Penalizing Top Performers: Conservative Loss for Semantic Segmentation Adaptation

Xinge Zhu[1], Hui Zhou[2], Ceyuan Yang[1], Jianping Shi[2], Dahua Lin[1]

[1]CUHK-SenseTime Joint Lab, CUHK, Hong Kong S.A.R.
[2]SenseTime Research, Beijing, China
zhuxinge123@gmail.com

Abstract. Due to the expensive and time-consuming annotations (e.g., segmentation) for real-world images, recent works in computer vision resort to synthetic data. However, the performance on the real image often drops significantly because of the domain shift between the synthetic data and the real images. In this setting, domain adaptation brings an appealing option. The effective approaches of domain adaptation shape the representations that (1) are discriminative for the main task and (2) have good generalization capability for domain shift. To this end, we propose a novel loss function, i.e., Conservative Loss, which penalizes the extreme good and bad cases while encouraging the moderate examples. More specifically, it enables the network to learn features that are discriminative by gradient descent and are invariant to the change of domains via gradient ascend method. Extensive experiments on synthetic to real segmentation adaptation show our proposed method achieves state of the art results. Ablation studies give more insights into properties of the Conservative Loss. Exploratory experiments and discussion demonstrate that our Conservative Loss has good flexibility rather than restricting an exact form.

1 Introduction

Deep convolutional neural networks have brought impressive advances to the state of the art across a multitude of tasks in computer vision [1,2,3]. At the same time, these significant leaps require a large amount of labeled data. For some pixel-level tasks, e.g., semantic segmentation, obtaining a fine-grained label is expensive and time-consuming. In [4], they report that it takes more than 90 minutes for manually labeling a single image. Recent advances in Computer Graphics [5] offer an alternative solution to address the data issue. In [5], they automatically capture both images and fine-grained labels from GTAV game with the speed faster than human in several orders of magnitude.

However, models trained on the synthetic data fail to perform well on the real-world images. The main reason is the shift between training and test domains [6]. In the presence of the domain shift, the model trained on the synthetic data often tends to be biased towards the source domain (synthetic images), making them incapable to generalize to the target domain (real images).

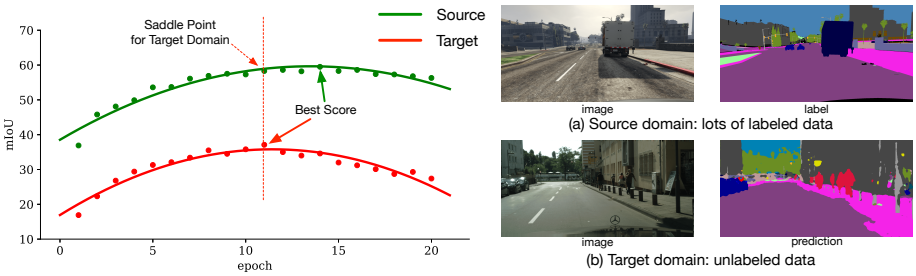


Fig. 1: We show the tendency of mIoU on the source domain and target domain. The curves indicate the trends and points denote the actual mIoU. Besides, we display the samples from source domain (GTAV) and target domain (Cityscapes)

Traditional approaches for domain adaptation mainly focus on the image classification task, which can be summarized as two lines: (1) minimizing the distance between the source and target distributions [7,8,9]; (2) explicitly ensuring that two distributions close to each other by adversarial learning [10,11]. Existing works [12,13] used the similar idea, i.e., gradient reversal layer, to our proposed loss in the domain adaptation for image classification, which was achieved by multiplying a negative scalar during the backpropagation. However, since there exist large category discrepancies between pixels in one image, the manner of uniformly reversing the gradients for all pixels with same scalar is not suitable for the structured prediction in the segmentation. Those drawbacks limit the gradient reversal layer to generalize to the segmentation adaptation.

Semantic segmentation provides pixel-level label for input image, which carries more dense and structured information than image classification, and thus making its domain adaptation difficult. Hence, the domain adaptation techniques in the classification task which focus on sparsely high-level features do not translate well to the segmentation adaptation [14]. Few works have explored the domain adaptation for segmentation [14,15,16]. Orthogonal to those works focusing on manipulating the data statistics [15] or applying the curriculum learning [14] to adaptation, we propose the novel Conservative Loss to realize it without introducing extra computational overhead.

