Significance-aware Information Bottleneck for Domain Adaptive Semantic Segmentation

Yawei Luo[1,2], Ping Liu[2], Tao Guan[1,4], Junqing Yu[1,5]*, Yi Yang[2,3]

[1]School of Computer Science & Technology, Huazhong University of Science & Technology
[2]ReLER, University of Technology Sydney [3]Baidu Research [4]Farsee2 Tech. Co.
[5]Center of Network and Computation, Huazhong University of Science & Technology

Abstract

For unsupervised domain adaptation problems, the strategy of aligning the two domains in latent feature space through adversarial learning has achieved much progress in image classification, but usually fails in semantic segmentation tasks in which the latent representations are overcomplex. In this work, we equip the adversarial network with a "significance-aware information bottleneck (SIB)", to address the above problem. The new network structure, called SIBAN, enables a significance-aware feature purification before the adversarial adaptation, which eases the feature alignment and stabilizes the adversarial training course. In two domain adaptation tasks, i.e., GTA5 → Cityscapes and SYNTHIA → Cityscapes, we validate that the proposed method can yield leading results compared with other feature-space alternatives. Moreover, SIBAN can even match the state-of-the-art output-space methods in segmentation accuracy, while the latter are often considered to be better choices for domain adaptive segmentation task.

1. Introduction

Semantic segmentation aims to assign each image pixel a category label. The recent adoption of Convolutional Neural Networks (CNNs) yields various of best-performing methods [26, 6, 31] for this task, but the achievement is at the price of a huge amount of dense pixel-level annotations obtained by expensive human labor. An alternative would be resorting to simulated data, such as computer-generated scenes [37, 38, 29], which can make unlimited amounts of labels available. However, models trained with the simulated images, no matter how perfect they perform on the simulation environment, fail to achieve a same or even similar satisfactory performance on realistic images. The reason behind this performance drop lies in the different data distributions of the two domains, typically known as domain shift [43].

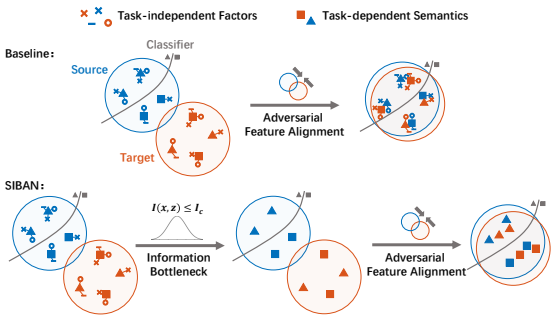


Figure 1: Comparison of the baseline method and SIBAN. The baseline method aligns the latent features directly. As the crude features contain various of task-independent factors, these features are prone to be wrongly aligned between two domains. SIBAN addresses this issue by employing an information bottleneck before the adversarial feature adaptation. The information bottleneck filters out the nuisance factors and maintains pure semantic information. Since the two domains essentially overlap in semantic-level, such purified features can facilitate the following alignment and stabilize the adversarial training course.

Domain Adaptation (DA) approaches [41, 49, 15, 53, 54, 52] are proposed to bridge the gap between the source and target domains. These methods tend to align the two domains in latent feature space so that a classifier trained on source data can also be applied to target samples. Despite the fact that great success has been made on the image level task [27, 50, 34, 11, 23, 28, 33], applying the latent space adaptation to semantic segmentation is non-trivial. The reasons are summarized as twofold. On the one hand, latent space adaptation for semantic segmentation may suffer from the complexity of high-dimensional features which encode

*Corresponding author (yjqing@hust.edu.cn).

This work was done when Yawei Luo (royalvane@hust.edu.cn) was a visiting student at University of Technology Sydney. Part of this work was done when Yi Yang (yee.i.yang@gmail.com) was visiting Baidu Research during his Professional Experience Program.

various visual cues: appearance, shape and context, *etc*. Some of the task-independent nuisance factors might be easily involved in the encoded representation and mislead the domain alignment. On the other hand, in the adversarial domain adaptation framework [12] which becomes popular in this field, the redundant information from the task-irrelevant factors might give excessive cues to the discriminator. The excessive cues lead the discriminator "unnecessary" high accuracy at the wrong time and produce uninformative gradients. All of this, unfortunately, will make the adversarial training process unstable, as pointed out in [31, 19, 35].

Being hampered by the difficulties in feature-space adaptation, the current tendency turns to explore the DA possibility in other spaces, including pixel (input) space and segmentation (output) space. The common idea of the pixel-space adaptations is to force the input images to look like from the same domain, thus decreasing the domain shift from the headstream. While segmentation-space approaches are based on the observation that the segmentation results usually share a significant amount of similarities on the spatial layout and local context. Currently, these two lines of work have produced leading results on the semantic segmentation task while the feature-space adaptation appears eclipsed in front of them. Taking the DA task GTA5 [37] $\rightarrow$ Cityscapes [10] as an example, there is a big difference in segmentation accuracy between the feature-space and output-space adaptation method: 29.2% vs 34.8% [15] on VGG-16 [26], 31.7% vs 37.0% [15] on DRN-26 [51], and 39.3% vs 41.4% [49] on ResNet-101 [14], respectively. The performance gap is so significant that it is justifiable the previous methods choose output-space adaptation as their first choice.

Now a question arises: is the feature-space adaptation really infeasible for the semantic segmentation task? This paper gives a negative answer. As previously analyzed, the obstacles in feature-space adaptation consist in 1) the difficulty of aligning the complicated latent representations between two domains and 2) the difficulty of training the adversarial network stably because of the overly accuracy of the discriminator. Accordingly, we propose Significance-aware Information Bottlenecked Adversarial Network (**SIBAN**), which overcomes the two obstacles above.

Our approach is inspired by the *information bottleneck (IB)* theory [48], where the learned latent representation Z needs to make a consistent prediction with the ground-truth labels Y while simultaneously contains the least mutual information $I(X, Z)$ with the given input X. In our framework, the information bottleneck is employed to compact the complicated latent representations to facilitate the feature alignment and adversarial training.

