A Prototypical Knowledge Oriented Adaptation Framework for Semantic Segmentation

Haitao Tian, Shiru Qu, and Pierre Payeur

Abstract—**A prevalent family of fully convolutional networks are capable of learning discriminative representations and producing structural prediction in semantic segmentation tasks. However, such supervised learning methods require a large amount of labeled data and show inability of learning cross-domain invariant representations, giving rise to overfitting performance on the source dataset. Domain adaptation, a transfer learning technique that demonstrates strength on aligning feature distributions, can improve the performance of learning methods by providing inter-domain discrepancy alleviation. Recently introduced output-space based adaptation methods provide significant advances on cross-domain semantic segmentation tasks, however, a lack of consideration for intra-domain divergence of domain discrepancy remains prone to over-adaptation results on the target domain. To address the problem, we first leverage prototypical knowledge on the target domain to relax its hard domain label to a continuous domain space, where pixel-wise domain adaptation is developed upon a soft adversarial loss. The development of prototypical knowledge allows to elaborate specific adaptation strategies on under-aligned regions and well-aligned regions of the target domain. Furthermore, aiming to achieve better adaptation performance, we employ a unilateral discriminator to alleviate implicit uncertainty on prototypical knowledge. At last, we theoretically and experimentally demonstrate that the proposed prototypical knowledge oriented adaptation approach provides effective guidance on distribution alignment and alleviation on over-adaptation. The proposed approach shows competitive performance with state-of-the-art methods on two cross-domain segmentation tasks.**

Index Terms—**Domain adaptation, semantic segmentation, prototypical knowledge, unsupervised learning, transfer learning.**

I. Introduction

THE vision and cognition systems endow humans the ability to not only precisely segment objects from familiar scenes, but also transfer learnt visual knowledge to an unknown scene. However, for computer vision models, such as semantic segmentation, cross-domain knowledge transfer remains challenging. With the goal of further boosting state-of-the-art performance [1], [2] for structural prediction, fully convolutional networks (FCNs) [3] have recently dominated the field of semantic segmentation, while involving training on huge pixel-wise labeled datasets in a supervised, end-to-end way. In spite of such massive training, FCN-based segmentation networks yet exhibit inherent limitations for practical use. In applications such as autonomous driving and robotic navigation that need to operate in changing scenarios, trained networks must deal with large appearance gaps, and thus may need to be retrained using a large number of pixel-wise labeled data in the new scenarios to transfer visual knowledge. However, it is difficult to collect and annotate such amounts of data in practice. A recently introduced and particularly appealing workaround [4] consists of utilizing photo-realistic synthetic street-scene images rendered by graphic engines which can simulate various scenarios for supervised network training. Nevertheless, an FCN pre-trained on synthetic datasets will generally perform poorly on real-world datasets (as shown in Fig. 1), which is mainly caused by a domain shift [5] in the data associated with different scenarios. The domain shift cannot be eliminated even with a large number of synthetic data.

Domain adaptation (DA), a transfer learning method [6] that is effective for handling label distribution mismatch between different domains, has been employed for computer vision tasks such as image classification to transfer learnt knowledge from a source domain to a target domain without using any labeled data from that target domain. In semantic segmentation tasks, this idea has been well leveraged to address domain-shift problems between synthetic datasets (source domain) and real-world datasets (target domain). Early researches [7]–[9] looked at the feature-level alignment that aims to align high-dimensional and suppressed features by employing a discrimination network (discriminator) on the feature space. More recently, researchers [10]–[13] paid closer attention to the use of structural information on the output space of FCN and hence obtained better adaptation performance than feature-space based methods. It is also revealed that feature-level adaptation provides limited improvement because DA suffers from encoding various complex visual cues that have been suppressed in the feature space. Given this, feature-based methods tend to resort to more extensive prior statistical knowledge from the target domain such as object proportion,

Manuscript received January 27, 2021; revised July 21, 2021 and September 14, 2021; accepted October 27, 2021. Date of publication November 22, 2021; date of current version November 30, 2021. This work was supported in part by the China Scholarship Council (CSC) under Grant 201906290105 and in part by the Natural Sciences and Engineering Research Council of Canada under the Discovery Grants Program. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sebastian Bosse. *(Corresponding author: Haitao Tian.)*

Haitao Tian is with the School of Automation, Northwestern Polytechnical University, Xi'an 710129, China, and also with the School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON K1N 6N5, Canada (e-mail: htian026@uottawa.ca).

Shiru Qu is with the School of Automation, Northwestern Polytechnical University, Xi'an 710129, China (e-mail: qushiru@nwpu.edu.cn).

Pierre Payeur is with the Sensing and Machine Vision for Automation and Robotic Intelligence Research Laboratory, University of Ottawa, Ottawa, ON K1N 6N5, Canada (e-mail: ppayeur@uottawa.ca).

Digital Object Identifier 10.1109/TIP.2021.3128311

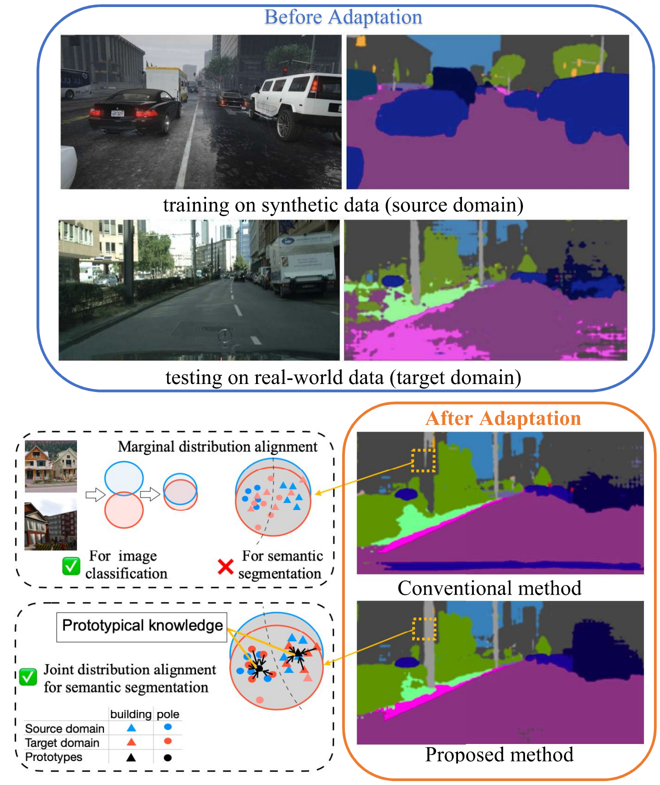

Before Adaptation
training on synthetic data (source domain)
testing on real-world data (target domain)
Marginal distribution alignment
✓ For image classification
✗ For semantic segmentation
After Adaptation
Conventional method
Prototypical knowledge
✓ Joint distribution alignment for semantic segmentation
building
pole
Source domain ▲ ●
Target domain ▲ ●
Prototypes ▲ ●
Proposed method

Fig. 1. Illustration of prototypical knowledge oriented adaptation. Top 2 rows: samples from source and target domain (left), a source-domain pretrained model for semantic segmentation inferred on samples without adaptation (right). An obvious performance drop can be observed on the prediction result of the target domain. **Bottom first row:** marginal distribution alignment is used for domain adaptation for image classification (left), where image labels are adapted by the alignment over global domain features ($P(F[$○$]) = P(F[$○$])$). However, marginal distribution alignment would have negative effectiveness on output-space domain adaptation (right) because the overlook on local appearance variance causes visible over-adaptation results on the target domain, i.e., the segmentations are corrupted after domain adaptation. **Bottom second row:** The proposed adaptation method (left) utilizes prototypical knowledge to develop a joint distribution alignment (e.g., category features from the prototype "building" are aligned by: $P(|$▲$F[$○$]) = P(|$▲$F[$○$])$) and obtains better improvement while it alleviates over-adaptation (right).

street scene layout, static object priors, etc., to reduce the inter-domain discrepancy.

Both DA models for image classification and semantic segmentation tasks assign hard domain labels to samples (images or pixels) and operate an adaptation process by optimizing a binary cross-entropy loss (further detailed in Section III.B.) However, we argue that the hard domain label assignment hinders the DA performance for semantic segmentation tasks and gives rise to over-adaptation performance on the target domain. First of all, different from cross-domain image classification tasks that operate marginal distribution alignment while handling an entire image as an instance, cross-domain semantic segmentation tasks need to align joint distributions by handling each pixel as an instance, to encode local appearance variance in a single image sample (see Fig. 1). Accordingly, inter-domain discrepancy of instances varies widely and causes intra-domain divergence among pixels in a single image sample. However, hard domain label assignment lacks consideration for this observation. Second, given the existence of the possible prior domain semantic consistency between the source and the target domain, a source-trained FCN is able to correctly segment parts of aligned target regions without any adaptation. But binary cross-entropy adaptation loss cannot exempt those regions from over-adaptation. One simple solution to the hard domain label assignment is to deploy specific discriminators for each class on the feature space [9], [14], which aims to disentangle feature encodings for independent adaptation. Nevertheless, such an approach not only introduces a huge number of parameters to the DA network which gives rise to a computational burden during training, but also is limited by adaptation performance on the feature space.

In this paper, we introduce a new domain adaptation loss for cross-domain semantic segmentation tasks, named the soft adversarial adaptation loss, to tackle the over-adaptation problem existing in conventional adaptation methods and improve the adaptation performance. First, prototypical knowledge is used to relax hard domain labels to a continuous domain space where constraints from binary adversarial loss are disentangled and a soft domain label based adversarial adaptation loss is developed. Second, with the soft domain label, we treat local regions on the target domain separately and particularly, thereby proposing distinct adaptation strategies for under-aligned (domain-specific) and well-aligned (domain-invariant) regions, in order to conduct pixel-specific adaptation on the target domain. Lastly, given the uncertainty associated with prototypical knowledge on the output space, we introduce discrimination confidence yielded from a unilateral discriminator to refine the prototypical knowledge. In this way, the adaptation performance is further improved.

The proposed prototypical knowledge oriented adaptation framework draws on two key observations: (1) visual appearance changes, such as illumination and object context differences between the source and target datasets, may give rise to different degrees of inter-domain discrepancy among domain pixels; and (2) there exist embedding spaces in which this variance may be interpreted and measured in a DA network. The main contributions of this work are threefold:

- The introduction of prototypical knowledge for output-space based DA models, which allows DA models to treat target domain pixels particularly and precisely.
- The introduction of an unsupervised output-space based adversarial domain adaptation approach that formulates a soft adversarial adaptation loss without introducing additional parameters beyond that of conventional adversarial loss;
- The introduction of a unilateral discriminator on the feature space, which is trained in a unilateral way instead of adversarial one. It preserves a persistent capability on interpreting inter-domain discrepancy.

II. Related Work

A. Supervised Pixel-Wise Structural Prediction

Semantic image segmentation is important for vision-based applications, e.g., scene understanding, autonomous driving, medical image processing and robotic navigation. Fully convolutional networks [3] have brought significant leaps forward in semantic segmentation against traditional non-CNN models [15], [16]. Some state-of-the-art FCN-based models (e.g., Deeplab v2 [1], PSPNet [2]) achieved impressive performance on semantic segmentation. Meanwhile, benefitting from the increasing number of pixel-wise annotated datasets (e.g., CamVid [17], KITTI [18], MS COCO [19], PASCAL VOC 2012 [20], Mapillary [21], Cityscapes [22]), FCN-based models leverage a supervised network training approach and obtain high accuracy segmentation results. However, the side effect of supervised training is its heavy dependence on well annotated data, while it is impractical to collect datasets with widespread variability that cover various testing scenarios in the real world. Moreover, the annotation of datasets at the pixel level is extremely time-consuming (e.g., annotating one image in Cityscape would require more than 90 minutes). Based on the cost of access to labeled datasets, FCN-based segmentation networks were alternatively trained on synthetic datasets [4], [23], which are readily rendered and annotated by Graphic engines. Nevertheless, due to the domain discrepancy in the visual appearance, models trained on synthetic data tend to perform poorly when running on real-world datasets and scenarios [5].

