Unsupervised Scene Adaptation with Memory Regularization \textit{in vivo}

Zhending Zheng \quad Yi Yang

Centre for Artificial Intelligence, University of Technology Sydney
zhedong.zheng@student.uts.edu.au, yi.yang@uts.edu.au

Abstract

We consider the unsupervised scene adaptation problem of learning from both labeled source data and unlabeled target data. Existing methods focus on minoring the inter-domain gap between the source and target domains. However, the intra-domain knowledge and inherent uncertainty learned by the network are under-explored. In this paper, we propose an orthogonal method, called memory regularization \textit{in vivo} to exploit the intra-domain knowledge and regularize the model training. Specifically, we refer to the segmentation model itself as the memory module, and minor the discrepancy of the two classifiers, \textit{i.e.,} the primary classifier and the auxiliary classifier, to reduce the prediction inconsistency. Without extra parameters, the proposed method is complementary to most existing domain adaptation methods and could generally improve the performance of existing methods. Albeit simple, we verify the effectiveness of memory regularization on two synthetic-to-real benchmarks: GTA5 $\rightarrow$ Cityscapes and SYNTHIA $\rightarrow$ Cityscapes, yielding $+11.1\%$ and $+11.3\%$ mIoU improvement over the baseline model, respectively. Besides, a similar $+12.0\%$ mIoU improvement is observed on the cross-city benchmark: Cityscapes $\rightarrow$ Oxford RobotCar.

1 Introduction

Due to the unaffordable cost of the segmentation annotation, unsupervised scene adaptation is to adapt the learned model to a new domain without extra annotation. In contrast to the conventional segmentation tasks, unsupervised scene adaptation reaches one step closer to the real-world practice. In the real-world scenario, the annotation of the target scene is usually hard to acquire. In contrast, abundant source data is easy to access. To improve the model scalability on the unlabeled target domain, most researchers resort to transfer the common knowledge learned from the source domain to the target domain.

The existing scene adaptation methods typically focus on reducing the discrepancy between the source domain and the target domain. The alignment between the source and target domains could be conducted on different levels, such as pixel level [Hoffman \textit{et al.}, 2018; Wu \textit{et al.}, 2018], feature level [Hoffman \textit{et al.}, 2018; Huang \textit{et al.}, 2018; Yue \textit{et al.}, 2019; Luo \textit{et al.}, 2019b; Zhang \textit{et al.}, 2019a] and semantic level [Tsai \textit{et al.}, 2018; Tsai \textit{et al.}, 2019; Wang \textit{et al.}, 2019]. Despite the great success, the brute-force alignment drives the model to learn the domain-agnostic shared features of both domains. We consider that this line of methods is sub-optimal in that it ignores the domain-specific feature learning on the target domain, and compromise the final adaptation performance.

Since the domain-specific knowledge is ignored for the target unlabeled data, the regularization by the data itself does not aid in the domain adaptation. To qualitatively verify this, we leverage the auxiliary classifier of the baseline model [Tsai \textit{et al.}, 2018] as a probe to pinpoint the inconsistency. As shown in Fig. 1, we observe that the model predicts one consistent supervised result of the source labeled data, while the unlabeled target data suffers from the inconsistency. The predicted result of the primary classifier is different from the auxiliary classifier prediction, especially in the target domain. It implies that the intra-domain consistency has not
been learned automatically, when we minor the inter-domain discrepancy.

To effectively exploit the intra-domain knowledge and reduce the target prediction inconsistency, we propose a memory mechanism into the deep neural network training, called memory regularization in vivo. Different from the previous works focusing on the inter-domain alignment, the proposed method intends to align the different predictions within the same domain to regularize the training. As shown in Fig. 2(c), we consider the inputs as key and the output prediction as the corresponding value. In other words, the proposed method deploys the model itself as the memory module, which memorizes the historical prediction and learns the key-value projection. Since we have the auxiliary classifier and the primary classifier, we could obtain two values for one key. We note that the proposed method is also different from other semi-supervised works deploying the extra memory terms. Since the proposed method does not require additional parameters or modules, we use the term “in vivo” to differentiate our method from [Chen et al., 2018; Tarvainen and Valpola, 2017; Zhang et al., 2018b]; these methods deploy external memory modules.