On the one hand, by enforcing a constraint on the mutual information $I(X, Z)$, we encourage the feature extractor to filter out those task-independent nuisance factors while only keeping the task-dependent factors. In our semantic segmentation task, the task-dependent factor corresponds to the pure semantic information. Since in our *simulated* $\rightarrow$ *real* setting, the two domains vary a lot at *visual level*, but overlap at *semantic level*, such pure semantic information is usually domain-invariant. On the other hand, in the adversarial learning-based framework for adaptation, utilizing the information bottleneck prevents D from the distractions introduced by task-irrelevant factors, which is difficult for the vanilla generator G to depress. As a matter of fact, our proposed network effectively modulates the D's behavior, thus can stabilize the adversarial training process.

Moreover, to deal with the long-tailed data distribution problem [47] introduced by the unbalanced pixel number between different classes, we propose a novel layer, which is named "Significance Aware Layer". By introducing this layer into the IB module, our framework takes the channel-wise significance of each semantic feature into consideration and keeps balanced information constraints between them based on their respective significance. We call this newly designed module as Significance-aware Information Bottleneck (**SIB**), the whole framework as Significance-aware Information Bottlenecked Adversarial Network (SIBAN).

On the whole, our contributions are summarized below.

- We propose a significance-aware information bottlenecked adversarial network (SIBAN) for feature-space domain adaptive semantic segmentation, which combines the advantages from Information Bottleneck theory and Adversarial Learning framework respectively. To our knowledge, this is the first time to successfully utilize information bottleneck strategy for this challenging, dense labeling task.
- We propose a Significance-aware IB (SIB) module and integrate it into our framework. By taking advantage of this module, our framework is able to balance the information constraint between different classes, for maintaining the final performance on the classes which are rare among datasets.
- We theoretically and experimentally prove the effectiveness of our approach, which achieves the leading adaptation result in feature space and performs on par with the state-of-the-art input/output-space adaptations.

2. Related Work

2.1. Domain Adaptive Semantic Segmentation

Ben-David *et al.* [2] have proven that the adaptation loss is bounded by three terms, *e.g.*, the expected loss on source domain, the domain divergence, and the shared error of the ideal joint hypothesis on the source and target domain. Because the first term corresponds to the well-studied supervised learning problems and the third term is considered

sufficiently low, the majority of recent works lay emphasis on the second term. In this spirit, some approaches focus on the distribution shift in the latent feature space [46, 16, 25, 20, 50, 42, 17]. Nevertheless, most of such methods only achieve in classification task while failing in segmentation. With a few exceptions, Hoffman *et al.* [16] employed adversarial network to align the feature representations between domains and additionally appended category statistic constraints to the adversarial model. Apart from the feature-space DA, some methods address the problem in the pixel space [24, 4], which relates to the style transfer approaches [56, 9] to make images indistinguishable across domains. Joint consideration of pixel- and feature-space domain adaptation is studied in [15]. For segmentation task, it is also found that aligning the segmentation space is a more effective DA strategy [49, 7]. Besides the adversarial training-based DA methods [15, 49, 24], other lines of work on semantic segmentation borrow the idea from self-training [39] or co-training [55]. The self-training-based DA [40, 57] attempts to assign pseudo labels to target images and then use these labels to train the target model directly. While the co-training-based DA [41, 30] aims to detect the domain-invariant features by maximizing the consensus of the multiple classifiers.

2.2. Information Bottleneck

Information bottleneck [48] (IB) tends to enforce an upper bound on the mutual information $I(X, Z)$ between the latent representation Z learned by the encoder and the original input X. As pointed in [48], for a supervised learning task, IB encourages Z to be predictive of the label Y, and simultaneously, push the Z to "forget the original input X as much as possible. This is equivalent to upperbound a $Kullback - Leibler(KL-)$ divergence between the joint probability $P(X, Z)$ and the product of the marginals $P(X) \times P(Z)$ to a specific bottleneck value I_c. Although the information bottleneck principle is appealing, it suffers from the fact that mutual information computation is computationally challenging [45], which is especially hard to be instantiated in the context of CNNs. Inspired by a similar approach in variational autoencoders (VAE) [22], recent methods [1, 35] implemented the IB in practical deep models by leveraging a variational bound and the reparameterization trick. This paper follows such strategy to instantiate the IB in the context of adversarial learning-based domain adaptation.

3. Method

3.1. Problem Settings and Overall Idea

We focus on the problem of unsupervised domain adaptation (UDA) in semantic segmentation, where we have access to the labeled source dataset $\{\mathbf{x}_i^s, \mathbf{y}_i^s\}$ and the unlabeled target dataset $\{\mathbf{x}_i^t\}$. The goal is to learn a model G that can correctly predict the pixel-level labels for the target data $\{\mathbf{x}_i^t\}$ by the information from $\{\mathbf{x}_i^s, \mathbf{y}_i^s\}$ and $\{\mathbf{x}_i^t\}$. To facilitate the discussion, we divide G into a feature extractor F and a classifier C, where $G = C \circ F$. Accordingly, we denote the latent representation z as $z = F(\mathbf{x})$ and the final segmentation prediction as $\hat{\mathbf{y}} = C \circ F(\mathbf{x})$.

Traditional feature-level adaptations [16, 15, 49] consider two aspects in dealing with the problem discussed above. First, these methods train a model G to distill knowledge from labeled data by minimizing the task loss in the source domain, which is formalized as a supervised problem:

$$\mathcal{L}_{seg}(F, C) = \mathbb{E}_{\mathbf{x},\mathbf{y}\sim p(\mathbf{x}^S,\mathbf{y}^S)}[\ell(C \circ F(\mathbf{x}), \mathbf{y})] \ , \quad (1)$$

where $\mathbb{E}[\cdot]$ denotes statistical expectation and $\ell(\cdot, \cdot)$ is an appropriate loss function, such as multi-class cross entropy.

Second, during the training process, those feature-level adaptation methods also make F, the submodule in G, to learn domain-invariant features. Ideally, the domain-invariant features should confuse a domain discriminator D which aims at distinguishing the features extracted between the source and target domains. This is achieved by minimaxing an adversarial loss:

$$\begin{aligned}\mathcal{L}_{adv}(F, D) = &- \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}^S)}[\log(D(F(\mathbf{x})))] \\ &- \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}^T)}[\log(1 - D(F(\mathbf{x})))] \ .\end{aligned} \quad (2)$$

However, as mentioned above, there is a significant limitation in previous feature-space adversarial learning methods [16, 15, 49]: there is no explicit constraint to prevent the network from encoding task-independent nuisance factors into the latent features, which makes the adaptation difficult and the adversarial training unstable. To handle the issue, we propose to distill the task-dependent parts from the crude features and conduct the adaptation based on these "purified" representations, thus helping the feature adaptation and stabilizing the adversarial training.