B. Unsupervised Pixel-Wise Adversarial Domain Adaptation

Domain adaptation (DA) [6] has been successfully leveraged to approach the generalization issue of computer vision tasks such as cross-domain image classification [24], [25] and detection tasks [26], [27]. It is intuitive to implement DA on FCN-based segmentation networks to tackle the problem of domain-shift. A prevalent family of approaches [7]–[14], [28]–[42] specialize in leveraging adversarial adaptation methods to address the pixel-wise adaptation problem by embedding a discrimination network (discriminator) into FCN-based segmentation networks as an adaptation component. For the DA model, the segmentation network not only learns discriminative representations, but also acts as a generative network (generator) interacting with the discriminator for invariant representation learning. In other words, adaptation progress can be realized by adversarial learning [43], where the discriminator is trained to best distinguish the source and target domains encodings, and the generator simultaneously tries to produce invariant outputs that emphasize cases where the discriminator is confused on their classification. As a result of adversarial interaction with the discriminator, the generator manages to produce invariant- discriminative features, such that the segmentation network can be applied on both domains.

There are three major types of DA models for segmentation networks, i.e., image-space based adaptation [8], [28], [30], [40] that considers each pixel of the input image as an adapting instance aiming to generate cross-domain similar looking images; feature-space based adaptation [7]–[9], [14], [29] that considers each spatial grid on the feature layer as an adapting instance; and output-space based adaptation [10]–[13], which considers each element on the soft-max output layer as an adapting instance. For example, the method of Vu *et al.* [12] is an output-space adaptation method which utilizes entropy-based adversarial loss, while Zhou *et al.* [13] recently proposed an output-space adaptation method that introduces affinity space adaptation for semantic segmentation. In comparison to image-space and feature-space based approaches, adversarial DA relying on output space recently yielded better performance due to the usage of spatial information in the output space, though the method shows a certain degree of overfitting adaptation on the target domain.

C. Cluster Assumption for Domain Adaptation

The method proposed in this paper also relates to cluster assumption-based domain adaptation approaches. The cluster assumption states that the feature projections of samples should be clustered around category prototypes so that decision boundaries can go through low-density regions with low entropy. In cross-domain image classification tasks, Shu *et al*. [44] incorporate a virtual adversarial training and conditional entropy minimization to push the decision boundaries away from the target domain. Saito *et al*. [45] utilize the cluster assumption to develop a semi-supervised domain adaptation strategy, while Pan *et al*. [46] compute category prototypes on the source domain, target domain and source-target domain, in order to push the prototypes to be close in the feature space. With regard to cross-domain semantic segmentation tasks, Vu *et al*. [12] utilize entropy minimization and adversarial adaptation simultaneously to encourage unambiguous cluster assignments. The approach was evolved by Chen *et al*. [47] who introduce a maximum squares loss based on cluster assumption. Besides, Lee *et al*. [48] employ adversarial dropout to enforce the cluster assumption on the target domain, by which the decision boundaries are pushed away from the target domain features.

III. Preliminaries

In this section, we introduce mathematical preliminaries on supervised segmentation model settings and unsupervised output-space based domain adaptation settings. Building upon that, the proposed approach is detailed in the next section.

Under the cross-domain adaptation for semantic segmentation settings, let's consider a source domain $\mathfrak{D}_S = \{(\boldsymbol{X}_s^{[i]}, \boldsymbol{Y}_s^{[i]})\}_i^{n_S}$, where $\boldsymbol{X}_s^{[i]} \in \mathbb{R}^{W \times H \times 3}$ is the i th of n_S synthetic dataset images with a size of $W \times H \times 3$ in $\mathfrak{D}_S$, and $\boldsymbol{Y}_s^{[i]} \in \mathbb{R}^{W \times H \times L}$ is a corresponding pixel-wise annotation label with L categories. Also, we denote the target domain in the same way, i.e., $\mathfrak{D}_T = \{(\boldsymbol{X}_t^{[i]})\}_i^{n_T}$, where $\boldsymbol{X}_t^{[i]} \in \mathbb{R}^{W \times H \times 3}$ is a real-world dataset image. Note that there are no available labels in $\mathfrak{D}_T$. The goal of this work is to train an FCN-based segmentation network on the source domain and transpose it to the target domain unsupervisedly. For clarity, $(\boldsymbol{X}_s, \boldsymbol{Y}_s)$ and $\boldsymbol{X}_t$ correspond to a random sample from $\mathfrak{D}_S$ and $\mathfrak{D}_T$ respectively in the rest of this paper. We use $m^{(w,h)}$ to denote the element of a matrix $\boldsymbol{M}$, e.g., $x_s^{(w,h)}$ is the pixel of $\boldsymbol{X}_s$ at the location (w, h).

A. Supervised Semantic Segmentation Problem Settings

An FCN-based segmentation network trains a feature encoder $F(\cdot)$ that learns discriminative features and a dense classifier $C(\cdot)$ that produces structural prediction (i.e., classifying each pixel into a specific category). During network training, a source domain pair $(\boldsymbol{X}_s, \boldsymbol{Y}_s)$ is fed into F, then C takes high-dimensional feature encodings from F and produces a final label soft-max output $C(F(\boldsymbol{X}_s))$. The segmentation loss is the multi-class cross entropy, formulated as:

$$\mathcal{L}_{\text{seg}}^{\text{S}} = -\mathbb{E}_{(\boldsymbol{X}_s,\boldsymbol{Y}_s)\in\mathfrak{D}_S}\Big[\sum\nolimits_{(w,h)}\sum\nolimits_{l} \mathbb{1}_{\left[l=y_s^{(w,h)}\right]} \times \log C(F(x_s^{(w,h)}))^{(l)}\Big] \quad (1)$$

where $l \in \{1,\ldots,\text{L}\}$, $\mathbb{E}[\cdot]$ is the statistical expectation. Ground truth label $\boldsymbol{Y}_s$ is correspondingly encoded into one-hot vector by the indication function $\mathbb{1}[\cdot]$.

As for a target domain sample, $\boldsymbol{X}_t \in \mathfrak{D}_T$, the source-trained segmentation network under the parameter distribution of $\mathfrak{D}_S$ can also generate a direct label soft-max output $C(F(\boldsymbol{X}_T))$, while it would hardly reflect the genuine label distribution of $\mathfrak{D}_T$ because of the presence of domain-shift. Meanwhile, the absence of pixel-wise annotation in $\mathfrak{D}_T$ does not allow for fine-tuning on the trained segmentation network.

B. Unsupervised Output-Space Based Domain Adaptation Problem Settings

The output-space adaptation scheme considers the entire FCN-based segmentation model as a generator $(G = C[F(\cdot)])$, which generates as invariant soft-max output distributions as possible with regard to the two domains. This generalizing capability of G is endowed by a discriminator D, which is embedded on the output space, at the end of the segmentation network, to classify the structured prediction from different domains into a specific domain label. The G and D are updated alternatively by adversarial optimization as follows:

First, the discriminator D has the capability of classifying each element on the soft-max outputs $G(\boldsymbol{X}_s)$ and $G(\boldsymbol{X}_T)$ into its original domain label (we denote source domain label as 1, and target domain as 0). During training, samples from the two domains with hard domain label $(\boldsymbol{X_s}, \mathbf{1})$ and $(\boldsymbol{X_t}, \mathbf{0})$ are propagated forward into the generator and the discriminator successively. In this step, the segmentation network G just takes part in the forward propagation and the discrimination network D is updated with binary cross entropy by back propagation, formulated as:

$$\mathcal{L}_{adv}(D) = -\mathbb{E}_{\boldsymbol{X_s}\in\mathfrak{D}_S}\Big[\sum\nolimits_{(w,h)} \log(D(G(x_s^{(w,h)})))\Big] - \mathbb{E}_{\boldsymbol{X_t}\in\mathfrak{D}_T}\Big[\sum\nolimits_{(w,h)} \log(1 - D(G(x_t^{(w,h)})))\Big] \quad (2)$$

Second, as part of the DA model, the segmentation network G is also expected to generate invariant cross-domain distributions in order to eliminate the inter-domain discrepancy. Optimizing Eq. (3) with sample $(\boldsymbol{X_t}, \mathbf{1})$, G must manage to produce "source-style" soft-max output while it is actually trained with data from the target domain. In this stage, the back propagation is only applied on G, while D is fixed. Combining with supervised training using Eq. (1), the segmentation network G manages to generate discriminative-invariant distributions among pixels within the source and target domains.

$$\mathcal{L}_{adv}(G) = -\mathbb{E}_{\boldsymbol{X_t}\in\mathfrak{D}_T}\Big[\sum\nolimits_{(w,h)} \log(D(G(x_t^{(w,h)})))\Big] \quad (3)$$

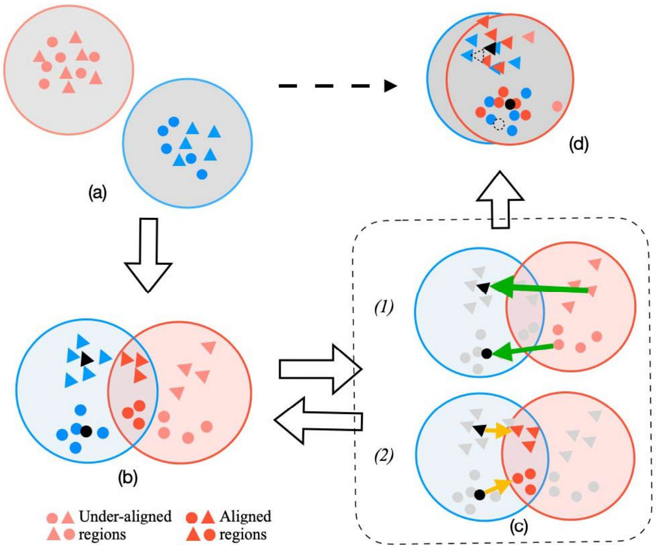


Fig. 2. Visualization of the prototypical knowledge adaptation (PKA) process. Only distribution projections of two classes are illustrated for clarity, and symbols are defined in Fig. 1. Four stages are considered in the adaptation process: (a) before adaptation; (b) supervised learning with prototypes on the source domain; (c) domain adaptation on the target domain; and (d) after adaptation. The training interaction between (b) and (c) is detailed in Section IV.C. The green and orange arrows denote the optimization of objective functions (7) and (9) respectively.

IV. Method

In this section, the proposed prototypical knowledge oriented adaptation framework is defined with the objective to seek more precise guidance on the adaptation process in two ways. First, it extends the idea of conventional output-space DA approaches to a soft domain label-based pixel-specific adaptation strategy. Second, it resorts to a unilateral discriminator to refine the prototypical knowledge to further improve its adaptation performance.

A. Prototypical Knowledge Adaptation (PKA)

In conventional adaptation methods to approach domain shift problems for semantic segmentation, target domain distributions, along with hard domain label, are treated under the same distance from the source domain, by which gradients of the adaptation network are backpropagated in a global manner. After adaptation, the segmentation model would segment aligned target domain regions incorrectly and cause over-adaptation performance (see an example in Fig. 1). In this section, we first investigate the use of prototypical knowledge as an interpreter of soft domain label, then introduce a soft adversarial loss on under-aligned regions and a semi-supervised learning strategy on aligned regions. Our model is built upon conventional output-space DA models without introducing extra parameters.

1) Prototypical Knowledge Interpretation (PKI): After the supervised training of a FCN on the source domain using Eq. (1), a pixel-wise classifier C is obtained and can be degraded into L estimated category prototypes

$\{C_1, \dots C_l, \dots, C_{\mathrm{L}}\}$, such that $\sum_{l=1}^{L} C_l(F(\mathrm{x}_{\mathrm{s}}^{(w,h)})) = 1$. From the perspective of cluster assumption [44], [45], source instances are clustered around class prototypes on the feature embedding (as shown in Fig. 2(b)) with supervision from the source domain ground truth. The category label of each instance $\mathrm{x}_{\mathrm{s}}^{(w,h)}$ is assigned to the nearest prototype according to the confidence $C_l(F(\mathrm{x}_{\mathrm{s}}^{(w,h)}))$. Due to the absence of labels on the target domain, each estimated prototype can be regarded as a representative point of a specific category within the source domain while keeping in the center of the source domain distributions. Therefore, given a target instance projected onto the feature embedding, even though not allowing to make precise prediction, prototypes measure the confidence of a source-trained FCN when generating source-like predictions on the target domain, which is referred to as $C_l(F(\mathrm{x}_{\mathrm{t}}^{(w,h)})$.