Our contribution is two-fold: (1) We propose to leverage the memory of model learning to pinpoint the prediction uncertainty and exploit the intra-domain knowledge. This is in contrast to most existing adaption methods focusing on the inter-domain alignment. (2) We formulate the memory regularization in vivo as the internal prediction discrepancy between the two classifiers. Different from the existing memory-based models, the proposed method does not need extra parameters, and is compatible with most scene segmentation networks.

2 Related Works
2.1 Domain Adaptation for Segmentation
Most existing works typically focus on minoring the domain discrepancy between the source domain and the target domain to learn the shared knowledge. Some pioneering works [Hoffman et al., 2018; Wu et al., 2018] apply the image generator to transfer the source data to the style of the target data, and intend to reduce the low-level visual appearance difference. Similarly, Yu et al.[Yue et al., 2019] and Wu et al.[Wu et al., 2019] generate the training images of different styles to learn the domain-agnostic feature. Adversarial loss is also widely studied. Tsai et al.[Tsai et al., 2018; Tsai et al., 2019] apply the adversarial losses to different network layers to enforce the domain alignment. Luo et al.[Luo et al., 2019b] leverage the attention mechanism and the class-aware adversarial loss to further improve the performance. Besides, some works also focus on mining the target domain knowledge, which is close to our work. Zou et al.[Zou et al., 2018; Zou et al., 2019] leverage the confident pseudo labels to further fine-tune the model on the target domain, yielding a competitive benchmark. Recently, Shen et al.[Shen et al., 2019] propose to utilize the discriminator to find the confident pseudo label. Different from the pseudo label based methods, the proposed method focuses on target domain knowledge by mining the intrinsic uncertainty of the model learning on the unlabeled target-domain data. We note that the proposed method is orthogonal to the existing methods, including the inter-domain alignment [Tsai et al., 2018; Tsai et al., 2019; Luo et al., 2019b] and self-training with pseudo labels [Zou et al., 2018; Zou et al., 2019]. In Section 4.2, we show the proposed method can be integrated with other domain adaption methods to further improve the performance.

2.2 Memory-based Learning
As one of the early works, Weston et al.[Weston et al., 2014] propose to use external memory module to store the long-term memory. In this way, the model could reason with the related experience more effectively. Chen et al.[Chen et al., 2018] further apply the memory to the semi-supervised learning to learn from the unlabeled data. In this work, we argue that the teacher model, which is applied in many frameworks, also could be viewed as one kind of external memory terms. Because the teacher model distills the knowledge of
the original setting, and memorizes the key concepts to the final prediction [Hinton et al., 2015]. For instance, one of the early work, called temporal ensemble [Laine and Aila, 2016], uses the historical models to regularize the running model, yielding the competitive performance. The training sample could be viewed as the key, and the historical models are the memory model to find the corresponding value for the key. Since the historical models memorize the experience from the previous training samples, the temporal ensemble could provide stable and relatively accurate predictions of the unlabeled data. Except for [Laine and Aila, 2016], there are different kinds of external memory models. Mean Teacher [Tarvainen and Valpola, 2017] leverages the weight moving average model as the memory model to regularize the training. Further, French et al. [French et al., 2017] extend Mean Teacher for visual domain adaptation. Zhang et al. [Zhang et al., 2018b] propose mutual learning, which learns the knowledge from multiple student models.