3.2. Information Constrained Domain Adaptation

The pipeline of our network is shown in Fig. 2 where we utilize a simple feature-space adversarial network as the backbone. To purify the encoded latent representation, we adopt an information constraint on the latent space, encouraging F to encode only task-dependent semantic features into the representations. Built upon the recently developed information theories for deep learning [1, 35], we achieve such constraint by employing a variational information bottleneck into the feature extractor F, which is shared among the source domain and target domain respectively. In this case, we obtain the following objective function:

$$\begin{aligned}F^*, C^*, D^* = &\arg\min_{F,C}\max_{D} \mathcal{L}_{seg}(F, C) + \lambda\mathcal{L}_{adv}(F, D) \\ s.t. \quad &\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}^S)}(\mathrm{KL}[F(\mathbf{z}|\mathbf{x})||r(\mathbf{z})]) \leq I_c, \\ &\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}^T)}(\mathrm{KL}[F(\mathbf{z}|\mathbf{x})||r(\mathbf{z})]) \leq I_c.\end{aligned} \quad (3)$$

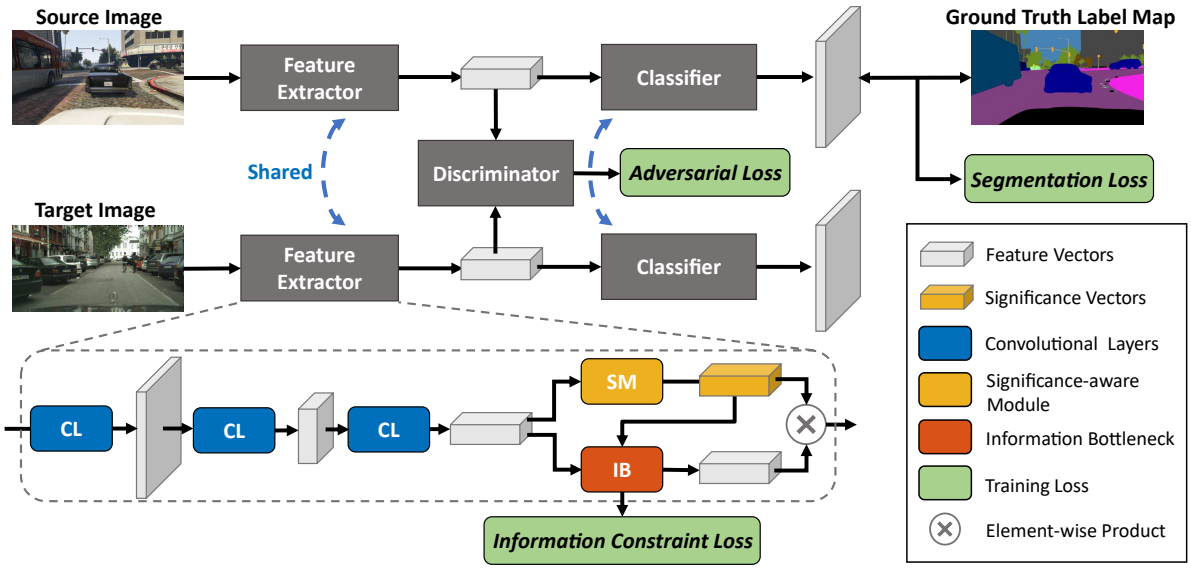


Figure 2: The network architecture of the proposed SIBAN.

where $r(\mathbf{z})$ denotes a prior marginal distribution of $\mathbf{z}$, which is modeled as a standard Gaussian $\mathcal{N}(0; I)$ in this paper. The intuitive meaning of variational IB is clear: the larger the KL-divergence between $F(\mathbf{z}|\mathbf{x})$ and $r(\mathbf{z})$, the stronger the dependence between $\mathbf{x}$ and $\mathbf{z}$, indicating that $\mathbf{z}$ encodes more information from $\mathbf{x}$, in which case some of them might be not task-related and therefore harmful to the adaptation. Therefore, by enforcing the KL- divergence to a threshold I_c and minimizing the task loss, we can explicitly remove the task-independent factors from $\mathbf{z}$.

We can equivalently optimize Eq. 3 by introducing two Lagrange multipliers: $\beta^S \geq 0$ for the source domain, and $\beta^T \geq 0$ for the target domain:

$$\begin{aligned} F^*, C^*, D^* = \arg\min_{F,C}\max_{D} \mathcal{L}_{seg}(F,C) + \lambda\mathcal{L}_{adv}(F,D) + \\ \beta^S(\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}^S)}(\mathrm{KL}[F(\mathbf{z}|\mathbf{x})||r(\mathbf{z})]) - I_c) + \\ \beta^T(\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}^T)}(\mathrm{KL}[F(\mathbf{z}|\mathbf{x})||r(\mathbf{z})]) - I_c). \end{aligned} \tag{4}$$

To simplify the formulation, we define the last two items of Eq. 4 as the information constraint losses $\mathcal{L}_{ic}^S$ and $\mathcal{L}_{ic}^T$ for source and target domain, respectively. Accordingly, the overall training loss can be rewritten as:

$$\begin{aligned} \mathcal{L}_{Overall}(F,C,D) = &\mathcal{L}_{seg}(F,C) + \lambda\mathcal{L}_{adv}(F,D) + \\ &\beta^S\mathcal{L}_{ic}^S(F) + \beta^T\mathcal{L}_{ic}^T(F). \end{aligned} \tag{5}$$

Unlike the traditional information bottleneck methods [5, 1, 13] that uses a fixed β, we follow the suggestion of [35] to adaptively update β^S/β^T via dual gradient descent. The motivation behind is intuitive: the more bias should be given to the feature purification when the encoded information overflow the bottleneck and vice versa, to enforce a specific constraint I_c on the mutual information. Specifically, we train the network to minimax the overall loss $\mathcal{L}_{Overall}(F,C,D)$ by alternating between optimizing F, C, D, β^S and β^T until the loss converges.

$$\begin{aligned} C, F &\leftarrow \arg\min_{C,F} \mathcal{L}_{Overall}(F,C,D) \\ D &\leftarrow \arg\max_{D} \mathcal{L}_{Overall}(F,C,D) \\ \beta^S &\leftarrow \max(0, \beta^S + \alpha\mathcal{L}_{ic}^S) \\ \beta^T &\leftarrow \max(0, \beta^T + \alpha\mathcal{L}_{ic}^T), \end{aligned} \tag{6}$$

where α denotes the step length for updating β^S/β^T.