We observe that with training step goes by, the performance on the target domain *first rises and then falls*. We show the trends of mIoU on the experiment of synthetic (GTAV data [5]) to real (Cityscapes data [4]) segmentation adaptation in Fig 1. It can be observed that the performance on source domain and target domain would not reach the best at the same time because of the domain shift. Since there is no ground truth for target domain during training, it is required to find the saddle point of target domain on the source domain. It is note-worthy that the saddle point for target domain does bias to the best score on the source domain but not reach, which delivers a balance between the discriminativeness and domain-invariant. This phenomenon is consistent with many domain adap-

tation theories [17,18,19]. Therefore, we focus on learning representations with two following characteristics which are: (i) discriminative for semantic segmentation on the source domain (corresponding to the '*first rises*') and (ii) invariant to the change of domains.

In this paper, this is achieved by training with the Conservative Loss in an adversarial framework. The Conservative Loss is extremely simple. It holds two attributes corresponding to the properties of desired representations. First, when the probability of ground truth label on the source domain is low, the Conservative Loss enforces the network to learn more discriminative features via gradient descent, which corresponds to the first property of discriminativeness. Second, when the probability of ground truth label is much high, our loss penalizes this case by giving a negative value, which prevents the model from biasing to source domain training data further increasing the generalization capability. This corresponds to the second property of domain-invariant. Our loss function can be seen to seek the optimal parameters that deliver a saddle point of those two objectives. Furthermore, the generative adversarial network (GAN) [20] is also introduced to our model. Unlike some works [10,15] where they apply the feature-level discriminator, we utilize the GAN to further supplement the domain alignment by enforcing reconstructed images to be indistinguishable for the discriminator.

We conduct extensive experiments on synthetic to real segmentation adaptation. The proposed method considerably improves over previous state-of-the-art and achieves **9.3** points of mIoU gain on Synthia [21] to Cityscapes [4] experiment without introducing any extra computational overhead during evaluation. Ablation studies verify the effect of different components to our performance and give more insights into properties of our Conservative Loss. More discussions and visualization demonstrate the Conservative Loss has good flexibility rather than limiting to a fixed instantiation.

2 Related Work

Semantic Segmentation. Semantic segmentation is a highly active field, which is a task of assigning object label to each pixel of image. With the surge of deep segmentation model [3], most recent top-performing methods are built on the CNNs [1,22,23].

Huge amount of human effort is required to annotate the fined-grained semantic segmentation ground truth. According to [5], it did take about 60 minutes to manually segment each image. On the contrary, collecting data from video games such as GTAV [5] is much faster and cheaper compared with the human annotator. For example, [5] extracted 24,966 GTAV images with annotations within 49 hours by using a GPU parallel method. However, it is hard to apply the model trained on the synthetic image to the real-world image because of their discrepant data distributions.

Domain Adaptation. Many machine learning methods rely on the assumption that the training and test data are in the same distribution. However, it is

often the case that there exists some discrepancies [17,19], which leads to significant performance drop on the test data. Domain adaptation aims to alleviate the impact of the discrepancy between training and test data.

Domain Adaptation for Image Classification. Existing works on domain adaptation mostly focus on image classification problem. Conventional methods include Maximum Mean Discrepancy (MMD) [7,8,9], geodesic flow kernel [24], sub-space alignment [25], asymmetric metric learning [26], *etc.* Recently, domain adaptation approaches aim to improve the adaptability of deep neural networks [7,13,27,28,29,30,31,32].

Domain Adaptation for Semantic Segmentation. Much less attention has been given to domain adaptation for semantic segmentation task. The pioneering work in this task is [15], which combines the global and local alignment methods with a domain adversarial training. Another work [14] applies the curriculum learning to solve the domain adaptation from easy to hard. In [16], they propose an unsupervised learning to adapt road scene segmenters across different cities. In [33], they perform output space adaptation at feature level by an adversarial module. Unlike them constraining the distribution [15] or the output of the network [33], we propose the Conservative Loss to naturally seek the discriminative and domain-invariant representations.