Building upon the above observation, we define the prototypical knowledge of the target domain in Eq. (4).

$$\rho_l^{(w,h)} = \Phi[\mathrm{C}_l(\mathrm{F}(\mathrm{x}_{\mathrm{t}}^{(w,h)})) \cdot \tau_l] \tag{4}$$

where $\Phi(\cdot)$ represents a nonlinear transformation followed with sigmoid activation. To relieve the influence of the dominance of high frequency classes, we introduce the balance coefficient, $\tau_l = \sqrt{N^{\psi_l} / \sum_{\psi_l} C_l(F(\mathrm{x}_{\mathrm{t}}^{(w,h)})}$, proposed in [14], where ψ^l is the l-th layer of one-hot outputs of $\boldsymbol{X_t}$, and N^{ψ_l} is the nonzero pixel number of ψ^l. The prototypical knowledge indicates the extent of source-consistent semantic knowledge that a target instance learns from the source domain prototypes on the output space. We also illustrate this prototypical knowledge interpretation (PKI) process in Fig. 3.

2) Soft Adversarial Loss: As mentioned in Section III.B, conventional adaptation methods utilize hard domain labels ($(\boldsymbol{X_s}, \mathbf{1})$ and $(\boldsymbol{X_t}, \mathbf{0})$) to develop the binary adversarial loss function. However, the usage of hard domain labels assigns the membership of domains to a Boolean set regardless of the existence of semantically consistent regions across domains. With the access to prototypical knowledge, it is possible to relax the binary domain membership to a continuous space, where we take the prototypical knowledge of the nearest prototype as the soft domain label of the target domain as:

$$\rho^{*(w,h)} = \max\{\rho_1^{(w,h)}, \dots, \rho_L^{(w,h)}\} \tag{5}$$

With the target domain label $(\boldsymbol{X_t}, \boldsymbol{\rho^*})$, the binary cross-entropy term over a target domain sample in Eq. (2) is transformed into a soft cross entropy loss as defined in Eq. (6):

$$\mathcal{L}_{\mathrm{soft}}^{\mathrm{T}} = -\sum\nolimits_{(w,h)} [\rho^{*(w,h)} \cdot \log(D(G(x_t^{(w,h)}))] \tag{6}$$

3) Domain Adaptation Process: In order to deploy precise guidance on adaptation, we split the soft-max output of target domain into two regions by a threshold T: the under-aligned regions (where $\rho^{*(w,h)} < T$) and the aligned regions (where $\rho^{*(w,h)} \geq T$). We then consider specific adaptation processes on under-aligned regions and on aligned regions respectively, as follows and shown in Fig. 2.

a) PKA on under-aligned regions using soft adversarial loss: Combined with soft adversarial loss $\mathcal{L}_{\mathrm{soft}}^{\mathrm{T}}$, we evolve the conventional adaptation loss from Eq. (2) to an adversarial adaptation loss defined as:

$$\begin{aligned} &\mathcal{L}_{\mathrm{soft\_adv}}(G, D) \\ &\quad = -\mathbb{E}_{\boldsymbol{X_s} \in \mathfrak{D}_S}[\sum\nolimits_{(w,h)} \log(D(G(x_s^{(w,h)})))] \\ &\qquad - \mathbb{E}_{\boldsymbol{X_t} \in \mathfrak{D}_T}[\sum\nolimits_{(w,h) \in [UN]} \rho^{*(w,h)} \cdot \log(D(G(x_t^{(w,h)})))] \end{aligned} \tag{7}$$

This loss function (7) allows to localize adaptation operation over the source domain and the under-aligned regions [UN] of the target domain, meanwhile preventing aligned regions from adaptation by extinguishing gradients backpropagation on these regions.

b) PKA on aligned regions using semi-supervised loss: The semi-supervised learning strategy is used for adapting aligned regions in a non-adversarial way. Semi-supervised learning [41] is already commonly leveraged in DA methods as an efficient supplementary alignment approach with adversarial adaptation. Collaborating with adversarial adaptation that proceeds on under-aligned regions, semi-supervised learning operates the adaptation process on aligned regions.

To conduct the semi-supervised learning of prototypical knowledge adaptation (PKA) on aligned regions [AL], pseudo labels $\hat{y}_t^{(w,h)}$ are first distilled from prototypical knowledge by Eq. (8), where the indication function $\mathbb{1}[\cdot]$ encodes prototypical knowledge into one-hot vectors.

$$\hat{y}_t^{(w,h)} = \mathbb{1}_{\left[l=argmax\{\rho_1^{(w,h)}, \dots, \rho_L^{(w,h)}\}\right]} \tag{8}$$

Second, since entropy minimization has demonstrated complementary effectiveness [42], [47] on the development of semi-supervised learning by encouraging unambiguous cluster alignments, we employ the negative entropy [42] on prototypical knowledge, $H^{(w,h)} = -\sum_{l=1}^{\mathrm{L}} \rho_l^{(w,h)} \cdot log(\rho_l^{(w,h)})$, to penalize over-confident prototypical knowledge values so as to prevent the network from early overfitting on easy-to-transfer regions.

Finally, the semi-supervised learning objective function is formulated in Eq. (9):

$$\begin{aligned} \mathcal{L}_{\mathrm{seg}}^{\mathrm{T}} = -\mathbb{E}_{X_t \in \mathfrak{D}_T}[\sum\nolimits_{(w,h) \in [\mathrm{AL}]} (\hat{y}_t^{(w,h)} \log G(x_t^{(w,h)}) \\ + \alpha H^{(w,h)})] \end{aligned} \tag{9}$$

where α controls the strength of entropy minimization. The overall network training is operated by minimaxing optimization among loss functions, which will be further detailed in Section IV.C.

B. Prototypical Knowledge Refinement (PKR)

The development of soft domain labels builds upon the cluster assumption that prototypical knowledge adaptation is highly related to the intra-domain semantic inconsistency. Nevertheless, the prototypical knowledge would contain implicit uncertainty information when involving uncertain predictions [53] on the output space. For instance, a target-domain pixel might be segmented incorrectly by a FCN learnt on the source domain because of the FCN's inherent inability in segmenting some regions (e.g., class boundaries and regions

[redacted]

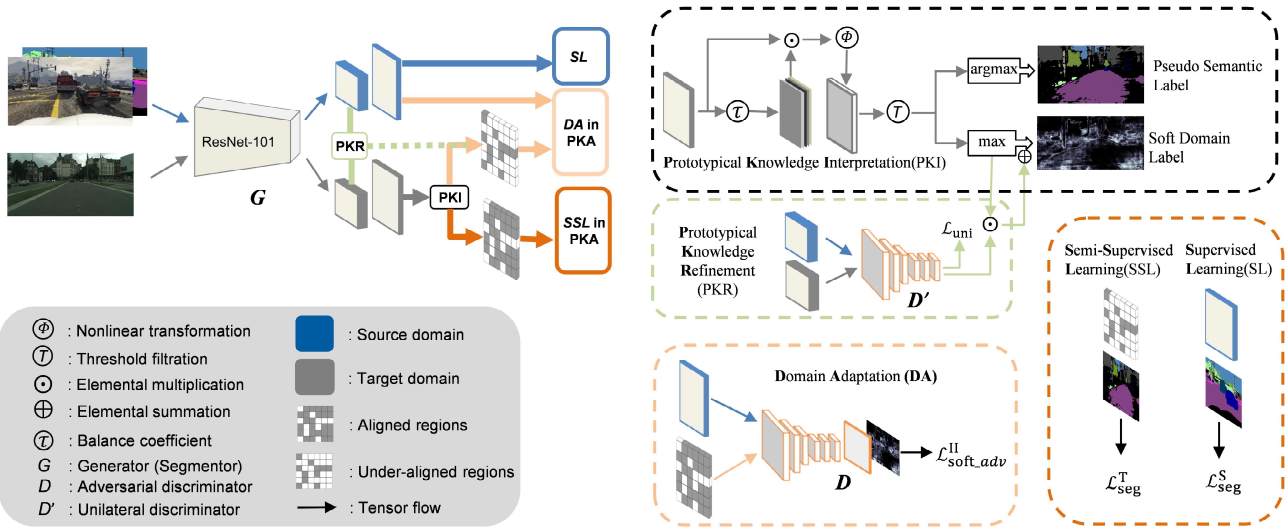

ResNet-101
G
PKR
PKI
SL
DA in PKA
SSL in PKA
Prototypical Knowledge Interpretation(PKI)
argmax
max
Pseudo Semantic Label
Soft Domain Label
Prototypical Knowledge Refinement (PKR)
$\mathcal{L}_{\text{uni}}$
D'
Semi-Supervised Learning(SSL)
Supervised Learning(SL)
Domain Adaptation (DA)
D
$\mathcal{L}^{\text{II}}_{\text{soft\_adv}}$
$\mathcal{L}^{\text{T}}_{\text{seg}}$
$\mathcal{L}^{\text{S}}_{\text{seg}}$
Φ : Nonlinear transformation
τ : Threshold filtration
⊙ : Elemental multiplication
⊕ : Elemental summation
τ : Balance coefficient
G : Generator (Segmentor)
D : Adversarial discriminator
D′ : Unilateral discriminator
: Source domain
: Target domain
: Aligned regions
: Under-aligned regions
: Tensor flow

Fig. 3. A conceptual overview of the proposed prototypical knowledge oriented adaptation (**PKA**) framework. Entire network is parameterized by 3 different fully convolutional networks: generator G (ResNet-101 based segmentation network), and 2 discrimination networks, D and D'.

having large inter-class confusion [49]). In other words, under-aligned regions would contain aligned pixels that are far away from all prototypes, which eventually causes a side effect on the PKA process. To alleviate the influence of uncertain information on the prototypical knowledge, we propose a prototypical knowledge refinement strategy to enhance the performance of PKA.

1) Unilateral Discriminator: A unilateral discriminator D' on the feature space builds upon the assumption that discrimination confidence encoded by this unilateral discriminator can provide intra-domain discrepancy information from a different perspective. Different from conventional discriminators [7], [9], the unilateral discriminator D' is not trained in an adversarial manner but is trained separately by circumventing the interaction with the generator network. As a result, D' will remain discriminative to interpret intra-domain discrepancy (further discussed in Section V.F and Fig. 8).

In detail, the unilateral discriminator D' is trained by Eq. (10) to best discriminate the source and target domains encodings on the feature space.

$$\begin{aligned}\mathcal{L}_{\text{uni}}\left(D'\right) = &-\mathbb{E}_{X_s\in\mathfrak{D}_S}[\sum\nolimits_{(w,h)}\log\left(D'(F(x_s^{(w,h)}))\right)]\\ &-\mathbb{E}_{X_t\in\mathfrak{D}_T}[\sum\nolimits_{(w,h)}\log(1-D'(F(x_t^{(w,h)})))] \quad (10)\end{aligned}$$

In each iteration, it allocates discrimination confidence $D'(F(x_t^{(w,h)})) \in [0, 1]$ to measure the intra-domain discrepancy of each feature grid. With discrimination confidence, we realize a control gate to refine prototypical knowledge, as defined in Eq. (11):

$$\hat{\rho}^{*(w,h)} = (1 + D'(F(x_t^{(w,h)})) \cdot \rho^{*(w,h)} \quad (11)$$

Finally, the soft adversarial loss from Eq. (7) with $\hat{\rho}^{*(w,h)}$ is updated as in Eq. (12).

$$\begin{aligned}&\mathcal{L}^{\text{II}}_{\text{soft\_adv}}\left(G, D\right)\\ &= -\mathbb{E}_{X_s\in\mathfrak{D}_S}[\sum\nolimits_{(w,h)}\log\left(D(G(x_s^{(w,h)}))\right)]\\ &\quad -\mathbb{E}_{X_t\in\mathfrak{D}_T}[\sum\nolimits_{(w,h)\in[\text{UN}]}\hat{\rho}^{*(w,h)}\cdot\log(D(G(x_t^{(w,h)})))] \quad (12)\end{aligned}$$

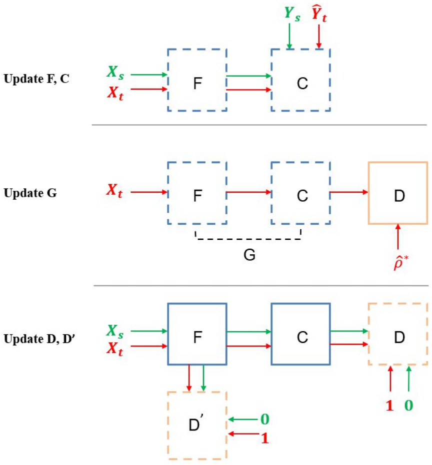


Fig. 4. Illustration of iterative updates of the proposed network. Input data and tensor flow on the source domain are denoted in green. Input data and tensor flow on the target domain are denoted in red. Network embeddings is denoted with dotted lines during back-propagation.