3.3. Significance-aware Information Bottleneck

The starting point for our significance-aware information bottleneck (SIB) is the observation that segmentation of those infrequent classes is prone to be hurt by the standard IB. We analyze the reason from two folds. On the one hand, for the infrequent classes, the supervision is insufficient to support the network to learn a good representation under the constraint from the bottleneck. On the other hand, from the view of information entropy, the actual encoding of an infrequent sample would span more channels in a feature vector. As the KL-divergence is calculated by summing up the channel-wise losses, the features from those infrequent classes are usually suffered from more powerful constraint. The problem is severe in semantic segmentation task because the class occupations in a scene are highly unbalanced and the latent features are usually high-dimensional. The proposed SIB aims to address such limitation by incorporating the significance-aware mechanism.

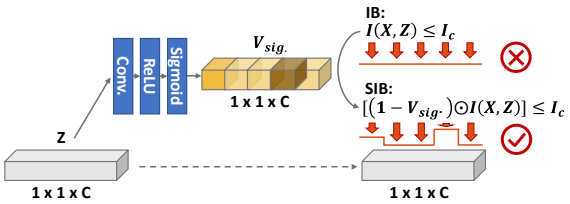


Figure 3: Significance-aware information bottleneck (SIB). We use a significance-aware module to detect the channel-wise significance $\mathcal{V}_{sig.}$ for each pixel-level feature, with which the original information constraint loss is adaptively weighted. The different sizes of red arrows indicate SIB attaches different compression to each channel according to their significance, while the standard IB compress each channel equally.

Fig. 3 details our proposed SIB module. Firstly, we detect the channel-wise significance vector $\mathcal{V}_{sig.}$ for the latent feature. Since we adopt a 1×1 kernel-sized convolutional layer in SIB, here we use a $1 \times 1 \times C$ shaped feature vector within the $w \times h \times C$ shaped feature map for illustration. The information constraint is then adaptively weighted by multiplying $1 - \mathcal{V}_{sig.}$. Taking the source domain features as an example, the significance-aware IB loss can be obtained as

$$\mathcal{L}_{ic}^{S} = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}^S)}[(1-\mathcal{V}_{sig.}) \odot (\mathrm{KL}[F(\mathbf{z}|\mathbf{x})||r(\mathbf{z})] - I_c)], \tag{7}$$

where $\odot$ denotes the channel-wise product. The intuition is that the more significant channels should get less constraint.[1] Such SIB can adaptively decrease the constraint effect on important channels, thus preventing the critical information from being eliminated. Experimental results show the proposed SIB brings a significant improvement over the standard IB in segmentation task, especially for those infrequent classes.

3.4. Network Architecture

Our network architecture is illustrated in Fig. 2. It is composed of a generator G and a discriminator D. G can be any FCN-based segmentation network [44, 26, 6], which is further divided into a feature extractor F and a classifier C. We attach the SIB on the output from last convolutional layer of F. D is a CNN-based binary classifier with a fully-convolutional output [12], which attempts to distinguish whether a latent feature is from source or target domain.

Given a source domain image and the annotation (x_i^S, y_i^S), F is used to extract a latent representation z_i^S and SIB is applied to z_i^S to conduct the significance-aware feature purification. Specifically, we firstly forward z_i^S to the significance-aware module to yield a channel-wise significance vector $V_{sig.}$ for each pixel-level features. Then $V_{sig.}$ together with z_i^S are fed into IB to calculate a significance-weighted KL-divergence between $p(z_i^S)$ and $\mathcal{N}(0; I)$, which is named "information constraint loss". Finally, we multiply z_i^S with $V_{sig.}$ to produce z_{sig}^S, which denotes the final representation of x_i^S. On the one hand, z_{sig}^S is forwarded to C to yield a segmentation loss under the supervision of the ground-truth label y_i^S. On the other hand, z_{sig}^S is input to D to generate an adversarial loss.

Given a target domain image x_i^T, we also forward it to F through SIB and obtain a purified latent representation z_{sig}^T. Different from the source data flow, since we have no access to the target annotation, we only use the adversarial loss and information constraint loss to train the network.

3.5. Theoretical Insight

In this section, we show the relationship between our method and the theory of domain adaptation proposed by Ben-David *et al* [2].

Theorem. Let $\mathcal{H}$ be the hypothesis class, $\mathcal{S}$ and $\mathcal{T}$ denote two different domains, we have the theory bounds the expected error on the target samples $\epsilon^T(h)$ by three terms as follows:

$$\forall h \in \mathcal{H}, \epsilon_{\mathcal{T}}(h) \leq \epsilon_{\mathcal{S}}(h) + \frac{1}{2} d_{\mathcal{H}}(\mathcal{S}, \mathcal{T}) + \lambda, \tag{8}$$

where

$$d_{\mathcal{H}}(\mathcal{S},\mathcal{T}) \triangleq 2 \sup_{h\in\mathcal{H}} \left| \Pr_{\mathbf{x}\sim\mathcal{S}}[h(\mathbf{x})=1] - \Pr_{\mathbf{x}\sim\mathcal{T}}[h(\mathbf{x})=1] \right|,$$

$$\lambda \triangleq \min\left[\epsilon_{\mathcal{S}}(h) + \epsilon_{\mathcal{T}}(h)\right]$$

Here $\epsilon_{\mathcal{S}}(h)$ is the expected error on the source samples which can be minimized easily in a fully-supervised manner, $d_{\mathcal{H}}(\mathcal{S},\mathcal{T})$ denotes a discrepancy distance between source and target distributions *w.r.t.* a hypothesis set $\mathcal{H}$. λ is the shared expected loss and is expected to be negligibly small. This theorem proven by Ben-David *et al* [2] emphasizes the importance of decreasing domain discrepancy for adaptation problem and forms the theoretical basis of our paper.