Adversarial Learning. Recently, Generative Adversarial Network (GAN) [20] has raised great attention. Some works extend this framework for domain adaptation. CoGAN [11] achieves the domain adaptation by generating cross-domain instances. Domain adversarial neural networks [12] consider adversarial training for suppressing domain biases. In [10], they incorporate adversarial discriminative setting to help mitigate performance degradation. In our work, we also incorporate the GAN into our model, whose discriminator drives the source image towards the target one for promoting domain alignment.

3 Methodology

As presented above, the key to realize unsupervised domain adaptation is the discriminative and domain-invariant representations. The Conservative Loss is proposed to penalize the extreme cases and its goal is to deliver a balance between the discriminative and the domain-invariant representations. Furthermore, we introduce the generative adversarial networks to align the source and target embedding. Below, we first describe the framework of our model and its network blocks. Then, the Conservative Loss and its background are presented in details. Finally, the alternative optimization is provided.

3.1 Framework Overview

Our framework is illustrated in Figure 2. In our setting, there are two domains: source domain (image and label) and target domain (image only). Our framework aims to achieve good performance on the target domain by applying the model trained on the source domain.

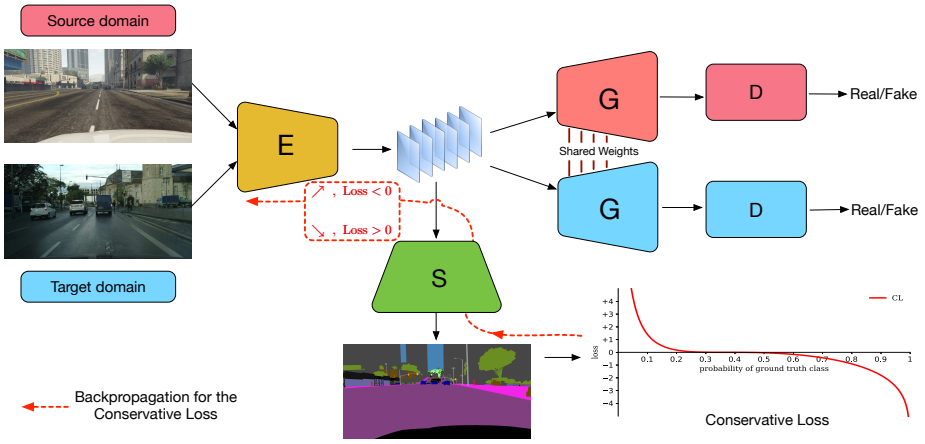


Fig. 2: The pipeline of our framework. E denotes the encoder, G denotes the generator, and D is the discriminator. S is the pixel-wise classifier for semantic segmentation. The red color represents the network blocks for the source domain, and the blue for the target domain. We also display the Conservative Loss and its backpropagation. ↗ represents the gradient ascend and ↘ denotes the gradient descend

Our model consists of two major parts, i.e., GAN and Segmentation part. The GAN aims to align the source and target embedding. More specifically, the generator and discriminator are playing a minimax game [20], in which the generator takes source embedding as input and generates the target-like image to fool the discriminator, while the discriminator tries to classify the reconstructed image [10,11]. The segmentation part can be seen as a regular segmentation model. For each part, the detailed components are shown in the following:

- The encoder(E) performs the feature embedding given source or target image, whose architecture is a fully convolutional network. The generator(G) reconstructs the image based on the embedding. The discriminator(D) does classify the reconstructed images as real or fake. S is the pixel-wise classifier.
- The GAN consists of encoder, generator and discriminator.
- The segmentation part consists of encoder and pixel-level classifier. Note that the encoder does work in both GAN and Segmentation.

The detailed architecture of generators and discriminators is described in the supplementary material because of the limited page space.

3.2 Background

In this section, we briefly introduce the theory of domain adaptation and present its relation to our proposed loss.