Different from existing memory-based methods [Chen et al., 2018; Tarvainen and Valpola, 2017; Zhang et al., 2018b], the proposed method leverages the memory of the model itself to regularize the running model. The proposed memory regularization does not introduce extra parameters and external modules. (see Fig. 2)

3 Method

3.1 Algorithm Overview

Formulation. We denote the images from the source domain and the target domain as 
\[ X_s = \{x^s_i\}^M_{i=1} \] and \[ X_t = \{x^t_i\}^N_{i=1} \], where \( M, N \) are the number of the source images and target images. Every source domain data in \( X_s \) is annotated with corresponding ground-truth segmentation maps \( Y_s = \{y^s_i\}^M_{i=1} \). Given one unlabeled target domain image \( x^t_i \), we intend to learn a function to project the image to the segmentation map \( y^t_i \). Following the practice in [Tsai et al., 2018; Luo et al., 2019b], we adopt the modified DeepLabv2 as our baseline model, which contains one backbone model and two classifiers, i.e., the primary classifier \( C_p \) and the auxiliary classifier \( C_a \). To simplify, we denote the two functions \( F_p \) and \( F_a \) as the segmentation functions, where \( F_p \) projects the image to the prediction of the primary classifier, and \( F_a \) maps the input to the prediction of the auxiliary classifier.

Overview. As shown in Fig. 3, the proposed method has two training stages, i.e., Stage-I and Stage-II, to progressively transfer the learned knowledge from the labeled source data to the unlabeled target data. In the Stage-I, we follow the conventional domain adaptation methods to minor the inter-domain discrepancy between the source domain and the target domain. When training, we regularize the model by adding the memory regularization. The memory regularization helps to minor the intra-domain inconsistency, yielding the performance improvement. In the Stage-II, we leverage the trained model to predict the label for the unlabeled target data. Then the model is further fine-tuned on the target domain. With the help of pseudo labels, the model could focus on learning domain-specific knowledge on the target domain. The pseudo labels inevitably contain noise, and the memory regularization in Stage-II could prevent the model from overfitting to the noise in pseudo labels. Next we introduce different objectives for the model adaptation in detail. We divide the losses into two classes: (1) Domain-agnostic learning to learn the shared inter-domain features from the source domain; (2) Domain-specific learning to learn the intra-domain knowledge, especially the features for the target domain.

3.2 Domain-agnostic learning

Segmentation loss. First, we leverage the annotated source-domain data to learn the source-domain knowledge. The segmentation loss is widely applied, and could be formulated as
the pixel-wise cross-entropy loss:

\[ L_{p\text{seg}}^{p} = - \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{c=1}^{C} y_{h}^{c} \log(F_{p}(x_{h}^{c})), \]  

(1)

\[ L_{p\text{seg}}^{a} = - \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{c=1}^{C} \hat{y}_{h}^{c} \log(F_{a}(x_{h}^{c})), \]  

(2)

where the first loss is for the primary prediction, and the second objective is for the auxiliary prediction. \( H \) and \( W \) denote the height and the width of the input image, and \( C \) is the number of segmentation classes.

**Adversarial loss.** Segmentation loss only focuses on the source domain. We demand one objective to minor the discrepancy of the target domain and the source domain, and hope that the model could transfer the source-domain knowledge to the target domain. We, therefore, introduce the adversarial loss [Tsai et al., 2018] to minor the discrepancy of the source domain and the target domain. The adversarial loss is applied to both predictions of the primary classifier and the auxiliary classifier:

\[ L_{p\text{adv}}^{a} = \mathbb{E} \left[ \log(D_{p}(F_{p}(x_{h}^{c}))) + \log(1 - D_{p}(F_{a}(x_{h}^{c}))) \right], \]  

(3)

\[ L_{a\text{adv}}^{a} = \mathbb{E} \left[ \log(D_{a}(F_{a}(x_{h}^{c}))) + \log(1 - D_{a}(F_{a}(x_{h}^{c}))) \right], \]  

(4)

where \( D \) denotes the discriminator. In this work, we deploy two different discriminators, i.e., \( D_{p} \) and \( D_{a} \), for the primary prediction and the auxiliary prediction, respectively. The discriminator is to find out whether the target prediction \( F(x_{h}) \) is close to the source prediction \( F(x_{h}) \) in the semantic space. By optimizing the adversarial loss, we force the model to bridge the inter-domain gap on the semantic level.