Corollary. Information bottleneck is trying to optimize the upper bound for $d_{\mathcal{H}}(\mathcal{S},\mathcal{T})$.

Proof. As a distance metric, $d_{\mathcal{H}}(.,.)$ satisfies the following triangle inequality:

$$d_{\mathcal{H}}(\mathcal{S},\mathcal{T}) \leq d_{\mathcal{H}}(\mathcal{S}, r(z)) + d_{\mathcal{H}}(\mathcal{T}, r(z)), \tag{9}$$

where $r(Z) \sim \mathcal{N}_k(0, I)$ and k is the dimension of Z.

Recall that variational IB (see Eq. 4) enforces the distribution of latent feature to approximate a multivariate normal distribution via purification: $\mathcal{S} \rightarrow \mathcal{N}_k(0, I)$ and $\mathcal{T} \rightarrow \mathcal{N}_k(0, I)$, hence forcing the last two terms of Eq. 9

[1] It is noteworthy that we do not back-propagate the information constraint loss to the significance-aware layer. Hence the $\mathcal{V}_{sig.}$ is only trained to minimize the task loss and does not subject to the IB.

Table 1: Adaptation from GTA5 [37] to Cityscapes [10]. We present the results in terms of per-class IoU and mean IoU. "F", "P" and "S" represent the DA applied in feature space, pixel space and semantic space, respectively. "VGG-16" and "ResNet" represent the VGG16-FCN8s and ResNet-101 backbones, respectively. IBAN denotes the baseline network equipped with a standard IB. *Gain* indicates the mIoU improvement over using the source only.

GTA5 → Cityscapes																							
	Space	Arch.	road	side.	buil.	wall	fence	pole	light	sign	vege.	terr.	sky	pers.	rider	car	truck	bus	train	motor	bike	**mIoU**	**gain**
Source only	-		26.0	14.9	65.1	5.5	12.9	8.9	6.0	2.5	70.0	2.9	47.0	24.5	0.0	40.0	12.1	1.5	0.0	0.0	0.0	17.9	-
CyCADA (pixel only) [15]	P		83.5	38.3	76.4	20.6	16.5	22.2	26.2	21.9	80.4	28.7	65.7	49.4	4.2	74.6	16.0	26.6	2.0	8.0	0.0	34.8	16.9
AdaptSeg (seg. only) [49]	S		87.3	29.8	78.6	21.1	18.2	22.5	21.5	11.0	79.7	29.6	71.3	46.8	6.5	80.1	23.0	26.9	0.0	10.6	0.3	35.0	17.1
Source only	-	VGG-16	26.0	14.9	65.1	5.5	12.9	8.9	6.0	2.5	70.0	2.9	47.0	24.5	0.0	40.0	12.1	1.5	0.0	0.0	0.0	17.9	-
FCNs in the wild (feat. only) [16]	F	VGG-16	70.4	**32.4**	62.1	14.9	5.4	10.9	14.2	2.7	79.2	21.3	64.6	**44.1**	4.2	70.4	8.0	7.3	0.0	3.5	0.0	27.1	9.2
CyCADA (feat. only) [15]	F	VGG-16	**85.6**	30.7	74.7	14.4	13.0	17.6	13.7	5.8	74.6	15.8	69.9	38.2	3.5	72.3	16.0	5.0	0.1	3.6	0.0	29.2	11.3
Baseline (feat. only) [49]	F	VGG-16	81.8	23.5	75.2	17.6	12.8	20.3	16.9	**10.8**	76.4	22.6	71.3	43.8	6.5	72.1	20.0	**19.5**	1.2	9.6	0.3	31.7	13.8
IBAN (Ours)	F	VGG-16	84.0	11.1	**80.2**	16.4	14.5	21.1	19.0	7.9	80.6	27.5	76.0	43.8	4.9	**78.5**	16.9	17.3	1.7	8.6	0.0	32.1	14.2
SIBAN (Ours)	F	VGG-16	83.4	13.0	77.8	**20.4**	**17.5**	**24.6**	**22.8**	9.6	**81.3**	**29.6**	**77.3**	42.7	**10.9**	76.0	**22.8**	17.9	**5.7**	**14.2**	**2.0**	**34.2**	**16.3**
Source only	-	ResNet	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	-
AdaptSeg (seg. only) [49]	S	ResNet	86.5	25.9	79.8	22.1	20.0	23.6	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	4.8
Source only	F	ResNet	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	-
Baseline (feat. only) [49]	F	ResNet	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	2.7
IBAN (Ours)	F	ResNet	88.2	33.7	**80.1**	23.4	21.8	27.7	27.9	16.3	**83.2**	38.3	76.2	57.5	20.3	81.1	25.9	33.4	1.9	22.4	20.7	40.7	4.1
SIBAN (Ours)	F	ResNet	**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**	**6.0**

Table 2: Adaptation from Synthia [38] to Cityscapes [10]. The table setting is the same as Table 1.