C. Network Overview and Optimization

We detail the proposed prototypical knowledge oriented adaptation framework in Fig. 3. The framework is parameterized by three convolutional networks: a generator G and two discriminators D and D'. In particular, D' is embedded on the final feature layer, and D is embedded on the soft-max output space. Note, G is the identical expression of $C \cdot F$ in the context of an output-space adaptation network.

The interactions between different network embeddings and adversarial optimization are critical to the network performance. During the network training, Eq. (1), (9), (10) and (12) are used to alternatively optimize the adaptation network according to the stages below. Note that there is no specific optimization order for stages in each iteration. The optimization stages are also shown in Fig. 4.

- **Segmenter updating**. At this stage, there are available labeled data from $\mathfrak{D}_S$ and pseudo labels (produced by

Eq. (8)) from $\mathfrak{D}_T$. The segmentation network is trained in a supervised way such that parameters in F and C are updated by minimizing the loss function Eq. (1) and Eq. (9) as follows:

$$\min_{F,C}\,[\mathcal{L}_{\mathrm{seg}}^{\mathrm{S}}(\mathrm{F},\mathrm{C})+\mathcal{L}_{\mathrm{seg}}^{\mathrm{T}}(\mathrm{F},\mathrm{C})] \tag{13}$$

- **Generator updating**. At this stage, unlabeled data in $\mathfrak{D}_T$ are used to optimize the generator G, by maximizing the loss function Eq. (12) as in Eq. (14). The D is fixed at this step. λ_{adv} is the trade-off weight used to balance the segmenter and generator updating processes.

$$\max_{G}\,[\lambda_{\mathrm{adv}}\mathcal{L}_{\mathrm{soft\_adv}}^{\mathrm{II}}(G,D)] \tag{14}$$

- **Discriminators updating**. Data and domain labels in $\mathcal{D}_S$ and $\mathcal{D}_T$ are used to update D and D' simultaneously, so as to best distinguish soft-max output and feature encodings from the source and target domains. At this stage, the generator network is fixed. The loss functions Eq. (10) and Eq. (12) are jointly minimized as in Eq. (15).

$$\min_{D,D}\,[\mathcal{L}_{\mathrm{soft\_adv}}^{\mathrm{II}}(G,D)+\mathcal{L}_{\mathrm{uni}}\left(F,D'\right)] \tag{15}$$

D. Analysis

In this section, we theoretically evaluate the efficacy of the proposed prototypical knowledge oriented adaptation framework compared to conventional adaptation approaches.

Proposition 1: Semi-supervised learning in the PKA introduces a class-wise threshold filtration strategy in the process of pseudo label generation.

Proof: The class-wise threshold filtration strategy was initially proposed in [41], showing better performance than a global thresholding strategy. It sets a class-specific threshold k_l to filtrate pseudo labels for class l by the inequality operation $\mathrm{C}_l\left(\mathrm{F}(\mathrm{x}_{\mathrm{t}}^{(w,h)}\right) > k_l$.

In the PKA, prototypical knowledge on pseudo labels complies with the inequality (16):

$$\rho_l^{(w,h)} > T \tag{16}$$

$$\Leftrightarrow \Phi[\mathrm{C}_l(\mathrm{F}(\mathrm{x}_{\mathrm{t}}^{(w,h)}))\cdot\tau_l] > T$$

$$\Leftrightarrow \mathrm{C}_l(\mathrm{F}(\mathrm{x}_{\mathrm{t}}^{(w,h)})) > \Phi^{-1}(T)/\tau_l \tag{17}$$

The derivation in (17) shows that, even though a global threshold, T, is used in semi-supervised learning, the PKA utilizes a class-specific threshold $k_l=\Phi^{-1}(T)/\tau_l$ for pseudo label generation.

Proposition 2: PKA relaxes constraints from hard domain labels to a soft semantic similarity space, thereby achieving a tighter bound than conventional output-space based adaptation approaches.

Proof: We first revisit the domain adaptation theory which is initially proposed in [50]:

Theorem: Let H be a hypothesis space. $\epsilon_S(\cdot)$ and $\epsilon_T(\cdot)$ are corresponding generalization error functions for domains $\mathfrak{D}_S$ and $\mathfrak{D}_T$. For all $h\in H$:

$$\epsilon_T(h)\le\epsilon_S(h)+\frac{1}{2}d_H(\mathfrak{D}_S,\mathfrak{D}_T)+\lambda \tag{18}$$

where: $\lambda=\epsilon_S(h^*)+\epsilon_T(h^*)$, $\exists h^*=\operatorname*{argmin}_{h\in H}\epsilon_S(h)+\epsilon_T(h)$. In Eq. (18), distance $d_H(\mathfrak{D}_S,\mathfrak{D}_T)$ measures the supremum of the domain distribution divergence over $\mathfrak{D}_S$ and $\mathfrak{D}_T$, which can be further formalized by:

$$d_H(\mathfrak{D}_S,\mathfrak{D}_T)$$
$$=2\sup_{h\in H}\left|Pr_{\mathfrak{D}_S}\left[h\left((\boldsymbol{X_s})=1\right]-Pr_{\mathfrak{D}_T}[h\left((\boldsymbol{X_t})=1\right]\right| \tag{19}$$

In conventional output-space based adaptation tasks, the minimization of $\epsilon_S(h)$ in Eq. (18) is solved by cross-entropy minimization (as formulated in Eq. (1)). Domain distribution divergence $d_H(\mathfrak{D}_S,\mathfrak{D}_T)$ in Eq. (19) is measured by learning a discriminator using Eq. (2), and then $d_H(\mathfrak{D}_S,\mathfrak{D}_T)$ is minimized by optimizing the generator G (as formulated in Eq. (3)). The interactions between the discriminator and the generator form the iterative adversarial learning process in conventional methods.

Taking a closer look on the relationship between $d_H(\mathfrak{D}_S,\mathfrak{D}_T)$ and adversarial adaptation loss in Eq. (2), hypothesis $h(\cdot)$ in Eq. (19) can be denoted as $D(G(\cdot))\in H$, and given hard domain labels $(\boldsymbol{X_s},\boldsymbol{1})$ and $(\boldsymbol{X_t},\boldsymbol{0})$, the domain distribution divergence $d_H(\mathfrak{D}_S,\mathfrak{D}_T)$ can be rewritten as in Eq. (20):

$$d_H(\mathfrak{D}_S,\mathfrak{D}_T)=2\sup_{D\in H}\left|\begin{array}{l}Pr_{\mathfrak{D}_S}\left[D\left(G\left((\boldsymbol{X_s})\right)=1\right]\right.\\ \left.+Pr_{\mathfrak{D}_T}\left[D\left(G\left((\boldsymbol{X_t})\right)=0\right]-1\right.\end{array}\right| \tag{20}$$

We shall see that PKA will achieve a tighter bound in $H\Delta H$ space [51] than conventional adversarial adaptation methods do in H space.

In PKA, soft domain labels $(\boldsymbol{X_s},\boldsymbol{1})$ and $(\boldsymbol{X_t},\boldsymbol{\rho^*})$ are used to evolve binary adversarial loss. If we set:

$$\Gamma(\boldsymbol{X})=\begin{cases}\boldsymbol{1}, & if\ \boldsymbol{X}\in\mathfrak{D}_S\\ \boldsymbol{\rho^*}, & if\ \boldsymbol{X}\in\mathfrak{D}_T,\end{cases} \tag{21}$$

the proposed soft adversarial loss (7) is subject to the $H\Delta H$ distance [51]:

$$d_{H\Delta H}(\mathfrak{D}_S,\mathfrak{D}_T)$$
$$=2\sup_{D\in H,\Gamma\in H}\left|\begin{array}{l}Pr_{\mathfrak{D}_S}\left[D\left(G\left(x_s(\boldsymbol{X_s})\right)=\Gamma\left(\boldsymbol{X_s}\right)\right]\right.\\ \left.-Pr_{\mathfrak{D}_T}[D\left(G\left((\boldsymbol{X_t})\right)=\Gamma\left(\boldsymbol{X_t}\right)\right]\right.\end{array}\right| \tag{22}$$

If we define $D(G(\boldsymbol{X}))\triangle\Gamma(\boldsymbol{X})=\{D^*:D^*(X^*)=1\}$, where $X^*=\{\boldsymbol{X}:D(G(\boldsymbol{X}))\oplus\Gamma(\boldsymbol{X})\}$, $\oplus$ is *XOR* operator, we can denote $d_{H\Delta H}(\mathfrak{D}_S,\mathfrak{D}_T)$ as in Eq. (23):

$$d_{H\Delta H}(\mathfrak{D}_S,\mathfrak{D}_T)$$
$$=2\sup_{D^*\in H\Delta H}\left|\begin{array}{l}Pr_{\mathfrak{D}_S}\left[D^*\left(G\left((\boldsymbol{X_s})\right)=1\right]\right.\\ \left.+Pr_{\mathfrak{D}_T}\left[D^*\left(G\left((\boldsymbol{X_t})\right)=0\right]-1\right.\end{array}\right| \tag{23}$$
$$\le 2\sup_{D\in H}\left|\begin{array}{l}Pr_{\mathfrak{D}_S}\left[D\left(G\left((\boldsymbol{X_s})\right)=1\right]\right.\\ \left.+Pr_{\mathfrak{D}_T}\left[D\left(G\left((\boldsymbol{X_t})\right)=0\right]-1\right.\end{array}\right| \tag{24}$$

In Eq. (23), PKA shows a similar theoretical basis with conventional adaptation methods which is formulated in Eq. (22), aiming to develop a discriminator D^* that is able to best distinguish encodings from source and target domains. However, in the inequality of Eq. (24), by reducing $d_{H\Delta H}$, PKA yields a tighter bound [51] over conventional methods that are developed on H hypothesis space. In this way, by relaxing constraints from hard domain labels, PKA obtains better adaptation performance compared with conventional methods.

V. Experiments

In this section, we experimentally evaluate the efficacy of the proposed framework by exploiting cross-domain semantic segmentation tasks, and analyze the results qualitatively and quantitatively with comparisons to state-of-the-art approaches.

A. Datasets

Cityscapes [22] is a real-world street scene dataset collected by dash cameras mounted on a moving car wandering in European cities. It contains 2,975 training images and 500 validation images, with high resolution (2048 × 1024), and pixel-wise labels in 34 categories of street objects. This dataset is not used for training but rather the training set (without labels) is considered as the target domain.

GTA5 [23] is a synthetic street scene dataset extracted from a realistically rendered computer game: Grand Theft Auto V. As rendered and annotated by a Graphic engine, it forms a large dataset with 24,966 images with high resolution (1914 × 1052), and pixel-wise labels in 19 of the 34 categories of Cityscape. The entire image set with ground truth labels is used as the source domain in the task "GTA5 to Cityscapes."