### 3.3 Domain-specific learning

However, the segmentation loss and the adversarial loss do not solve the intra-domain inconsistency, especially in the target domain. In the Stage-I, we consider leveraging the uncertainty in the target domain and propose the memory regularization in vivo to enforce the consistency. In the Stage-II, we further utilize the memory to regularize the training and prevent the model overfitting to the noisy pseudo labels.

**Memory regularization.** In this paper, we argue that the model itself could be viewed as one kind of memory module, in that the model memorizes the historical experience. Without introducing extra parameters or external modules, we enforce the model to learn from itself. In particular, we view the input image as the key, and the model as the memory module. Given the input image (key), the model could generate the value by simply feeding forward the key. We could obtain two values by the primary classifier and the auxiliary classifier, respectively. To minor the uncertainty of the model learning on the target domain, we hope that the two values of the same key could be as close to each other as possible, so we deploy the KL-divergence loss:

\[ L_{mr} = - \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{c=1}^{C} F_{a}(x_{h}^{c}) \log(F_{p}(x_{h}^{c})) \]

- \[ - \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{c=1}^{C} F_{p}(x_{h}^{c}) \log(F_{a}(x_{h}^{c})). \]  

(5)

We only apply the memory regularization loss on the target domain \( X_{t} \) and ask the mapping functions \( F_{p} \) and \( F_{a} \) to generate a consistent prediction on the unlabeled target data.

#### Discussion

1. **What is the advantage of the memory regularization?** By using the memory regularization, we enable the model to learn the intra-domain knowledge on the unlabeled target data with an explicit and complementary objective. As discussed in the [Tarvainen and Valpola, 2017; Chen et al., 2018], we could not ensure that the memory always provides a right class prediction for the unlabeled data. The memory mechanism is more likely to act as a teacher model, providing the class distribution based on the historical experience.

2. **Will the auxiliary classifier hurt the primary classifier?** As shown in many semi-supervised methods [Zhang et al., 2018; Tarvainen and Valpola, 2017], the bad-student model also could provide essential information for the top-student models. Our experiment also verifies that the sub-optimal auxiliary classifier could help the primary classifier learning, and vice versa (see Section 4.2).

**Self-training with pseudo labels.** In the Stage-II, we do not use the source data anymore. The model is fine-tuned on the unlabeled target data and mine the target domain knowledge. Following the self-training policy in [Zou et al., 2018; Zou et al., 2019], we retrain the model with the pseudo label \( \tilde{y}_{h}^{c} \). The pseudo label combines the output of \( F_{p}(x_{h}^{c}) \) and \( F_{a}(x_{h}^{c}) \) from the trained model in the Stage-I. In particular, we set the \( \tilde{y}_{h}^{c} = \arg \max (F_{p}(x_{h}^{c}) + 0.5F_{a}(x_{h}^{c})) \). The pseudo segmentation loss could be formulated as:

\[ L_{p\text{seg}}^{p} = - \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{c=1}^{C} \tilde{y}_{h}^{c} \log(F_{p}(x_{h}^{c})), \]  

(6)

\[ L_{p\text{seg}}^{a} = - \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{c=1}^{C} \tilde{y}_{h}^{c} \log(F_{a}(x_{h}^{c})). \]  

(7)

We apply the pixel-wise cross-entropy loss as the supervised segmentation loss. Since most pseudo labels are correct, the model still could learn from the noisy labels. In Section 4.2, we show the self-training with pseudo labels further boosts the performance on the target domain despite the noise in pseudo labels.

**Discussion.** **What is the advantage of the memory regularization in the Stage-II?** In fact, we treat the pseudo labels as the supervised annotations in the Stage-II. However, the pseudo labels contain the noise and may mislead the model to overfit the noise. The proposed memory regularization in the Stage-II works as a smoothing term, which enforces the consistency in the model prediction, rather than focusing on fitting the pseudo label extremely.

