_2 with samples from \mathcal{S}

 train F, F_1, F_2, F_t with samples from \mathcal{T}_l

end for

end for

return F, F_t

4. EXPERIMENTS

We validate the proposed method by experimenting the adaptation from the recently built synthetic urban scene dataset GTA [3] to the commonly used urban scene semantic segmentation dataset Cityscapes [20].

Model	per-class IoU																			mIoU
	road	sidewlk	bldg.	wall	fence	pole	t. light	t. sign	veg.	terr.	sky	person	rider	car	truck	bus	train	mbike	bike	
No Adapt	31.9	18.9	47.7	7.4	3.1	16.0	10.4	1.0	76.5	13.0	58.9	36.0	1.0	67.1	9.5	3.7	0.0	0.0	0.0	21.1
FCN [13]	70.4	32.4	62.1	14.9	5.4	10.9	14.2	2.7	79.2	21.3	64.6	44.1	4.2	70.4	8.0	7.3	0.0	3.5	0.0	27.1
No Adapt	18.1	6.8	64.1	7.3	8.7	21.0	14.9	16.8	45.9	2.4	64.4	41.6	17.5	55.3	8.4	5.0	6.9	4.3	13.8	22.3
CDA [5]	26.4	22.0	74.7	6.0	11.9	8.4	16.3	11.1	75.7	13.3	66.5	38.0	9.3	55.2	18.8	18.9	0.0	16.8	14.6	27.8
No Adapt	59.7	24.8	66.8	12.8	7.9	11.9	14.2	4.2	78.7	22.3	65.2	44.1	2.0	67.8	9.6	2.4	0.6	2.2	0.0	26.2
Round 1	66.9	25.6	74.7	17.5	10.3	17.1	18.4	8.0	79.7	34.8	59.7	46.7	0.0	77.1	10.0	1.8	0.0	0.0	0.0	28.9
Round 2	72.2	28.4	74.9	18.3	10.8	24.0	25.3	17.9	80.1	36.7	61.1	44.7	0.0	74.5	8.9	1.5	0.0	0.0	0.0	30.5

Table 1: Adaptation from GTA to Cityscapes. All numbers are measured in %. The last three rows show our results before adaptation, after one and two rounds of curriculum learning using the proposed FCTN, respectively.

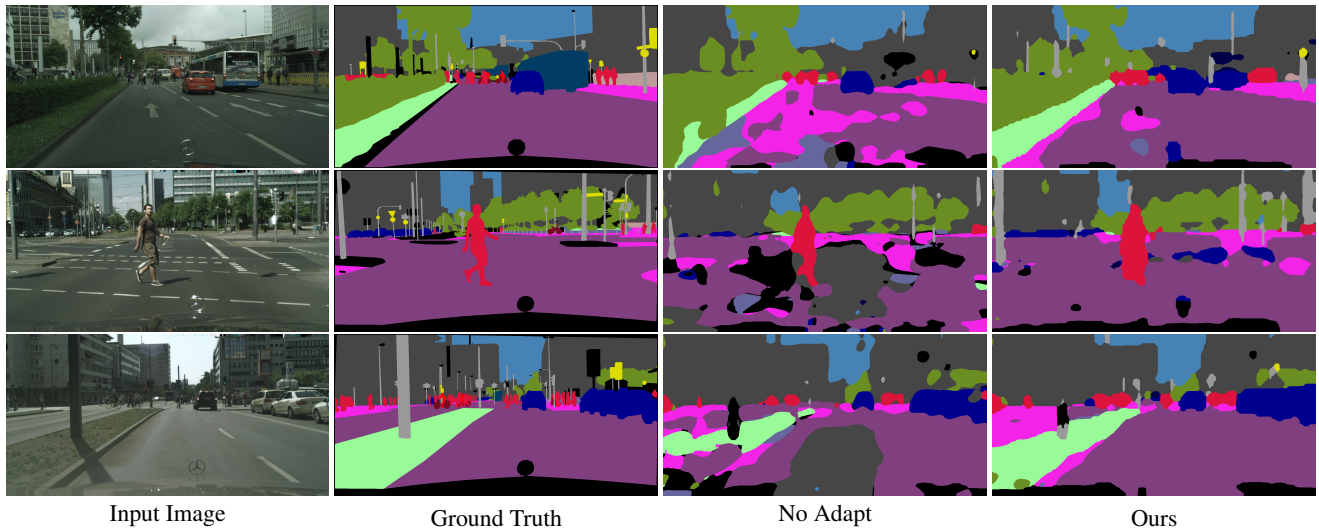


Fig. 3: Domain adaptation results from the Cityscapes Val set. The third column shows segmentation results using the model trained solely by the GTA dataset, and the fourth column shows the segmentation results after two rounds of the FCTN training (best viewed in color).