Many theoretical analyses of domain adaptation [17,18,19] have offered a upper bound on the expected risks of target domain, which depends on its source domain error (test-time) and the divergence between two domains. Formally,

$$\epsilon_{\mathcal{T}} \leq \epsilon_{\mathcal{S}} + \frac{1}{2} d(\mathcal{S}, \mathcal{T}) + \mathcal{C}, \tag{1}$$

where $\mathcal{S}$ and $\mathcal{T}$ denote the source domain and target domain, respectively. ϵ is the expected risk. d is the domain divergence, which has different notions, for example $\mathcal{H}$-divergence [19]. $\mathcal{C}$ is a constant term.

It can be observed that two terms $\epsilon_{\mathcal{S}}$ and $d(\mathcal{S}, \mathcal{T})$ closely relate to the properties in the desired representations. The first term $\epsilon_{\mathcal{S}}$ indicates that the model should produce discriminative representations for getting smaller expected risks on the source domain, which corresponds to the first property of discriminativeness. The second term $d(\mathcal{S}, \mathcal{T})$ defines the discrepancy distance between two distributions, in which the more similar the representations of both domains are, the smaller it is. This correlates with the second property of domain-invariant. More theoretical analyses are shown in the supplementary material.

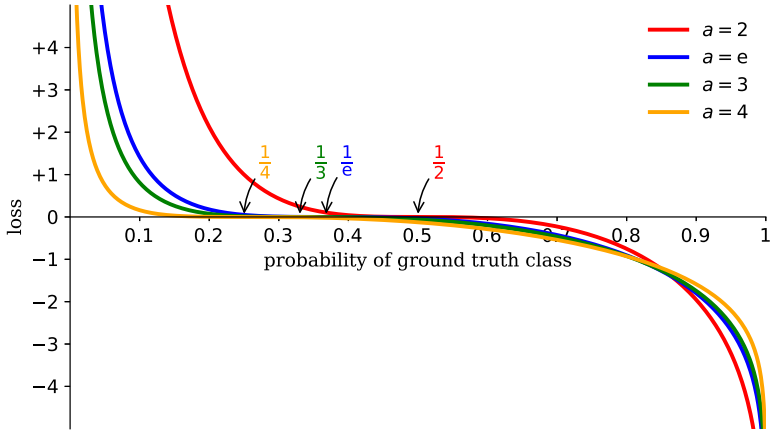


Fig. 3: The proposed Conservative Loss with different a. It can be observed that the Conservative Loss keeps low values in the middle level and punishes the extremely good or bad cases

3.3 Conservative Loss

As explained above, the desired representations should be discriminative for the main task on source domain and possess good generalization ability rather than getting into the overfitting. We thus propose the Conservative Loss for the

semantic segmentation on the source domain, which carries the two following properties:

- When the probability of ground truth class is low, the loss function gives a positive value, which enables the network to learn a more discriminative feature by using gradient descent method.
- When the probability is high, the loss function delivers the negative value, which makes the network avoid the bias towards the source domain via the gradient ascend further learning the better generalization.

The Conservative Loss is formulated as:

$$\mathrm{CL}(p_t) = (1 + \log_a(p_t))^2 * \log_a(-\log_a(p_t)), \tag{2}$$

where p_t is the probability of our prediction towards ground truth. a is the base of logarithmic function, which also indicates the intersection point with x-axis, that is $\frac{1}{a}$. The Conservative Loss is visualized for several values of $a \in [2, \mathrm{e}, 3, 4]$ in Figure 3, in which e is Euler's number and $\mathrm{e} \approx 2.718$. Specifically, $(1+\log_a(p_t))^2$ acts as a modulating factor, which delivers the large values when p_t is much low or high. $\log_a(-\log_a(p_t))$ is designed as the switch of gradient direction, in which when $p_t > \frac{1}{a}$ it is negative, otherwise it is positive.

In the following, we have raised two lemmas to analysis the appealing property of our Conservative Loss.