SYNTHIA-RAND-CITYSCAPES [4] contains 9,400 synthetic images with 16 of the 19 categories of GTA5. The resolution of each image is 760 × 1280. While this dataset is also rendered by a Graphic engine, it is less realistic than GTA5 and uses different viewing angles. That presents more severe domain shift that will challenge the adaptation capability of the proposed model. This dataset is used as the source domain in the task "SYNTHIA to Cityscapes."

B. Implementation Details

The proposed network is deployed using PyTorch on a NVIDIA GTX 1080Ti GPU. We use DeepLab-v2 [1] with ResNet-101 as FCN backbone for the segmentation network. Discriminator D comprises four convolutional layers with stride 2 and kernel size 4 followed by a leaky ReLU, and a classifier layer with same stride and kernel size. The discriminator D' comprises three convolutional layers with stride 1 and kernel size 1 followed by a leaky ReLU, and a classifier layer with kernel size 1. Before feeding into D and D', the encodings from output space and feature space are up-sampled to the size of the input image: $W \times H$. The hyperparameter λ_{adv} is initialized as 0.001 and decayed by a damping policy with the multiplier $(1 - \frac{iter.}{\max\_iter.})$. In our best model, α is set as 0.005.

C. Baselines

The proposed method is able to readily combine with vanilla DA paradigms, while keeping the same number of network parameters. We consider two output-space based domain adaptation models as baselines for our experiments:

- **ASNet** [10] (CVPR 2018) is the first output-space based DA method which has been widely adopted as the baseline model in the state-of-the-art works [11]–[13]. We use ASNet (single mode) with implementation details reported in [10] as our "Baseline I. "
- MRNet [53] (IJCAI 2020) is a recently proposed method to evolve the ASNet with a simple but efficient memory regularization module. It does not introduce any extra parameters compared to [10]. We adopt MRNet (State1) with corresponding implementations as our "Baseline II."

D. Adaptation From GTA5 to Cityscape

The overall quantitative experimental results over 19 classes are detailed in Table I. It shows that when the model is trained on data from the source domain (GTA5) only ("without adaptation") and inferred on the target domain (Cityscapes), the performance measured as the mean Intersection over Union (mIoU) [18] reaches 36.6%. The rest of Table I are adaptation strategies from the literature compared to the proposed PKA and PKA + PKR (denoted as PKA+ in the rest of this paper) respectively. In this section, we experimentally compare the proposed methods with the state-of-the-art approaches in two ways.

1) Comparison With the Baselines: Results in Table I show that both baselines [10] and [53] surpass the without-adaptation model by a clear margin. The latter achieves an 8.9% improvement on mean IoU from the without-adaptation model. The PKA models (PKA (Baseline I) and PKA (Baseline II)) are developed upon Baseline I [10] and Baseline II [53], respectively, while keeping the same number of network parameters. By exploiting the prototypical knowledge on the output space and proceeding with pixel-specific adaptation strategy, PKA outperforms the corresponding Baseline I and Baseline II by a margin of 5.7% and 3.7% on mean IoU, respectively. Close inspection of the specific classes reveals that, PKA outperforms the baselines in most categories (e.g., "wall," "terrain," "bus" and "bike").

2) Comparison With State-of-the-Art: Given that PKA relies on a different joint domain alignment strategy than another recently introduced pixel-wise weighted adaptation approach CLAN [11] that evolves the method in Baseline I [10], we also include a comparison with [11]. From the perspective of methodology, Luo *et al.* [11] employ two classifiers and consider the cousin-distance as the inter-domain discrepancy metric, while PKA only employs one classifier and considers more accessible prototypical knowledge on the output space, demonstrating a more intuitive and simpler adaptation network. From the experimental results in Table I, PKA (Baseline I) demonstrates better adaptation performance than [11] both in terms of overall mIoU and classes-specific IoU. In the overall comparison with other methods, PKA models outperform recently proposed output-space based methods [12], [13], [54]

TABLE I

EXPERIMENTAL RESULTS OF THE ADAPTATION TASK "GTA5 TO CITYSCAPES" UNDER THE METRICS OF IOU ON EACH CLASS AND MEAN IOU ON OVERALL PERFORMANCE. THE HIGHLIGHTED RESULTS SHOW BEST IOU AND BEST MEAN IOU IN EACH COLUMN

-GTA5 to Cityscapes-		road	side.	build.	wall	fence	pole	light	sign	vege.	terrain	sky	person	rider	car	truck	bus	train	moto.	bike	mIoU
Without adaptation		75.8	16.8	77.2	12.5	21.0	25.5	30.1	20.1	81.3	24.6	70.3	53.8	26.4	49.9	17.2	25.9	6.5	25.3	36.0	36.6
Adversarial Domain Adaptation																					
feature space	ASNet (fea.) [10]	83.7	27.6	75.5	20.3	19.9	27.4	28.3	27.4	79.0	28.4	70.1	55.1	20.2	72.9	22.5	35.7	8.3	20.6	23.0	39.3
	SIBAN [29]	88.5	35.4	79.5	26.3	24.3	28.5	32.5	18.3	81.2	40.0	76.5	58.1	25.8	82.6	30.3	34.4	3.4	21.6	21.5	42.6
	SSF-DAN [14]	90.3	**38.9**	81.7	24.8	22.9	30.5	37.0	21.2	84.8	38.8	76.9	58.8	30.7	85.7	30.6	38.1	5.9	28.3	36.9	45.4
output space	ASNet [10]	86.5	25.9	79.8	22.1	20.0	23.0	33.1	21.8	81.8	25.9	75.9	57.3	26.2	76.3	29.8	32.1	7.2	29.5	32.5	41.4
	CLAN [11]	87.0	27.1	79.6	27.3	23.3	28.3	35.5	24.2	83.6	27.4	74.2	58.6	28.0	76.2	33.1	36.7	6.7	31.9	31.4	43.2
	AdvEnt [12]	89.9	36.5	81.6	29.2	25.2	28.5	32.3	22.4	83.9	34.0	77.1	57.4	27.9	83.7	29.4	39.1	1.5	28.4	23.3	43.8
	AdvEnt+Minent [12]	89.4	33.1	81.0	26.6	**26.8**	27.2	33.5	24.7	83.9	36.7	78.8	58.7	30.5	84.8	**38.5**	44.5	1.7	31.6	32.4	45.5
	ASA [13]	89.2	27.8	81.3	25.3	22.7	28.7	36.5	19.6	83.8	31.4	77.1	59.2	29.8	84.3	33.2	45.6	**16.9**	**34.5**	30.8	45.1
	MRNet [53]	89.1	23.9	82.2	19.5	20.1	33.5	42.2	39.7	85.3	33.7	76.4	60.2	33.7	86.0	36.1	43.3	5.9	22.8	30.8	45.5
	IntraDA [54]	90.6	37.1	82.6	30.1	19.1	29.5	32.4	20.6	85.7	40.5	**79.7**	58.7	31.1	86.3	31.5	48.3	0.0	30.2	35.8	46.3
	PKA (baseline I)	88.1	30.7	82.2	**33.1**	23.3	32.2	35.6	28.0	84.4	**43.2**	77.2	61.2	28.2	85.2	34.2	**49.2**	0.0	34.2	**43.3**	47.1
	PKA (baseline II)	**90.4**	33.1	**84.6**	31.7	24.4	**36.4**	**42.2**	**41.3**	**86.8**	40.6	78.8	**64.1**	**36.5**	**87.9**	37.4	48.0	2.3	26.7	40.9	**49.2**
Non-adversarial Domain Adaptation																					
Context Aware [55]		91.3	46.0	84.5	34.4	29.7	32.6	35.8	36.4	84.5	**43.2**	83.0	60.0	32.2	83.2	35.0	46.7	0.0	33.7	42.2	49.2
PIT [56]		87.5	43.4	78.8	31.2	**30.2**	**36.3**	39.9	**42.0**	79.2	37.1	79.3	**65.4**	**37.5**	83.2	**46.0**	45.6	**25.7**	23.5	**49.9**	50.6
MaxSquare [47]		89.4	43.0	82.1	30.5	21.3	30.3	34.7	24.0	85.3	39.4	78.2	63.0	22.9	**84.6**	36.4	43.0	5.5	34.7	33.5	46.4
CBST [42]		91.8	53.5	80.5	32.7	21.0	34.0	28.9	20.4	83.9	34.2	80.9	53.1	24.0	82.7	30.3	35.9	16.0	25.9	42.8	45.9
CRST (MRENT) [42]		91.8	53.4	80.6	32.6	20.8	34.3	29.7	21.0	84.0	34.1	80.6	53.9	24.6	82.8	30.8	34.9	16.6	26.4	42.6	46.1
PKA+ (baseline I)		91.6	50.5	84.4	**34.5**	27.5	32.5	36.4	31.1	85.5	37.1	84.8	60.8	30.2	81.2	28.5	38.5	0.0	38.4	45.7	48.4
PKA + (baseline II)		**93.0**	**55.0**	**85.9**	32.7	30.0	34.5	**40.0**	36.3	**85.7**	40.6	**86.4**	62.5	32.9	82.8	26.7	**50.5**	0.2	**39.8**	49.4	**50.8**

and feature-space based methods [10], [14], [29] in terms of mean IoU, demonstrating the superior effectiveness of the proposed soft adversarial adaptation. When taking a close look on the results of IoU on infrequent classes and small objects in Table I, both the proposed PKA methods and other adversarial learning methods vary in DA performance on different classes, which reflects the different underlying adaptation strategies used by them. Besides, the PKA methods are not only more efficient on specific classes, such as "person," "wall" and "terrain," but also show effectiveness for over-adaptation alleviation, e.g., baselines [10] and [53] exhibit obvious over-adaptation on the class "bike" and "fence" compared to the without-adaptation model, while the PKA methods address it efficiently. Nevertheless, it also can be observed that the PKA methods occasionally underperform other approaches, as exhibited on the class "train." We infer that prototypical knowledge on that category is not effectively utilized as there is a large inter-class confusion [49] on that category in the source domain, which inspires our future works to improve this problem.

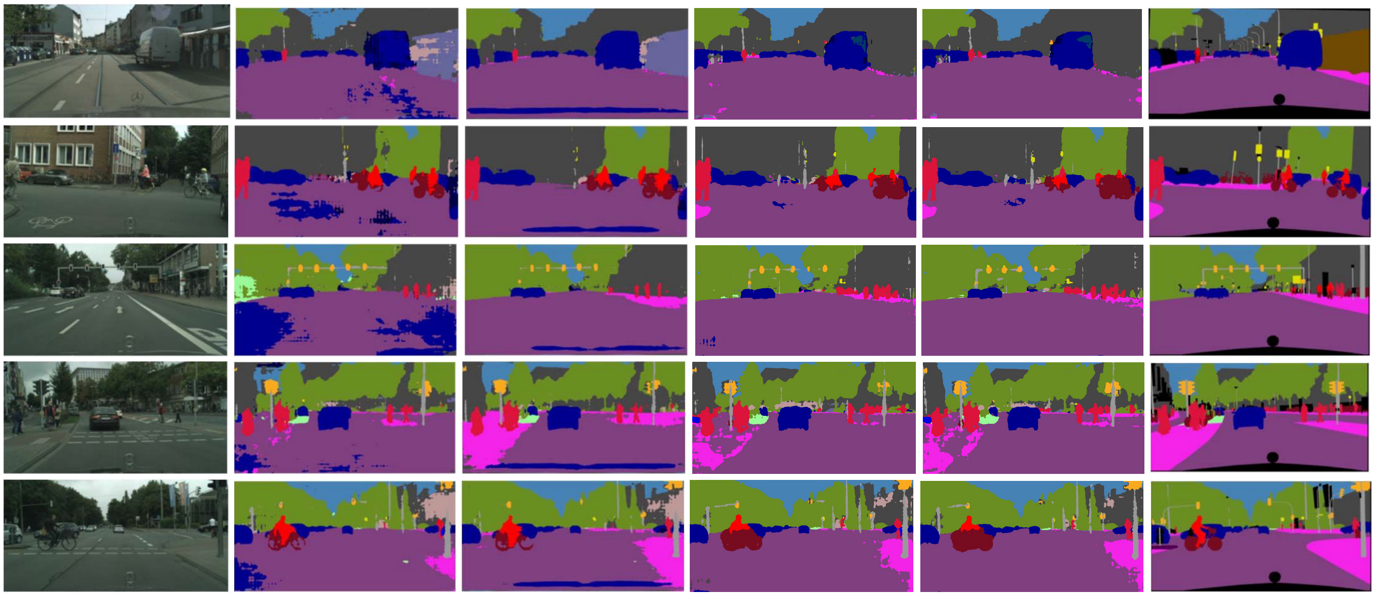
With regard to PKA+, which is proposed to further boost the performance of PKA by employing a unilateral discriminator, the study compares PKA+ (PKA+ (Baseline I) and PKA+ (Baseline II)) with a variety of state-of-the-art non-adversarial adaptation methods. As shown in Table I, the two PKA+ models yield comparable results with non-adversarial DA methods [42], [47], [55], and [56] in terms of mean IoU, showing the convincing effectiveness of the proposed method in the domain adaptation community.