### 3.4 Optimization

We integrate the above-mentioned losses. The total loss of the Stage-I and Stage-II training could be formulated as:

\[ L_{S1}(F_{a}, F_{p}, D_{a}, D_{p}) = L_{\text{seg}} + L_{\text{adv}} + \lambda_{mr} L_{mr}, \]  

(8)

\[ L_{S2}(F_{a}, F_{p}) = L_{\text{pseg}} + \lambda_{mr} L_{mr}, \]  

(9)

where \( \lambda_{mr} \) is the weight for the memory regularization. We follow the setting in PSPNet [Zhao et al., 2017] to set 0.5
for segmentation losses on the auxiliary classifier. $L_{seg} = L_{seg}^p + 0.5L_{seg}^p$, $L_{pseg} = L_{pseg}^p + 0.5L_{pseg}^p$. For adversarial losses, we follow the setting in [Tsai et al., 2018; Luo et al., 2019b], and select small weights for adversarial loss terms $L_{adv} = 0.001L_{adv}^p + 0.0002L_{adv}^p$. Besides, we fix the weight of memory regularization as $\lambda_{mr} = 0.1$ for all experiments.

### 3.5 Implementation Details

**Network Architectures.** We deploy the widely-used Deeplab-v2 [Chen et al., 2017] as the baseline model, which adopts the ResNet-101 [He et al., 2016] as the backbone model. Since the auxiliary classifier has been widely adopted in the scene segmentation frameworks, such as PSPNet [Zhao et al., 2017] and modified DeepLab [Tsai et al., 2018; Luo et al., 2019b], for fair comparison, we also applied the auxiliary classifier in our baseline model as well as the final full model. We also insert the dropout layer before the classifier layer, and the dropout rate is 0.1. Besides, we follow the PatchGAN [Isola et al., 2017] and deploy the multi-scale discriminator model.

**Implementation Details.** The input image is resized to $1280 \times 640$, and we randomly crop $1024 \times 512$ for training. We deploy the SGD optimizer with the batch size 2 for the segmentation model, and the initial learning rate is set to 0.0002. The optimizer of the discriminator is Adam and the learning rate is set to 0.0001. Following [Zhao et al., 2017; Zhang et al., 2019b], both segmentation model and discriminator deploy the ploy learning rate decay by multiplying the factor $(1 - \frac{iter}{total-iter})^{0.9}$. We set the total iteration as 100k iterations and adopt the early-stop policy. The model is first trained without the memory regularization for 10k to avoid the initial prediction noise, and then we add the memory regularization to the model training. For Stage-I, we train the model with 25k iterations. We further finetune the model in the Stage-II for 25k iterations. We also adopt the class balance policy in the [Zou et al., 2018] to increase the weight of the rare class, and the small-scale objects. When inference, we combine the outputs of both classifiers $y_i = \arg \max(F_p(x_i) + 0.5F_a(x_i))$. Our implementation is based on Pytorch. We will release our code for reproducibility.

### 4 Experiment

#### 4.1 Dataset and Evaluation Metric

We mainly evaluate the proposed method on the two unsupervised scene adaption settings, i.e., GTA5 [Richter et al., 2017] → Cityscapes [Cordts et al., 2016] and SYNTHIA [Ros et al., 2016] → Cityscapes [Cordts et al., 2016]. Both source datasets, i.e., GTA5 and SYNTHIA, are the synthetic datasets. GTA5 contains 24,966 training images, while SYNTHIA has 9,400 images for training. The target dataset, Cityscapes, is collected in the realistic scenario, including 2,975 unlabeled training images. Besides, we also evaluate the proposed method on the cross-city benchmark: Cityscapes [Cordts et al., 2016] → Oxford RobotCar [Madern et al., 2017]. We follow the setting in [Tsai et al., 2019] and evaluate the model on the Cityscapes validation set/RobotCar validation set. For the evaluation metric, we report the mean Intersection over Union (mIoU), averaged over all classes.