Lemma 1: *The objective function of domain adaptation system contains a saddle point, which relates to the zero point of Conservative Loss.*

As the pipeline in Fig 2 shown, the full objective consists of two parts, including the loss $\mathcal{L}^s_{seg,p_t}$ for Segmentation and the loss $\mathcal{L}_{GAN}$ for GAN . The sign of $\mathcal{L}^s_{seg,p_t}$ dynamically depends on p_t. When p_t is much high, the negative value leads to the gradient ascend for escaping the bias to source domain. Otherwise, the positive value makes the features discriminative. It can be seen that our loss balances the two objectives (discriminativeness and domain-invariant) that shape the representations during learning, and its zero point acts as the saddle point. More details are shown in the supplementary material.

Lemma 2: *Our loss encourages the moderate examples in large range, which makes the overall optimization more stable.*

From the loss form, it can be observed that the loss focuses on the hard negatives and positives, and tends to give the low value for the probability in the middle level. For instance, with $a = \mathrm{e}$, the loss values of $p_t = 0.9$ and $p_t = 0.1$ are -1.8 and 1.4, respectively, while the loss values of $p_t = 0.5$ and $p_t = 0.6$ are -0.03 and -0.06. In such setting, the loss extends the range in which an example receives low loss, which brings a stable optimization even in the case of the gradient descend and ascend frequently alternate due to the joint optimization of $\mathcal{L}^s_{seg,p_t}$ and $\mathcal{L}_{GAN}$.

In practice we use a λ-balanced variant of the Conservative Loss:

$$\mathrm{CL}(p_t) = \lambda(1 + \log_a(p_t))^2 * \log_a(-\log_a(p_t)). \tag{3}$$

As our experiments will show, different balanced factors λ yield slightly different performance. While in our main experiments we use the Conservative Loss defined above, its exact form is not crucial. In Section 4.5 we offer other forms of our loss which also maintain the two properties, and experimental results demonstrate that they can also be effective.

3.4 Model Objective

Our full objective is to alternatively update the three network blocks, i.e., discriminators(D), generators(G) and encoder(E). Note that S is a pixel-level classifier which has no learnable parameters in our model. Hence, the objective contains three terms: $\mathcal{L}_D$, $\mathcal{L}_G$ and $\mathcal{L}_E$. We then explain the various losses used in our method and describe the alternative optimization scheme.

Adversarial Loss. Inheriting from GAN [20], we apply the adversarial losses which are derived from the discriminator to all three blocks. We term them as $\mathcal{L}_{\text{GAN,D}}, \mathcal{L}_{\text{GAN,G}}$ and $\mathcal{L}_{\text{GAN,E}}$. For each adversarial loss it consists of two parts, i.e., $\mathcal{L}^s_{GAN}$ for the source image and $\mathcal{L}^t_{GAN}$ for the target image. Thus we can obtain the adversarial loss by $\mathcal{L}_{GAN} = \mathcal{L}^s_{GAN} + \mathcal{L}^t_{GAN}$. It is noted that for the encoder, the adversarial loss does a cross-domain update (i.e., classifying the image as real or fake from source domain to target domain and vice versa), which enforces the network to generate similar embeddings for two domains.

Reconstructed Loss. The generator performs the image reconstruction. We use L1 distance as $\mathcal{L}_{rec}$ because L1 encourages less blurring.

Segmentation Loss. As Section 3.3 introduced, the Conservative Loss is applied to the semantic segmentation in the domain adaptation setting.

During training, we iteratively optimize all three learnable parts (Encoder, Generator and Discriminator). During inference, only the encoder and segmentation classifier are used to produce the results on target domain. The alternating update scheme is described as following:

(1) Update discriminators: the overall loss is $\mathcal{L}_D = \mathcal{L}_{GAN,D}$.
(2) Update generators: the loss involves the adversarial loss and reconstructed loss. The overall loss is $\mathcal{L}_G = \mathcal{L}_{GAN,G} + \mathcal{L}_{rec}$.
(3) Update encoder: since the encoder does work in both two components, i.e., GAN and Segmentation, the overall loss is a combination of several losses, including adversarial loss and segmentation loss on source domain; $\mathcal{L}_E = \mathcal{L}_{GAN,E} + \mathcal{L}^s_{seg}$.