In terms of qualitative experimental evaluation, Fig. 5 illustrates the efficacy of the proposed method. First, it is seen that, compared to a trained model that does not involve adaptation, the proposed methods provide a noticeable improvement in the segmentation. In comparison to [10], there are visible improvements on specific classes in the segmentation results, demonstrating the fundamental adaptation capability of the proposed models on those classes. Second, we visualize the feature clusters formed in the embedding space in Fig. 6 by using t-SNE [57]. We can observe that even though visible adaptation results are achieved from ASNet [10], a certain degree of domain misalignment exists. The proposed PKA and PKA+ provide more uniform and separable feature clusters, resulting in better performance in the target domain.

E. Adaptation From SYNTHIA to Cityscapes

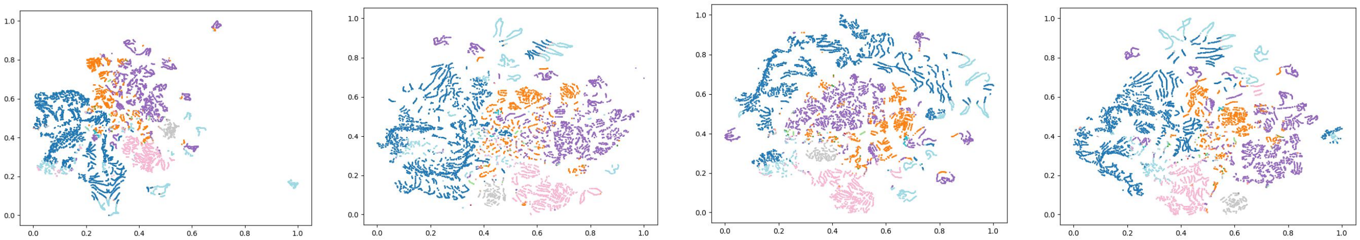
Table II refers to the experimental results of the PKA and PKA+ methods on the task "SYNTHIA to Cityscapes." Similar to [10]–[13], the proposed models are evaluated on 13 classes on Cityscapes. As detailed in Table II, without adaptation, the model (trained on SYNTHIA only) reaches 38.6% on mean IoU. PKA outperforms the two baselines consistently, demonstrating the fundamental efficacy of PKA on this different cross-domain scenario.

When compared to the adversarial methods ([11]–[13], and [54]), similar effectiveness as discussed in "GTA5 to Cityscapes" is observed in Table II. More specifically, the proposed PKA (Baseline I) method shows comparable overall performance to [11] on mean IoU with slightly different performance on IoU for specific classes while leveraging different underlying principles. Besides, the proposed PKA+ (Baseline II) outperforms the non-adversarial methods, Context Aware [55], PIT [56], MaxSquare [47], CBST [42], and



target domain samples | without adaptation | ASNet [10] | PKA (baseline I) | PKA+ (baseline I) | manual annotation

Fig. 5. Qualitative results obtained for semantic segmentation without and with domain adaptation. The third column is the adaptation results shown in ASNet. The fourth and fifth columns are adaptation results obtained with the proposed PKA and PKA+. The last column is the ground truth manual annotations.



without adaptation

ASNet [10]

PKA

PKA+

Fig. 6. Feature clusters visualization in the embedding space by t-SNE [57].

MRENT [42] on mean IoU while being more efficient on some classes, such as “building,” “sky,” and “car.”

F. Ablation Studies

Experimental ablation studies are also conducted in the task “GTA5 to Cityscapes” to investigate the efficacy of the proposed framework.

1) Learning of the Hyper-Parameter T: Parameter T in the formulation (16) is used to split the soft-max output of target domain into the under-aligned regions and the aligned regions. Setting an appropriate value of T not only involves sufficient aligned pixels for semi-supervised learning, but also alleviates the over-adaptation problem in the adversarial adaptation process. We tested PKA+ on both baselines by altering T over the range [0.7, 1]. As illustrated in Fig. 7, the adaptation performance is affected as T varies. The best performance is achieved when T is around 0.92. When $T = 1$, PKA+ degrades to a single adversarial adaptation strategy.

2) Evaluation of the Efficacy of Each Component: Without placing any constraints on the baseline networks and introducing additional parameters, PKA can facilitate the development of output-space based DA works with soft adversarial adaptation loss and lead to efficient adaptation improvements. Experimental results in Table III first show that the proposed PKA achieves a significant gain on mIoU from Baseline I and Baseline II, respectively. Second, a domain confidence (**DC**) based refinement strategy is proposed for **PKA** to improve the adaptation performance by refining prototypical knowledge. When combining with **DC** (shown as **PKA+** in Table III), the **PKA**’s mIoU elevates from 47.1% to 48.4% on Baseline I, and from 49.2% to 50.8% on Baseline II, which demonstrates the fundamental efficacy on the refinement component. Furthermore, given that discrimination confidence (**DC**) is able to measure the intra-domain discrepancy, we experimented

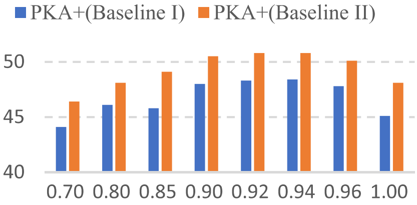


Fig. 7. Performance (on mIoU) of PKA+ while altering the hyperparameter T.

TABLE II

EXPERIMENTAL RESULTS OF THE ADAPTATION TASK "SYNTHIA TO CITYSCAPES" UNDER THE METRICS OF IOU ON EACH CLASS AND MEAN IOU ON OVERALL PERFORMANCE

-SYNTHIA to Cityscapes-		road	side.	build.	light	sign	vege.	sky	person	rider	car	bus	moto.	bike	mIoU
Without adaptation		55.6	23.8	74.6	6.1	12.1	74.8	79.0	55.3	19.1	39.6	23.3	13.7	25.0	38.6
Adversarial Domain Adaptation															
feature space	ASNet (fea.) [10]	62.4	21.9	76.3	11.7	11.4	75.3	80.9	53.7	18.5	59.7	13.7	20.6	24.0	40.8
	SIBAN [29]	82.5	24.0	79.4	16.5	12.7	79.2	82.8	58.3	18.0	79.3	25.3	17.6	25.9	46.3
	SSF-DAN [14]	84.6	41.7	80.8	11.5	14.7	**80.8**	**85.3**	57.5	21.6	82.0	36.0	19.3	34.5	50.0
output space	ASNet [10]	84.3	42.7	77.5	4.7	7.0	77.9	82.5	54.3	21.0	72.3	32.2	18.9	32.3	46.7
	CLAN [11]	81.3	37.0	80.1	16.1	13.7	78.2	81.5	53.4	21.1	73.0	32.9	22.6	30.7	47.8
	AdvEnt [12]	87.0	44.1	79.7	4.8	7.2	80.1	83.6	56.4	23.7	72.7	32.6	12.8	33.7	47.6
	AdvEnt+Minent [12]	85.6	42.2	79.7	8.7	5.4	80.4	84.1	57.9	23.8	73.3	36.4	14.2	33.0	48.0
	ASA [13]	**91.2**	**48.5**	80.4	5.5	5.2	79.5	83.6	56.4	21.0	80.3	36.2	20.0	32.9	49.3
	MRNet [53]	82.0	36.5	80.4	**18.0**	13.4	81.1	80.8	61.3	21.7	84.4	32.4	14.8	**45.7**	50.2
	IntraDA [54]	84.3	37.7	79.5	9.2	8.4	80.0	84.1	57.2	23.0	78.0	**38.1**	20.3	36.5	48.9
	PKA (Baseline I)	86.1	43.1	81.6	9.4	12.4	80.6	84.1	61.1	23.4	81.4	35.1	23.0	36.4	50.6
	PKA (Baseline II)	84.1	37.7	**82.2**	13.1	**14.1**	81.7	81.0	**63.4**	**24.9**	**87.2**	35.9	**24.9**	44.1	**51.9**
Non-adversarial Domain Adaptation															
Context Aware [55]		82.5	42.2	81.3	18.3	15.9	80.6	83.5	61.4	**33.2**	72.9	**39.3**	26.6	43.9	52.4
PIT [56]		83.1	27.6	81.5	**26.4**	**33.8**	76.4	78.8	**64.2**	27.6	79.6	31.2	**31.0**	31.3	51.8
MaxSquare [47]		82.9	40.7	80.3	12.8	18.2	82.5	82.2	53.1	18.0	79.0	31.4	10.4	35.6	48.2
CBST [42]		68.0	29.9	76.3	22.8	29.5	77.6	78.3	60.6	28.3	81.6	23.5	18.8	39.8	48.9
CRST (MRENT) [42]		69.6	32.6	75.8	23.3	29.5	77.7	78.9	60.0	28.5	81.5	25.9	19.6	41.8	49.6
PKA+ (Baseline I)		86.4	**42.0**	**84.4**	12.4	14.3	81.1	**84.9**	63.1	21.4	**85.8**	37.3	22.4	39.7	51.9
PKA+ (Baseline II)		**86.8**	40.3	83.3	21.6	18.4	**83.2**	84.2	61.0	25.6	85.7	34.5	20.4	**44.3**	**53.0**

TABLE III

ABLATION STUDIES. FDA REFERS TO FEATURE SPACE BASED ADVERSARIAL ADAPTATION [7]. CGAN REFERS TO *Cycle-GAN* [52]. PK AND DC ARE SHORTED FROM PROTOTYPICAL KNOWLEDGE AND DISCRIMINATION CONFIDENCE

	Vanilla DA			**Components**			
Method	ASNet	MRNet	FDA	PK	DC	CGAN	mIoU
Baseline I [10]	✓						41.4
PKA	✓			✓			47.1
PKA+	✓			✓	✓		**48.4**
Baseline II [53]		✓					45.5
PKA		✓		✓			49.2
PKA+		✓		✓	✓		**50.8**
Baseline with DC	✓				✓		43.3
PKA with FDA [7]	✓		✓	✓			47.5
PKA+ with CGAN[52]		✓		✓	✓	✓	**51.2**

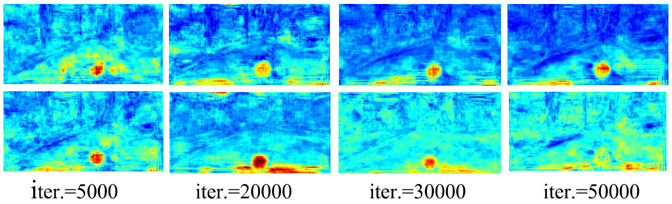


Fig. 8. Discrimination confidence visualization at different iteration steps (Unilateral discriminator (**first row**) vs. Conventional adversarial discriminator (**second row**)).

a separate adaptation approach, **Baseline with DC**, that utilizes the discrimination confidence as soft domain labels and operates the same adaptation strategy as **PKA**. As shown in Table III, **Baseline with DC** provides a mere improvement of 1.9% on mean IoU over Baseline I [10], and shows an inferior performance with respect to **PKA**(Baseline with prototypical knowledge (PK)). It indicates that discrimination confidence interpreted on the feature space is less reliable to estimate soft domain labels than prototypical knowledge.