#### 4.2 Ablation Studies

**Effect of the memory regularization.** To investigate how the memory helps both classifiers, we report the results of the single classifier in Table 1. The observation suggests two points: First, memory regularization helps both classifier learning and improves the performance of both classifiers, especially the auxiliary classifier. Second, the accuracy of the primary classifier prediction does not decrease due to the relatively poor results of the auxiliary classifier. The primary classifier also increases by 2.18% mIoU. It verifies that the proposed memory regularization helps to reduce the inconsistency and mine intra-domain knowledge. Furthermore, we report the results of the full model after Stage-I training, which combines the predictions of both classifiers. The full model achieves 45.46% mIoU accuracy, which is slightly higher than the prediction accuracy of the primary classifier. It also indicates that the predictions of the auxiliary classifier and primary classifier are complementary.

**Effect of different losses in Stage-I.** As shown in Table 2, the full model could improve the performance from 37.23% to 45.46% mIoU. When only using the adversarial loss $L_{adv}$, the model equals to the widely-used domain adaptation method [Tsai et al., 2018]. We note that the model only using the memory regularization $L_{mr}$ also achieves significant improvement comparing to the baseline model without adaption. We speculate that the memory regularization helps to mine

| Method           | without $L_{mr}$ | with $L_{mr}$ |
|------------------|------------------|--------------|
| Auxiliary Classifier | 40.04            | 44.45        |
| Primary Classifier      | 43.11            | 45.29        |
| Ours (Stage-I)              | 42.73            | 45.46        |

Table 1: Ablation study of the memory regularization on both classifiers, i.e., the auxiliary classifier and the primary classifier, in the Stage-I training. The result suggests that the memory regularization helps both classifiers, especially the auxiliary classifier. The final results of the full model combine the results of both classifiers, and therefore improve the performance further.

| Method           | $L_{seg}$ | $L_{adv}$ | $L_{mr}$ | mIoU  |
|------------------|-----------|-----------|----------|-------|
| Without Adaptation |   ✓       |   ✓       |   ✓      | 37.23 |
| Adversarial Alignment | ✓         |   ✓       |          | 42.73 |
| Memory Regularization | ✓         |   ✓       |   ✓      | 43.75 |
| Ours (Stage-I)              |   ✓       |   ✓       |   ✓      | 45.46 |

Table 2: Ablation study of different losses in the Stage-I training. We gradually add the adversarial loss $L_{adv}$ and the memory regularization $L_{mr}$ into consideration.

| Method           | $L_{pseg}$ | $L_{mr}$ | mIoU  |
|------------------|-----------|----------|-------|
| Ours (Stage-I)              |   ✓       |   ✓      | 45.46 |
| Pseudo Label         |   ✓       |   ✓      | 47.90 |
| Ours (Stage-II)              |   ✓       |   ✓      | 48.31 |

Table 3: Ablation study of different losses in the Stage-II training. The result suggests that the memory regularization could prevent the model from overfitting to the noise in the pseudo labels.
the target domain knowledge, yielding the better performance on the target domain. After combining all three loss terms, the full model arrives 45.40% mIoU on Cityscapes.

**Effect of different losses in Stage-II.** If we only deploy the pseudo segmentation loss $L_{pseg}$, the model equals to several previous self-training methods [Zou et al., 2018; Zou et al., 2019]. However, this line of previous methods usually demands a well-designed threshold for the label confidence. In contrast, we do not introduce any threshold, but apply the memory regularization to prevent overfitting toward noisy labels. As shown in Table 3, the full model with memory regularization arrives 48.31% mIoU accuracy, which is higher than the result of the model only trained on the pseudo labels. It verifies that the proposed memory regularization also helps the model learning from noisy labels.