4 Experiments

4.1 Dataset

Following previous works [14,15], we use GTAV [5] or Synthia [21] dataset as the source domain with pixel-level labels, and we use Cityscapes [4] dataset as the target domain. We briefly introduce the datasets as following:

GTAV has 24,966 urban scene images rendered by the gaming engine GTAV. The semantic categories are compatible with the Cityscapes dataset. We take the whole GTAV dataset with labels as the source domain data.
Synthia is a large dataset which contains different video sequences rendered from a virtual city. We take SYNTHIA-RAND-CITYSCAPES [21] as the source domain data which provides 9,400 images from all the sequences with Cityscape-compatible annotations. Inheriting from existing methods [14], we take 16 common object categories for the evaluation.
Cityscapes is a real-world image dataset focused on the urban scene, which consists of 2,975 images in training set and 500 images for validation. The resolution of images is 2048 $\times$ 1024 and 19 semantic categories are provided with pixel-level labels. We take the **unlabeled training set** as the target domain data. The adaptation results are reported on the validation set.

4.2 Training Setup

In our experiments, we use the FCN8s [34] as the semantic segmentation model. The backbone is VGG16 [2] which is pretrained on the ImageNet dataset [35]. We apply the PatchGAN [36] as the discriminator, in which the discriminator tries to classify whether overlapping image patches are real or fake. Similar to EBGAN [37], we add the Gaussian noise to the generator. During training, Adam [38] optimization is applied with β_1=0.9 and β_2=0.999. For the Conservative Loss, we apply $a = \mathrm{e}$ and the balanced weight $\lambda = 5$. The ablation study will give more detailed explanations. Due to the GPU memory limitation, the images used in our experiments are resized and cropped to 1024×512 and the batch size is 1. More experimental settings will be available in the supplementary material.
Warm Start. In our experiments, two different training strategies are employed, which are cold start and warm start. The cold start is that the whole model is trained by using the Conservative Loss from scratch. The warm start indicates the model is trained by first using cross entropy loss and then using our Conservative Loss. Many works [39,40,41] demonstrate that the warm start strategy to gradient update provides a more stable training compared with cold start. As the ablation study will show, the warm start performs better than the cold start. In the next domain adaptation experiments, the model is trained using warm start strategy for fairness.

4.3 Results

In this section, we provide a quantitative evaluation by performing two adaptation experiments, i.e., from GTAV to Cityscapes and from Synthia to Cityscapes. We compare our method with several existing models, including FCNWild [15], CDA [14] and [33]. FCNWild [15] applies the dilated network [42] as the backbone and the base model of [14] is the FCN8s-VGG19 [34]. Tsai *et al.* [33] adopts adversarial learning in the output space to perform feature adaptation. The detailed results of each category are available in the supplementary material.

Table 1: Results of domain adaptation from GTAV → Cityscapes. The bold values denote the best scores in the column.

Methods	Base	mIoU	mIoU gain
NoAdapt [15]	DilatedNet [42]	21.1	
FCNWild [15]	DilatedNet [42]	27.1	6.0
NoAdapt [14]	FCN8s [3]	22.3	
CDA [14]	FCN8s [3]	28.9	6.6
Tsai *et al.* [33]	FCN8s [3]	35.0	–
Ours-NoAdapt	FCN8s [3]	30.0	
Ours	FCN8s [3]	**38.1**	**8.1**

Table 2: Results of domain adaptation from Synthia → Cityscapes.

Methods	Base	mIoU	mIoU gain	mIoU-2
NoAdapt [15]	DilatedNet [42]	17.4		
FCNWild [15]	DilatedNet [42]	20.2	2.8	
NoAdapt [14]	FCN8s [3]	22.0		
CDA [14]	FCN8s [3]	29.0	7.0	
Tsai *et al.* [33]	FCN8s [3]	–	–	37.6
Ours-NoAdapt	FCN8s [3]	24.9		
Ours	FCN8s [3]	**34.2**	**9.3**	**40.3**

GTAV → Cityscapes. For a fairness, the result is evaluated over the 19 common classes. From Table 1 shown, our proposed method achieves the best performance (mIoU=**38.1**), which has **9.2** points higher than [14] and **11** points higher than [15]. Due to the different experimental settings and backbone network (baseline method [14] also mentions the difference), our own baseline performance is higher than other methods. However, the highlight is the **performance gain**. We can find that the proposed method yields an improvement of **8.1** points higher than 6.0 in [15] and 6.6 in [14].