SYNTHIA → Cityscapes																	
	Space	Arch.	road	side.	buil.	light	sign	vege.	sky	pers.	rider	car	bus	motor	bike	**mIoU**	**gain**
Source only	-		6.4	17.7	29.7	0.0	7.2	30.3	66.8	51.1	1.5	47.3	3.9	0.1	0.0	20.2	-
AdaptSeg (seg. only) [49]	S		78.9	29.2	75.5	0.1	4.8	72.6	76.7	43.4	8.8	71.1	16.0	3.6	8.4	37.6	17.4
Source only	-	VGG-16	6.4	17.7	29.7	0.0	7.2	30.3	66.8	**51.1**	1.5	47.3	3.9	0.1	0.0	20.2	-
FCNs in the wild (feat. only) [16]	F	VGG-16	11.5	18.3	33.3	0.0	**11.2**	43.6	70.5	45.5	1.3	45.1	4.6	0.1	0.5	22.0	1.8
Cross-city (feat. only) [15]	F	VGG-16	56.5	24.0	78.9	1.1	5.9	**77.8**	77.3	35.8	5.4	61.7	5.2	0.9	8.4	33.8	13.6
Baseline (feat. only) [49]	F	VGG-16	63.1	17.9	76.3	**4.7**	8.4	68.3	79.9	38.7	**8.5**	64.7	9.7	0.6	6.0	34.4	14.2
IBAN (Ours)	F	VGG-16	70.0	19.1	78.7	1.4	4.5	73.1	77.0	42.2	2.6	72.5	14.0	0.8	3.9	35.4	15.2
SIBAN (Ours)	F	VGG-16	**70.1**	**25.7**	**80.9**	3.8	7.2	72.3	**80.5**	43.3	5.0	**73.3**	**16.0**	**1.7**	**3.6**	**37.2**	**17.0**
Source only	-	ResNet	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	-
Baseline (seg. only) [49]	S	ResNet	79.2	37.2	78.8	9.9	10.5	78.2	80.5	53.5	19.6	67.0	29.5	21.6	31.3	45.9	7.3
Source only	F	ResNet	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	-
Baseline (feat. only) [49]	F	ResNet	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	2.2
IBAN (Ours)	F	ResNet	78.2	19.7	**80.5**	9.4	8.9	77.4	82.0	56.3	9.6	76.3	22.8	17.5	23.3	43.2	4.6
SIBAN (Ours)	F	ResNet	**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**	**7.7**

to be near zero. Consequently, our method attempts to optimize the upper bound for $d_{\mathcal{H}}(\mathcal{S},\mathcal{T})$, thus offering a tighter upper bound for $\epsilon_{\mathcal{T}}(h)$. The proof shows that our method is mathematically consistent with the theory of Ben-David *et al* [2]. □

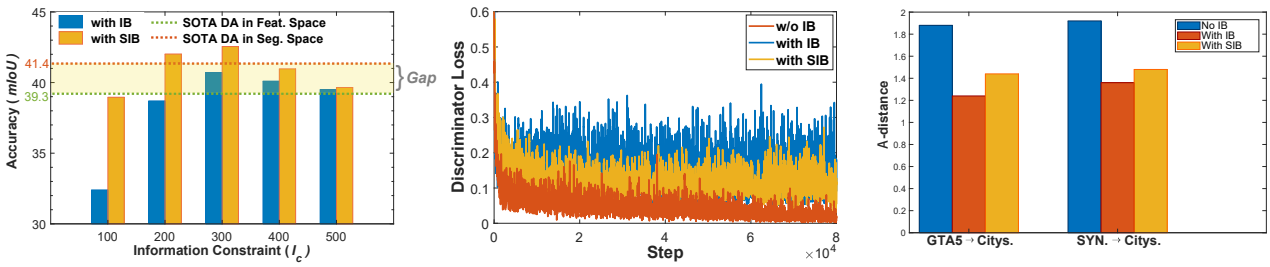
4. Experiments

4.1. Datasets

We evaluate our algorithm together with several state-of-the-art algorithms on two adaptation tasks, *e.g.*, SYNTHIA [38] → Cityscapes [10] and GTA5 [37] → Cityscapes. Cityscapes is a real-world dataset with $5,000$ street scenes which are divided into a training set with $2,975$ images, a validation set with 500 images and a testing set with $1,525$ images. We use Cityscapes as the target domain. GTA5 contains $24,966$ high-resolution images, automatically annotated into 19 classes. The dataset is rendered from a modern computer game, Grand Theft Auto V, whose labels are fully compatible with Cityscapes. SYNTHIA contains $9,400$ synthetic images compatible with the Cityscapes annotated classes. We use SYNTHIA or GTA5 as the source domain in the evaluation.

4.2. Implementation Details

We use PyTorch for our implementation. We utilize 1) DeepLab-v2 [6] framework with ResNet-101 [14] and 2) VGG-16-based FCN8s [26], as the two respective backbones for G. We use the feature-space adversarial DA method proposed in [49] as the baseline network. For significance-



(a) (b) (c)

Figure 4: (a). Adapted segmentation performance in terms of mIoU. (b). The training loss of D, where a complete balanced adversarial process is achieved when the loss converges to around 0.5. (c). $\mathcal{A}$-distance between source and target domain.

aware layer in SIB, we employ a convolution layer with kernel 1×1 and channel number $2,048$, followed by a ReLU and a Sigmoid to produce the channel-wise significance vector. We use the IB proposed in [1] as our bottleneck module. For network D, we adopt a similar structure with [36], which consists of 5 convolution layers with channel numbers $\{64, 128, 256, 512, 1\}$, the kernel 4×4 , and stride of 2. Each convolution layer is followed by a Leaky-ReLU [32] parameterized by 0.2 except the last layer. During training, we use SGD [3] as the optimizer for G with a momentum of 0.9, while using Adam [21] to optimize D with $\beta_1 = 0.9$, $\beta_2 = 0.99$. We set both optimizers a weight decay of $5e$-4. The initial learning rates for SGD and Adam are set to $2.5e$-4 and $1e$-4, respectively. Both learning rate are decayed by a poly policy, where the initial learning rate is multiplied by $(1-\frac{iter}{max\_iter})^{power}$ with $power = 0.9$. We train the network for a total of $100k$ iterations. We use a crop of $512 \times 1,024$ during training, and for evaluation we up-sample the prediction map by a factor of 2 and then evaluate mIoU. In our best model, we set hyper-parameters $\beta^S_{init} = \beta^T_{init} = 1e$-5, $\alpha = 1e$-8, $\lambda = 1e$-3 and $I_c = 300$, respectively.

4.3. Comparative Studies

Compared with SOTA. We present the adaptation results on tasks GTA5 $\rightarrow$ Cityscapes and SYNTHIA $\rightarrow$ Cityscapes in Table 1 and Table 2 respectively, with comparisons to the state-of-the-art feature-space DA methods [16, 15, 49, 8]. We also present the current state-of-the-art pixel-space and segmentation-space DA in the tables. In Table 1, not surprisingly, SIBAN significantly outperforms the source-only segmentation method by $+16.3\%$ on VGG-16 and $+6.0\%$ on ResNet-101 since the source-only segmentation method does not consider the domain shift. Besides, SIBAN outperforms the state-of-the-art feature-space methods, which improves the mIOU by over $+2.5\%$ compared with FCNs [16], AdaptSeg [49], and CyCADA [15]. Compared to the DA methods in the segmentation and pixel space [49, 15], SIBAN can also be on par with them. In some infrequent classes which are prone to suffer from the side effect of information bottleneck, *e.g.,* fence, traffic light, and rider, we can observe that SIBAN can significantly outperform IBAN. The results verify the effectiveness of SIB module to protect the uncommon classes from being eliminated. Similar results can be observed in Table 2. Some qualitative segmentation examples can be viewed in Fig. 5.