Cityscapes [20] is a large-scale urban scene semantic segmentation dataset. It provides over 5,000 finely labeled images (train/validation/test: 2,993/503/1,531), which are labeled with per pixel category labels. They are with high resolution of 1024×2048 . There are 34 distinct semantic classes in the dataset, but only 19 classes are considered in the official evaluation protocol.

GTA [3] contains 24,966 high-resolution labeled frames extracted from realistic open-world computer games, Grand Theft Auto V (GTA5). All the frames are vehicle-egocentric and the class labels are fully compatible with Cityscapes.

We implemented our method using Tensorflow[21] and trained our model using a single NVIDIA TITAN X GPU. We initialized the weights of shared base net F using the weights of the VGG-16 model pretrained on ImageNet. The hyper-parameter settings were $\alpha = 10^3, \beta = 100$. We used a constant learning rate 10^{-5} in the training. We trained the model for 70k, 13k and 20k iterations in the pre-training and two rounds of curriculum learning, respectively.

We use synthetic data as source labeled training data and Cityscapes train as an unlabeled target domain, while evaluating our adaptation algorithm on Cityscapes val using the predictions from the target specific branch F_t . Following Cityscapes official evaluation protocol, we evaluate our segmentation domain adaptation results using the per-class intersection over union (IoU) and mean IoU over the 19 classes. The detailed results are listed in Table. 1 and some qualitative results are shown in Fig. 3. We achieve

the state-of-the-art domain adaptation performance. Our two rounds of curriculum learning boost the mean IoU over our non-adapted baseline by 2.7% and 4.3%, respectively. Especially, the IoU improvement for the small objects (*e.g.* pole, traffic light, traffic sign etc.) are significant (over 10%).

5. CONCLUSION

A systematic way to address the unsupervised semantic segmentation domain adaptation problem for urban scene images was presented in this work. The FCTN architecture was proposed to generate high-quality pseudo labels for the unlabeled target domain images and learn from pseudo labels in a curriculum learning fashion. It was demonstrated by the DA experiments from the large-scale synthetic dataset to the real image dataset that our method outperforms previous benchmarking methods by a significant margin.

There are several possible future directions worth exploring. First, it is interesting to develop a better weight constraint for the two labeling branches so that even better pseudo labels can be generated. Second, we may impose the class distribution constraint on each individual image [5] so as to alleviate the confusion between some visually similar classes, *e.g.* road and sidewalk, vegetation and terrain etc. Third, we can extend the proposed method to other tasks, *e.g.* instance-aware semantic segmentation.

6. REFERENCES

- [1] Jonathan Long, Evan Shelhamer, and Trevor Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.
- [2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” *arXiv preprint arXiv:1606.00915*, 2016.
- [3] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun, “Playing for data: Ground truth from computer games,” in *European Conference on Computer Vision*. Springer, 2016, pp. 102–118.
- [4] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez, “The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3234–3243.
- [5] Yang Zhang, Philip David, and Boqing Gong, “Curriculum domain adaptation for semantic segmentation of urban scenes,” in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [6] Gabriela Csurka, Ed., *Domain Adaptation in Computer Vision Applications*, Advances in Computer Vision and Pattern Recognition. Springer, 2017.
- [7] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan, “Learning transferable features with deep adaptation networks,” in *International Conference on Machine Learning*, 2015, pp. 97–105.
- [8] Baochen Sun and Kate Saenko, “Deep coral: Correlation alignment for deep domain adaptation,” in *Computer Vision—ECCV 2016 Workshops*. Springer, 2016, pp. 443–450.
- [9] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko, “Simultaneous deep transfer across domains and tasks,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4068–4076.
- [10] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell, “Adversarial discriminative domain adaptation,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [11] Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada, “Asymmetric tri-training for unsupervised domain adaptation,” in *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 2988–2997.
- [12] Siyang Li, Xiangxin Zhu, Qin Huang, Hao Xu, and C.-C. Jay Kuo, “Multiple instance curriculum learning for weakly supervised object detection,” in *British Machine Vision Conference*, 2017.
- [13] 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.
- [14] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille, “Semantic image segmentation with deep convolutional nets and fully connected crfs,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.
- [15] Karen Simonyan and Andrew Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.
- [16] Xiaodan Liang, Yunchao Wei, Xiaohui Shen, Jianchao Yang, Liang Lin, and Shuicheng Yan, “Proposal-free network for instance-level object segmentation,” *arXiv preprint arXiv:1509.02636*, 2015.
- [17] Zhi-Hua Zhou and Ming Li, “Tri-training: Exploiting unlabeled data using three classifiers,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 11, pp. 1529–1541, 2005.
- [18] Avrim Blum and Tom Mitchell, “Combining labeled and unlabeled data with co-training,” in *Proceedings of the eleventh annual conference on Computational learning theory*. ACM, 1998, pp. 92–100.
- [19] David Eigen and Rob Fergus, “Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2650–2658.
- [20] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3213–3223.
- [21] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al., “Tensorflow: Large-scale machine learning on heterogeneous distributed systems,” *arXiv preprint arXiv:1603.04467*, 2016.