**Hyperparameter Analysis.** In this work, we introduce $\lambda_{mr}$ as the weight of the memory regularization. As shown in Fig. 4, we evaluate different weight values $\{0, 0.01, 0.05, 0.1, 0.2, 0.5\}$. We observe that the model is robust to the value of $\lambda_{mr}$. However, when the value $\lambda_{mr}$ is too large or small, the model may mislead to overfitting or underfitting the consistency. Therefore, without loss of generality, we use $\lambda_{mr} = 0.1$ for all experiments.

### 4.3 Comparisons with state-of-the-art methods

**Synthetic-to-real.** We compare the proposed method with different domain adaptation methods on GTA5 $\rightarrow$ Cityscapes (See Table 4). For a fair comparison, we mainly show the results on the same backbone, i.e., DeepLabv2. The proposed method has achieved 48.3% mIoU, which is higher than the competitive methods, e.g., pixel-level alignment [Hoffman et al., 2018], semantic level alignment [Tsai et al., 2018], as well as the self-training methods, i.e., [Zou et al., 2018; Zou et al., 2019]. Compared with the strong source-only model, the proposed method yields +11.1% improvement. Besides, we observe a similar result on SYNTHIA $\rightarrow$ Cityscapes (see Table 5). The proposed method arrives 53.8% mIoU* and 46.5% mIoU*, which is also competitive to other methods. We obtain +11.3% improvement in mIoU accuracy over the baseline.

**Cross-city.** We also evaluate the proposed method on adapting the model between different cities. The two real datasets, i.e., Cityscapes and Oxford RobotCar, are different from collection locations as well as weather conditions. Cityscapes is collected in the sunny days when Oxford RobotCar contains rainy scenarios. As shown in Table 6, the proposed method also achieves competitive results, i.e., 73.9% mIoU.
We propose a memory regularization method for unsupervised scene adaption. Our model leverages the intra-domain knowledge and reduces the uncertainty of model learning. Without introducing any extra parameters or external modules, we deploy the model itself as the memory module to regularize the training. Albeit simple, the proposed method is complementary to previous works and achieves competitive results on two synthetic-to-real benchmarks, i.e., GTA5 \rightarrow \text{Cityscapes and SYNTHIA} \rightarrow \text{Cityscapes}, and one cross-city benchmark, i.e., \text{Cityscapes} \rightarrow \text{Oxford RobotCar}.

### References

[Chen et al., 2017] Liang-Chieh Chen, George Papandreou, Ioannis Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. \textit{TPAMI}, 2017.

[Chen et al., 2018] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Semi-supervised deep learning with memory. In \textit{ECCV}, 2018.

[Cordts et al., 2016] 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 \textit{CVPR}, 2016.

[French et al., 2017] Geoffrey French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. \textit{arXiv:1706.05208}, 2017.

[He et al., 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In \textit{CVPR}, 2016.

[Hinton et al., 2015] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. \textit{NeurIPS}, 2015.

[Hoffman et al., 2018] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. \textit{ICML}, 2018.

[Huang et al., 2018] Haoshuo Huang, Qixing Huang, and Philipp Krahenbuhl. Domain transfer through deep activation matching. In \textit{ECCV}, 2018.

[Isola et al., 2017] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In \textit{CVPR}, 2017.

[Laine and Aila, 2016] Samuli Laine and Tero Aila. Temporal ensembling for semi-supervised learning. \textit{ICLR}, 2016.

[Luo et al., 2019a] Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Significance-aware information bottleneck for domain adaptive semantic segmentation. \textit{ICCV}, 2019.

[Luo et al., 2019b] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In \textit{CVPR}, 2019.

[Madden et al., 2017] Will Madden, Geoff Pascoe, Chris Linagar, and Paul Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. \textit{The International Journal of Robotics Research (IJRR)}, 36(1):3–15, 2017.

[Richter et al., 2016] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In \textit{ECCV}, 2016.

[Ros et al., 2016] 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 \textit{CVPR}, 2016.