Sensitivity to Constraint. We test the DA performance of IBAN / SIBAN in term of mIoU with varying I_c over a range $\{100, 200, 300, 400, 500\}$, where a smaller I_c indicating a more strict information constraint on the latent features. Fig. 4a presents the test results, in which we can see that SIBAN outperforms IBAN in all constraint cases. For SIBAN, the appropriate choice of I_c is between 200 and 400. An I_C with too small value would eliminate too much essential information, while an excessively large I_c would degrade SIBAN to the baseline model since it introduces too much noise. We can also observe that the IBAN is more sensitive to the constraint. When using $I_c = 300$, both IBAN and SIBAN surpass the feature-space baseline significantly and **SIBAN can even outperform the state-of-the-art segmentation-space DA methods [49].** From the result, we can conclude that our proposed SIBAN has bridged the performance gap between feature-space and segmentation-space DA methods [49].

Training Stability. Here we utilize the loss of D ($Loss_D$) as a proxy for the stability of adversarial training. In a stable adversarial course, G would learn to fool D successfully, and $Loss_D$ should converge to around 0.5. Fig. 4b reports $Loss_D$ over the course of training. We can see that $Loss_D$ quickly drops when the network is trained without IB, indicating D overpowers G substantially and learns to differentiate between features of the two domains accurately. We also observe that the introduction of IB / SIB into the adversarial network can significantly constrain the performance of D, thus stabilizing the adversarial training. Besides, we find the standard IB outperforms SIB, which seems contradictory to our standpoint. We ascribe it to the reason that a standard IB eliminates excessive information from features. Although making the training of D more stable, such comparatively less-informative features would also hurt the semantic segmentation task. On the contrary, our proposed SIB module can achieve both good training stability and outstanding segmentation performance.

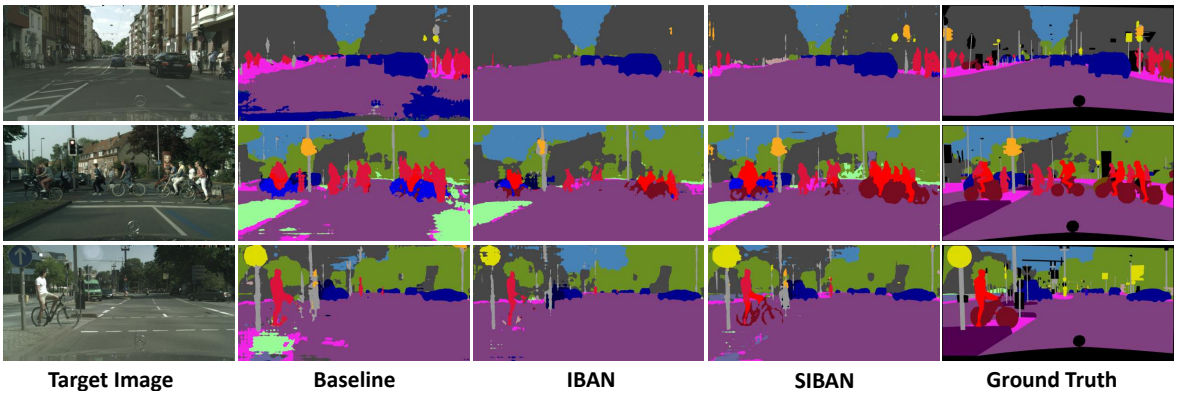


Figure 5: Qualitative results of UDA segmentation for GTA5 → Cityscapes. For each target image, we show the adaption result with baseline model, IBAN and SIBAN respectively, and the ground truth. More results are shown in Appendix.

$\mathcal{A}$-distance. Based on the theory of Ben-David *et al.* [2], $\mathcal{A}$-distance is used as a metric for the domain discrepancy, where a smaller $\mathcal{A}$-distance might indicate better DA performance. Generally, the $\mathcal{A}$-distance is computed as $d_{\mathcal{A}} = 2(1 - 2\epsilon)$, where ϵ is the generalization error of a classifier trained with the binary classification task of discriminating the source and target. In the adversarial training framework, we can just keep D as such a classifier. The comparative results are shown in Fig. 4c. From this figure, we can see that the introduction of IB / SIB significantly reduces $\mathcal{A}$-distances compared to the baseline. However, we can also observe that the $\mathcal{A}$-distance of IBAN is slightly smaller then SIBAN. Consistent with our previous analysis on training stability, we conclude that the discrepancy decrease of IB is at the cost of discarding some necessary information. This finding tells us that only reducing the global distribution discrepancy is far from enough for domain adaptation. The superior DA performance, as well as a relatively small $\mathcal{A}$-distance lead by SIBAN, show that our method can make a better trade-off between the feature purification and domain alignment.

4.4. Ablation Studies

To assess the importance of various aspects of the model, we run experiments on GTA5 → Cityscapes task on ResNet-101 backbone, deactivating one or a few modules at a time while keeping the others activated. Besides, we test the combination performance between SIBAN and other DA methods [30, 18], in which the author suggests the channel-wise significance [18] or the output [30] should also be aligned between domains. We simply implement these two methods by adding two extra discriminators D on significance tensors and segmentation maps, respectively. Table 3 shows the DA results under different settings. We observe that appending SA-layer can significantly improve the standard IB by 1.5%. Updating β_S/β_T adaptively brings extra 0.4% improvement as well. When employing two extra discriminators to the

Table 3: Ablation study on ResNet-101.

GTA5 → Cityscapes				
Module		**Extra D**		**mIoU**
$SA-layer$	$Ada.\beta$	$Sig.$ [18]	$Seg.$ [30]	
				40.7
√				42.2
√	√			42.6
√	√	√		43.2
√	√	√	√	**45.5**

significance tensors and the segmentation maps, the target segmentation accuracy would be further improved by 0.8% and 2.3%. The ablation study verifies the effectiveness of our SIB module as well as our "adaptive β" strategy for DA task. Furthermore, SIBAN can be expediently combined with other DA methods to yield even better segmentation results on target images.