3) Combination With Other DA Methods: We further explored the efficacy of the proposed method when combining with recent DA methods. First, given that [7]–[9] have demonstrated the effectiveness of DA on the feature space, we modified the D' from a unilateral training discriminator to a general adversarial discriminator for feature-space based domain adaptation [7] (**FDA**), to evaluate the superiority of the two kinds of discriminator when combined with **PKA**. Results in Table III show that, although **PKA with FDA** provides a slight improvement over the sole **PKA** model, it underperforms the proposed **PKA** + (PKA with DC) method. In Fig.8, we visualize discrimination confidence maps encoded by the unilateral discriminator and the adversarial discriminator at different iterations. It can be observed that the adversarial discriminator tends to be indiscriminate when the generator produces domain-invariant features with increasing adversarial training iterations number. However, the proposed unilateral discriminator is trained independently without interactions with the generator G, thereby being able to remain discriminative on domain discrepancy information. Lastly, we assembled the **PKA+** with an image-space based DA method [8], which transferred the GTA5 dataset into a Cityscapes-style dataset using *Cycle-GAN* [52]. With this strategy, the results in Table III show a further 0.4% improvement in the mean IoU over **PKA+** alone.

4) Choice of the Entropy Regularization Term: Chen *et al.* [47] propose Maximum Squares Loss as a variation of negative

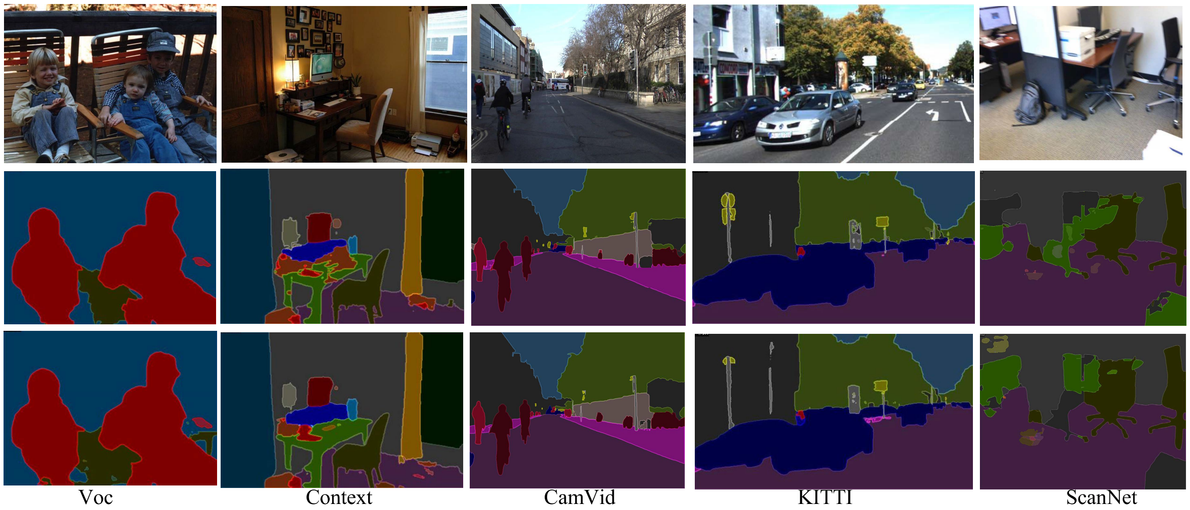


Fig. 9. PKA in real-world domain shifts. From top to bottom: target domain samples, segmentations without domain adaptation, adaptation results obtained with the proposed PKA.

TABLE IV

EXPERIMENTAL COMPARISON BETWEEN ENTROPY MINIMIZATION LOSS ($H(x)$) AND MAXIMUM SQUARES LOSS ($MS(x)$). *Vanilla Model* REFERS TO PKA MODELS TRAINED WITHOUT USING $H(x)$ NOR $MS(x)$

GTA5 to Cityscapes		
	PKA (Baseline I)	PKA (Baseline II)
Vanilla model	46.0%	48.4%
$H(x)$	**47.1%**	49.2%
$MS(x)$	46.9%	**49.3%**

entropy to achieve class-balanced cluster alignments. We also experiment PKA with the Maximum Squares term (defined in (25)) in our experiments.

$$H^{(w,h)} = -\frac{1}{2}\sum\nolimits_{l=1}^{L} (\rho_l^{(w,h)})^2 \tag{25}$$

Experimental results with two types of entropy minimization terms are summarized in Table IV. It can be observed that the effectiveness of the maximum squares loss is on par with that of the negative entropy loss over prototypical knowledge regularization. Therefore, we just use the common negative entropy in all our models.

G. Domain Adaptation Experiments on Practical Scenarios

In this section, we further explore the potential applications of the proposed prototypical knowledge adaptation in more practical and general scenarios. With experimental details, we discuss the effectiveness and generalization ability of the proposed framework.

1) PKA for Real-World Domains: In this part, we exploit PKA on multiple real-world domains with more challenging settings. *MSeg* [58] is a new cross-domain semantic segmentation composite dataset that unifies a variety of real-world datasets from different scenes (“Everyday objects,” “Driving,” and “Indoor”) at a unified taxonomy of 194 categories. It contains 190,231 training images and 12,561 validation images with flexible image resolutions, providing various realistic cross-domain shifts for real-world domain adaptation and generalization applications. Its elaborated taxonomy allows PKA to learn a universal segmentation model over multiple domains and adapt to the testing split (target domain) without using labels from that target domain. In our experiments, the training split [58] of *MSeg* with ground truth labels are set as the source domain. We revisit all training sets of the testing split of *MSeg* to build the target domain. Note that we do not use any images from the testing split of *MSeg* for training, e.g., the testing dataset of *CamVid* still is used for evaluation and the training dataset of *CamVid* is used in the target domain. In implementations, we follow [58] to train a HRNet-W48 [59] segmentation network and construct intra-scene DA pipelines with the proposed PKA method. Experimental results are summarized in Table V. The PKA method outperforms the source-only method on all target datasets with slight variations in different scenes. Meanwhile, qualitative results visualized in Fig 9 show that even though domain shift is less evident in the real-to-real scenarios, PKA achieves noticeable improvements on specific classes in the segmentation results.

PKA demonstrates its independent effectiveness while dealing with the domain shift across the large-scale label space. On the one side, PKA can readily combine with elaborated FCNs, e.g., HRNet-W48 [59], to aggregate semantic knowledge over a diversity of source domains. On the other side, although multiple adaptation scenes are involved over the large-scale label space, e.g., three distinct test scenes are predefined over 194 categories in *MSeg* [58], PKA allows to transfer semantics across domains by utilizing prototypical knowledge from the unlabeled target scenes without placing any constraints, thereby aligning intra-scene domain shifts and obtains better performance than the method without domain adaptation.

2) PKA for Cross-City Test Case: Oxford RobotCar [60] is a real-world street dataset containing 895 training images and 271 validation images collected in rainy scenes, providing another practical test case for our method. We follow

TABLE V

ADAPTATION RESULTS OVER THE *MSeg* DATASET UNDER THE METRICS MEAN IOU ON EACH DATASET AND THE HARMONIC MEAN OVER ALL DATASETS (THE TESTING SPLIT). WE DO NOT INCLUDE *WildDash* IN THE TARGET DOMAIN SINCE IT DOES NOT PROVIDE A TRAINING DATASET. THE SOURCE-ONLY MODEL IS TRAINED WHILE FULLY DEPENDING ON THE SOURCE DOMAIN DATASETS. BOTH THE SOURCE-ONLY MODEL AND PKA ARE TRAINED UNDER UNIFIED TAXONOMY AND EVALUATED ON TEST DATASET TAXONOMY AFTER LABEL MAPPING [57]. BALANCE COEFFICIENT τ_l IN EQ. (4) IS SET AS 1 DURING TRAINING

		Everyday objects		Driving		Indoor	
Method	Backbone	VOC	Context	CamVid	KITTI	ScanNet	*h*. mean
Source-only	*HRNet-W48*	66.1	40.3	83.3	61.5	44.8	55.2
PKA		**68.0**	**41.4**	**84.9**	**61.7**	**46.2**	**56.5**

TABLE VI

EXPERIMENTAL RESULTS OF THE ADAPTATION TASK "CITYSCAPES TO OXFORD ROBOTCAR" UNDER THE METRICS OF IOU ON EACH CLASS AND MEAN IOU ON OVERALL PERFORMANCE

Method	road	sidewalk	building	light	sign	sky	person	automobile	two-wheel	mIoU
Source-only	79.2	49.3	73.1	55.6	37.3	36.1	54.0	81.3	49.7	61.9
PatchAlign [61]	94.4	63.5	82.0	61.3	36.0	76.4	**61.0**	86.5	58.6	72.0
MRNet [53]	**95.9**	73.5	**86.2**	69.3	31.9	87.3	57.9	88.8	**61.5**	72.5
PKA	93.8	**77.9**	86.1	**70.2**	**40.3**	**89.1**	56.0	**91.1**	60.4	**73.9**

TABLE VII

EXPERIMENTAL RESULTS OF THE ADAPTATION TASK ON VIDEO SEMANTIC SEGMENTATION. WE SET THE 1ST FRAME AS OUR INITIAL KEY FRAME WITH A DURATION LENGTH OF 5, AND MEASURE MIOU ON THE 20TH FRAMES

	Per-frame approaches				Key-frame approach	
Backbone	*Resnet-101*		*MobileNet-V2*		*Resnet-101-FlowNetV2*	
Metric	mIoU	runtime	mIoU	runtime	mIoU	runtime
Source-only	36.6	1.5 fps	35.2	20.7 fps	36.5	10.5 fps
PKA	**47.1**		**44.5**		**46.9**	

Tsai *et al.* [61] to conduct the domain adaptation experiment: "Cityscapes to Oxford RobotCar" with PKA (Baseline II). Experimental results in terms of mean IoU are summarized in Table VI. It can be observed that the PKA method surpasses the source-only model by a clear margin, demonstrating the efficacy of our method on this different cross-domain scenario. When compared to state-of-the-art methods, PKA performs more efficiently on both general and class-specific adaptation improvements.

3) PKA for Video Semantic Segmentation: We further show that the proposed PKA can facilitate video segmentation scenario by being free from using any frame annotations. Due to the exceptional cost of frame-wise labelling for video sequences, video semantic segmentation models are generally trained on sparsely-labeled benchmark datasets in a semi-supervised manner [61]–[66]. To deploy PKA for video domain adaptation, the whole image set of *GTA5* [23] is taken as the source image domain. *Cityscapes* [22] contains 2,975 and 500 snippets for training and validation. Each snippet has 30 frames, of which only the 20$^{\text{th}}$ frame is annotated. We randomly extract one frame from each snippet to form the target video domain.

Based on the proposed PKA framework, we propose two schemes to develop unsupervised video domain adaptation. First, we propose a simple but efficient approach, per-frame inference, via combining PKA with a compact FCN backbone. In detail, we build PKA on a MobileNet-V2 [67] backbone, to train the **PKA-MobileNet** model over the source domain and the target domain. During the inference phase, we directly deploy the trained **PKA-MobileNet** model on each frame of the validation video set of Cityscapes. We observe the unsupervised video segmentation performance in terms of mean IoU and the average of frames per second (fps). Experimental results are summarized in Table VII. Second, we propose a key-frame based solution, **PKA-FlowNet**, to reduce redundant computation among neighboring frames by using frame-to-frame optical flow estimate. It utilizes a strong image segmentation model at keyframes but propagates the semantic features of keyframes to non-key ones. It is built upon the observation that available temporal correlations of high-level semantic concepts exist on consecutive frames [62]–[64]. To train such a key-frame based model, we reuse the off-the-shelf ResNet-101 model (yielded from the "GTA5 to Cityscapes" task) as the semantic backbone on key frames and deploy a pretrained optical flow network, FlowNet V2 [68], to encode the temporal consistency across frames, and then propagate semantic patterns from the temporal consistency to non-key frames by using a wrap operation [62]. Even though the absence of labels in the target domain, pseudo labels distilled from prototypical knowledge are utilized to train FlowNet V2 to obtain temporal correlations. During inference, we are able to deploy the **PKA-FlowNet** model for instant video sequences segmentation.