[Saito et al., 2018] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In \textit{CVPR}, 2018.

[Shen et al., 2019] Tong Shen, Dong Gong, Wei Zhang, Chunhua Shen, and Tao Mei. Regularizing proxies with multi-adversarial training for unsupervised domain-adaptive semantic segmentation. \textit{arXiv:1907.12282}, 2019.

[Tarvainen and Valpola, 2017] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In \textit{NeurIPS}, 2017.

[Tsai et al., 2018] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schlater, Kihyuk Sohn, Ming-Hsuan Yang, and Mannohman Chandraker. Learning to adapt structured output space for semantic segmentation. In \textit{CVPR}, 2018.

[Tsai et al., 2019] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schlater, and Mannohman Chandraker. Domain adaptation for structured output via discriminative representations. \textit{arXiv:1901.05427}, 2019.

[Vu et al., 2019] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In \textit{CVPR}, 2019.

[Wang et al., 2019] S Wang, L Zhang, W Zuo, and B Zhang. Class-specific reconstruction transfer learning for visual recognition across domains. \textit{TIP}, 2019.

[Wston et al., 2014] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. \textit{arXiv:1410.3916}, 2014.

[Wu et al., 2018] Zuxuan Wu, Xintong Han, Yen-Liang Lin, Mustafa Gokhan Uzunbas, Tom Goldstein, Ser Nam Lim, and Larry S Davis. Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation. In \textit{ECCV}, 2018.

[Wu et al., 2019] Zuxuan Wu, Xin Wang, Joseph E Gonzalez, Tom Goldstein, and Larry S Davis. Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation. In \textit{ECCV}, 2018.

[Yang et al., 2020] Jihan Yang, Ruijia Xu, Ruiyu Li, Xiaojian Qi, Xiaoyong Shen, Guanbin Li, and Liang Lin. An adversarial perturbation oriented domain adaptation approach for semantic segmentation. \textit{AAAI}, 2020.

[Yue et al., 2019] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In \textit{ICCV}, 2019.

### Table 6: Quantitative results on the cross-city benchmark: Cityscapes \rightarrow \text{Oxford RobotCar.}

| Method | mIoU | C|sidewalk | city | building | sky | sign | fence | road | pole | traffic_light | traffic_sign | vegetation | person | rider | car | truck | bus | train | motorcycle | bicycle | mIoU |
|--------|------|---|---------|------|----------|----|------|--------|------|------|-------------|-------------|------------|--------|-------|-----|------|-----|-------|-----------|---------|------|
| Ours (Stage-I) | 95.9 | 71.5 | 86.2 | 69.3 | 31.9 | 87.3 | 57.9 | 88.8 | 64.3 | 72.0 | 75.1 | 100.0 | 68.5 | 89.6 | 72.0 | 60.0 | 66.1 | 60.1 | 58.6 |

\[
\text{Oxford (Stage-I) mIoU:} \quad 95.9, 71.5, 86.2, 69.3, 31.9, 87.3, 57.9, 88.8, 64.3, 72.0.
\]
[Zhang et al., 2018a] Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, and Tao Mei. Fully convolutional adaptation networks for semantic segmentation. In CVPR, 2018.

[Zhang et al., 2018b] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In CVPR, 2018.

[Zhang et al., 2019a] Lei Zhang, Shanshan Wang, Guang-Bin Huang, Wangmeng Zuo, Jian Yang, and David Zhang. Manifold criterion guided transfer learning via intermediate domain generation. TNNLS, 2019.

[Zhang et al., 2019b] Li Zhang, Xiangtai Li, Anurag Arnab, Kuiyuan Yang, Yunhai Tong, and Philip HS Torr. Dual graph convolutional network for semantic segmentation. BMVC, 2019.

[Zhao et al., 2017] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In CVPR, 2017.

[Zou et al., 2018] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In ECCV, 2018.

[Zou et al., 2019] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In ICCV, 2019.