5. Conclusion

In this paper, we propose a novel significance-aware information bottlenecked adversarial network (SIBAN) for domain adaptive semantic segmentation. By conducting a significance-aware feature purification before the adversarial adaptation, SIBAN eases the following feature alignment and stabilizes the adversarial training course, thus significantly improving the feature-space adaptation performance. On two challenging $similated \rightarrow real$ DA tasks, SIBAN yields leading result compared with other feature-space methods, and can even match the state-of-the-art output-space methods in segmentation accuracy. For the semantic segmentation task, our proposed SIBAN brings the feature-/output-space UDA methods to the same starting line.

Acknowledgment. This work is partially supported by the National Natural Science Foundation of China (No. 61572211).

References

[1] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017. 3, 4, 7

[2] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010. 2, 5, 6, 8

[3] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010. 7

[4] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 7, 2017. 3

[5] Matthew Chalk, Olivier Marre, and Gasper Tkacik. Relevant sparse codes with variational information bottleneck. In *Advances in Neural Information Processing Systems*, pages 1957–1965, 2016. 4

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018. 1, 5, 6

[7] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. *arXiv preprint arXiv:1812.05040*, 2018. 3

[8] Yi-Hsin Chen, Wei-Yu Chen, Yu-Ting Chen, Bo-Cheng Tsai, Yu-Chiang Frank Wang, and Min Sun. No more discrimination: Cross city adaptation of road scene segmenters. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2011–2020. IEEE, 2017. 7

[9] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *arXiv preprint arXiv:1711.09020*, 2017. 3

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 2, 6

[11] Qianyu Feng, Guoliang Kang, Hehe Fan, and Yi Yang. Attract or distract: Exploit the margin of open set. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 1

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2, 5

[13] Anirudh Goyal, Riashat Islam, DJ Strouse, Zafarali Ahmed, Hugo Larochelle, Matthew Botvinick, Sergey Levine, and Yoshua Bengio. Transfer and exploration via the information bottleneck. In *ICLR*, 2019. 4

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 6

[15] 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. In *International conference on machine learning*, 2018. 1, 2, 3, 6, 7

[16] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016. 3, 6, 7

[17] Haoshuo Huang, Qixing Huang, and Philipp Krahenbuhl. Domain transfer through deep activation matching. In *ECCV*, 2018. 3

[18] Guoliang Kang, Liang Zheng, Yan Yan, and Yi Yang. Deep adversarial attention alignment for unsupervised domain adaptation: the benefit of target expectation maximization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 401–416, 2018. 8

[19] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 2

[20] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jungkwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, 2017. 3

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[22] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[23] Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. Improving image captioning transformer with entangled attention. In *ICCV*, 2019. 1

[24] Peilun Li, Xiaodan Liang, Daoyuan Jia, and Eric P Xing. Semantic-aware grad-gan for virtual-to-real urban scene adaption. *arXiv preprint arXiv:1801.01726*, 2018. 3

[25] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *Advances in neural information processing systems*, pages 469–477, 2016. 3

[26] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1, 2, 5, 6

[27] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*, 2015. 1

[28] Keyang Luo, Tao Guan, Lili Ju, Haipeng Huang, and Yawei Luo. P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo. In *ICCV*, 2019. 1

[29] Yawei Luo, Tao Guan, Hailong Pan, Yuesong Wang, and Junqing Yu. Accurate localization for mobile device using a multi-planar city model. In *ICPR*, 2016. 1

[30] 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 *CVPR*, 2019. 3, 8

[31] Yawei Luo, Zhedong Zheng, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Macro-micro adversarial network for human parsing. In *European Conference on Computer Vision*, pages 424–440. Springer, 2018. 1, 2

[32] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3, 2013. 7

[33] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *ICCV*, 2019. 1

[34] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *The IEEE International Conference on Computer Vision (ICCV)*, volume 2, page 3, 2017. 1

[35] Xue Bin Peng, Angjoo Kanazawa, Sam Toyer, Pieter Abbeel, and Sergey Levine. Variational discriminator bottleneck: Improving imitation learning, inverse rl, and gans by constraining information flow. In *International Conference on Learning Representations*, 2019. 2, 3, 4

[36] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 7

[37] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European Conference on Computer Vision*, pages 102–118. Springer, 2016. 1, 2, 6

[38] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3234–3243, 2016. 1, 6

[39] Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. Semi-supervised self-training of object detection models. 2005. 3

[40] Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Asymmetric tri-training for unsupervised domain adaptation. *arXiv preprint arXiv:1702.08400*, 2017. 3

[41] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. *arXiv preprint arXiv:1712.02560*, 2017. 1, 3

[42] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Unsupervised domain adaptation for semantic segmentation with gans. *arXiv preprint arXiv:1711.06969*, 2017. 3

[43] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000. 1

[44] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[45] Noam Slonim. *The information bottleneck: Theory and applications*. PhD thesis, Citeseer, 2002. 3

[46] Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):66, 2015. 3

[47] Ben Tan, Yu Zhang, Sinno Jialin Pan, and Qiang Yang. Distant domain transfer learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 2

[48] N TISHBY. The information bottleneck method. In *Proc. 37th Annual Allerton Conference on Communications, Control and Computing, 1999*, pages 368–377, 1999. 2, 3

[49] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. *arXiv preprint arXiv:1802.10349*, 2018. 1, 2, 3, 6, 7

[50] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 4, 2017. 1, 3

[51] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 636–644. IEEE, 2017. 2

[52] Yang Zhang, Philip David, Hassan Foroosh, and Boqing Gong. A curriculum domain adaptation approach to the semantic segmentation of urban scenes. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 1

[53] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identication. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[54] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camstyle: A novel data augmentation method for person re-identification. *IEEE Transactions on Image Processing*, 28(3):1176–1190, 2019. 1

[55] Zhi-Hua Zhou and Ming Li. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on knowledge and Data Engineering*, 17(11):1529–1541, 2005. 3

[56] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017. 3

[57] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 289–305, 2018. 3