Experimental results with the proposed approaches are reported in Table VII. First, when directly applying a trained Resnet-101 model onto each frame of the target domain (the

[redacted]

validation video set of *Cityscapes* [22]), both the source-only and the PKA method show slow inference runtime (1.5 fps). However, when trained with MobileNet, the source-only model is able to run faster with 20.7 fps. PKA further adapts MobileNet to the target domain and achieves higher accuracy (44.5% on mIoU) by learning prototypical knowledge from the target domain. Lastly, the key-frame based approach enables PKA to trade-off between inference accuracy and runtime. With estimated optical flow, the key-frame based model eventually achieves 46.9% of mIoU with 10.5 fps.

VI. Conclusion

In this work, we first propose a prototypical knowledge adaptation strategy by utilizing a soft adversarial loss to regularize the output-space adaptation process at a finer detail level. It realizes joint distribution alignment on the output space without introducing additional parameters beyond that of the conventional methods. In order to further improve the adaptation performance, we secondly propose a prototypical knowledge refinement strategy by introducing a unilateral discriminator in the proposed adaptation framework.

The theoretical analysis, experimental results on two baselines and comparison of performance against alternative DA strategies demonstrate that the proposed method is effective for adaptation improvements and is comparable to both adversarial and non-adversarial state-of-the-art methods. Besides, the proposed method is also evaluated under practical test settings with detailed experiment results on two real-world cross-domain scenarios. Future work will involve defining more efficient domain discrepancy metrics and regularization schemes to continue improve the performance of DA on cross-domain scenarios.

References

[1] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[2] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[3] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[4] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, “The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3234–3243.

[5] T. Tommasi, N. Patricia, B. Caputo, and T. Tuytelaars, “A deeper look at dataset bias,” 2015, *arXiv:1505.01257*.

[6] B. Gong, F. Sha, and K. Grauman, “Overcoming dataset bias: An unsupervised domain adaptation approach,” in *Proc. NIPS Workshop Large Scale Visual Recognit. Retr. (LSVRR)*, vol. 3, 2012, pp. 1–5.

[7] J. Hoffman, D. Wang, F. Yu, and T. Darrell, “FCNs in the wild: Pixel-level adversarial and constraint-based adaptation,” 2016, *arXiv:1612.02649*. [Online]. Available: https://arxiv.org/abs/1612.02649

[8] J. Hoffman *et al.*, “Cycada: Cycle-consistent adversarial domain adaptation,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 1989–1998.

[9] Y.-H. Chen, W.-Y. Chen, Y.-T. Chen, B.-C. Tsai, Y.-C.-F. Wang, and M. Sun, “No more discrimination: Cross city adaptation of road scene segmenters,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1–8.

[10] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, “Learning to adapt structured output space for semantic segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7472–7481.

[11] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, “Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2507–2516.

[12] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Perez, “ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2517–2526.

[13] W. Zhou, Y. Wang, J. Chu, J. Yang, X. Bai, and Y. Xu, “Affinity space adaptation for semantic segmentation across domains,” *IEEE Trans. Image Process.*, vol. 30, pp. 2549–2561, 2021.

[14] L. Du *et al.*, “SSF-DAN: Separated semantic feature based domain adaptation network for semantic segmentation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 982–991.

[15] C. Zhang, L. Wang, and R. Yang, “Semantic segmentation of urban scenes using dense depth maps,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2010, pp. 708–721.

[16] J. Tighe and S. Lazebnik, “Superparsing: Scalable nonparametric image parsing with superpixels,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2010, pp. 352–365.

[17] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, “Segmentation and recognition using structure from motion point clouds,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2008, pp. 44–57.

[18] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision meets robotics: The KITTI dataset,” *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.

[19] T.-Y. Lin *et al.*, “Microsoft coco: Common objects in context,” in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2014, pp. 740–755.

[20] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The PASCAL visual object classes challenge: A retrospective,” *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2014.

[21] G. Neuhold, T. Ollmann, S. R. Bulo, and P. Kontschieder, “The mapillary vistas dataset for semantic understanding of street scenes,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 22–29.

[22] M. Cordts *et al.*, “The cityscapes dataset for semantic urban scene understanding,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[23] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, “Playing for data: Ground truth from computer games,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2016, pp. 102–118.

[24] M. Long, Y. Cao, J. Wang, and M. I. Jordan, “Learning transferable features with deep adaptation networks,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 102–118.

[25] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, “Adapting visual category models to new domains,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2010, pp. 213–226.

[26] B. Sun and K. Saenko, “From virtual to reality: Fast adaptation of virtual object detectors to real domains,” in *Proc. Brit. Mach. Vis. Conf.*, 2014, p. 3.

[27] D. Vazquez, A. M. Lopez, J. Marin, D. Ponsa, and D. Geronimo, “Virtual and real world adaptation for pedestrian detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 797–809, Apr. 2014.

[28] S. Sankaranarayanan, Y. Balaji, A. Jain, S. N. Lim, and R. Chellappa, “Learning from synthetic data: Addressing domain shift for semantic segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3752–3761.

[29] Y. Luo, P. Liu, T. Guan, J. Yu, and Y. Yang, “Significance-aware information bottleneck for domain adaptive semantic segmentation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6778–6787.

[30] Y. Chen, W. Li, and L. V. Gool, “ROAD: Reality oriented adaptation for semantic segmentation of urban scenes,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7892–7901.

[31] Z. Wu *et al.*, “DCAN: Dual channel-wise alignment networks for unsupervised scene adaptation,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 535–552.

[32] W. Hong, Z. Wang, M. Yang, and J. Yuan, “Conditional generative adversarial network for structured domain adaptation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1335–1344.

[33] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, “Adversarial dropout regularization,” 2017, *arXiv:1711.01575*.

[34] H. Huang, Q. Huang, and P. Krahenbuhl, “Domain transfer through deep activation matching,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 611–626.

[35] Y. Zhang, Z. Qiu, T. Yao, D. Liu, and T. Mei, “Fully convolutional adaptation networks for semantic segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6810–6818.

[36] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, and K. Kim, “Image to image translation for domain adaptation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4500–4509.

[37] X. Zhu, H. Zhou, C. Yang, J. Shi, and D. Lin, “Penalizing top performers: Conservative loss for semantic segmentation adaptation,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 587–603.

[38] Y. Zhang, M. Ye, Y. Gan, and W. Zhang, “Knowledge based domain adaptation for semantic segmentation,” *Knowl.-Based Syst.*, vol. 193, Apr. 2020, Art. no. 105444.

[39] R. Li, W. Cao, Q. Jiao, S. Wu, and H.-S. Wong, “Simplified unsupervised image translation for semantic segmentation adaptation,” *Pattern Recognit.*, vol. 105, Sep. 2020, Art. no. 107343.

[40] Y. Li, L. Yuan, and N. Vasconcelos, “Bidirectional learning for domain adaptation of semantic segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6936–6945.

[41] Y. Zou, Z. Yu, B. V. Kumar, and J. Wang, “Unsupervised domain adaptation for semantic segmentation via class-balanced self-training,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 289–305.

[42] Y. Zou, Z. Yu, X. Liu, B. V. K. V. Kumar, and J. Wang, “Confidence regularized self-training,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5982–5991.

[43] I. Goodfellow *et al.*, “Generative adversarial nets,” in *Proc. NIPS*, 2014, p. 27.

[44] R. Shu, H. Bui, H. Narui, and S. Ermon, “A DIRT-T approach to unsupervised domain adaptation,” in *Proc. Int. Conf. Learn. Represent. (ICRL)*, Feb. 2018, pp. 1–19.

[45] K. Saito, D. Kim, S. Sclaroff, T. Darrell, and K. Saenko, “Semi-supervised domain adaptation via minimax entropy,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8050–8058.

[46] Y. Pan, T. Yao, Y. Li, Y. Wang, C.-W. Ngo, and T. Mei, “Transferrable prototypical networks for unsupervised domain adaptation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2239–2247.

[47] M. Chen, H. Xue, and D. Cai, “Domain adaptation for semantic segmentation with maximum squares loss,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2090–2099.

[48] S. Lee, D. Kim, N. Kim, and S.-G. Jeong, “Drop to adapt: Learning discriminative features for unsupervised domain adaptation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 91–100.

[49] Y. Zhang, P. David, H. Foroosh, and B. Gong, “A curriculum domain adaptation approach to the semantic segmentation of urban scenes,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 1823–1841, Aug. 2020.

[50] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, “Analysis of representations for domain adaptation,” in *Proc. NIPS*, 2007, p. 137.

[51] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, “A theory of learning from different domains,” *Mach. Learn.*, vol. 79, nos. 1–2, pp. 151–175, May 2010.

[52] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[53] Z. Zheng and Y. Yang, “Unsupervised scene adaptation with memory regularization *in vivo*,” in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 2–9.

[54] F. Pan, I. Shin, F. Rameau, S. Lee, and I. S. Kweon, “Unsupervised intra-domain adaptation for semantic segmentation through self-supervision,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3764–3773.

[55] J. Yang, W. An, C. Yan, P. Zhao, and J. Huang, “Context-aware domain adaptation in semantic segmentation,” in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 514–524.

[56] F. Lv, T. Liang, X. Chen, and G. Lin, “Cross-domain semantic segmentation via domain-invariant interactive relation transfer,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4334–4343.

[57] L. Van der Maaten and G. Hinton, “Visualizing data using t-SNE,” *J. Mach. Learn. Res.*, vol. 9, no. 1, pp. 2579–2605, 2008.

[58] J. Lambert, Z. Liu, O. Sener, J. Hays, and V. Koltun, “MSeg: A composite dataset for multi-domain semantic segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2879–2888.

[59] K. Sun *et al.*, “High-resolution representations for labeling pixels and regions,” 2019, *arXiv:1904.04514*.

[60] Y.-S. Xu, T.-J. Fu, H.-K. Yang, and C.-Y. Lee, “Dynamic video segmentation network,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6556–6565.

[61] Y.-H. Tsai, K. Sohn, S. Schulter, and M. Chandraker, “Domain adaptation for structured output via discriminative patch representations,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1456–1465.

[62] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei, “Deep feature flow for video recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2349–2358.

[63] Y. Liu, C. Shen, C. Yu, and J. Wang, “Efficient semantic video segmentation with per-frame inference,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Aug. 2020, pp. 352–368.

[64] P. Hu, F. Caba, O. Wang, Z. Lin, S. Sclaroff, and F. Perazzi, “Temporally distributed networks for fast video semantic segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8818–8827.

[65] Y. Li, J. Shi, and D. Lin, “Low-latency video semantic segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5997–6005.

[66] S. Jain, X. Wang, and J. E. Gonzalez, “Accel: A corrective fusion network for efficient semantic segmentation on video,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8866–8875.

[67] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[68] F. Reda, R. Pottor, J. Barker, B. Catanzaro. (2017). *Flownet2-Pytorch: Pytorch Implementation of Flownet 2.0: Evolution of Optical Flow Estimation With Deep Networks*. [Online]. Available: https://github.com/NVIDIA/flownet2-pytorch

Haitao Tian is currently pursuing the Ph.D. degree with the School of Electrical Engineering and Computer Science, University of Ottawa, Canada. He is also under a joint Ph.D. program (Cotutelle) with the School of Automation, Northwestern Polytechnical University, China. His research interests include domain adaptation, adversarial learning, deep learning, and image semantic segmentation.

Shiru Qu received the Ph.D. degree in automatic control from Northwestern Polytechnical University, China, in 2002. She is currently a Professor at the School of Automation, Northwestern Polytechnical University. Her research interests are computer vision, object detection and recognition, and image semantic segmentation.

Pierre Payeur received the Ph.D. degree in electrical engineering from Université Laval, Canada. Since 1998, he has been a Professor at the School of Electrical Engineering and Computer Science, University of Ottawa. He is currently the Director of the Sensing and Machine Vision for Automation and Robotic Intelligence Research Laboratory. His research interests include machine vision, smart sensing, automation, robotics, and computational intelligence.

[redacted]