Synthia → Cityscapes. We report the results of mIoU in Table 2. It is noted that [33] reported the results on Synthia [21] to Cityscapes adaptation with only 13 object categories (excluding wall, fense and pole). We also report this results as the mIoU-2. Our proposed model achieves a mIoU of **34.2**, and more importantly our model obtains a **9.3** points of performance gain which is higher than the performance gain of [14] (7.0) and [15] (2.8). Compared with [33] on 13 categories, our method also achieves the better performance. In particular, our model does not use any additional scene parsing data except the source domain and target domain data, while the [14] uses another dataset, i.e., PASCAL CONTEXT dataset, to obtain the superpixel label.

4.4 Ablation Study

In this section, we perform the thorough ablation experiments, including experiments with different components, different factors in the Conservative Loss and

Table 3: Results of ablation study for different components in the proposed model. CL means the Conservative Loss. CE means the cross entropy loss

Model	FCN8s+CE	FCN8s+GAN+CE	FCN8s+GAN+CL
mIoU	30.0	34.4	38.1

Table 4: Results of ablation experiments for a and λ in the Conservative Loss

a (with fixed $\lambda = 5$)	2	e	3	4
mIoU	37.5	38.1	37.3	36.8
λ (with fixed $a = \mathrm{e}$)	1	5	10	20
mIoU	37.2	38.1	37.9	37.8

different training strategies. Those experiments demonstrate different contributions of components and provide more insights of our method.

Effect of different components. In this experiment, we show how each component in our model affects the final performance. We consider several cases as following: (1): the baseline model, which contains only the base segmentation model (FCN8s in our model) and is trained using source data only. (2) the FCN8s and GAN component, which consists of base model and GAN and is trained using both source data and target data with the cross entropy loss. (3) the full model, which involves three parts, i.e., base model, GAN and Conservative Loss. We perform the ablation experiments on GTAV→Cityscapes setting.

The results of ablation study are shown in Table 3. It can be observed that each component plays an important role in performance improvement. More specifically, our full model achieves the best results and obtains **8.1** points performance gain. The GAN part also gets 4.4 performance gain compared with FCN8s+CE. Note that the GAN component could introduce the unlabeled target domain data into the whole model, so the Conservative Loss is applied based on the GAN and there is no variant of FCN8s+CL.

Effect of a and λ in the Conservative Loss. In this part, we design the ablation experiments for a and λ in the Conservative Loss. As shown in Equation 2, a is the base of logarithm and denotes the intersection point with x-axis. λ is a balanced factor. We show the impacts of different a and λ in Table 4.

Since there are two variables, we perform the ablation study for one variable with another fixed. For the ablation of a (with fixed $\lambda = 5$), it can be observed that $a = \mathrm{e}$ achieves the best result. Furthermore, we can find that all different a obtain much better performance compared with the cross entropy loss (34.4 in Table 3), which demonstrates that our loss performs consistently better and has high robustness. For the ablation of λ (with fixed $a = \mathrm{e}$), different λ show slightly different results and $\lambda = 5$ obtains the best performance.

Warm start & Cold start. As described in Section 4.2, we use a warm start strategy to train the proposed model. In this experiment, we compare the two training strategies. For the cold start strategy, we clamp the Conservative Loss

with [min $= -10$, max $= 10$], while this constraint is not exist in the warm start. We use the λ-balanced Conservative Loss with $\lambda = 5$ and $a = \mathrm{e}$.

Table 5: Results of two training strategies, i.e., cold start and warm start. CL means the Conservative Loss

Loss Function	[14]	CL with cold start	CL with warm start
mIoU	28.9	35.2	38.1

In Table 5, it can be observed that the Conservative Loss with cold start outperforms [14] with a large margin (6.3 points). The warm start performs better than the cold start because it enables the network to train stably.

4.5 Discussion

In this section, we design several experiments to verify the capability of the proposed method. We show the effect of adaptation on distribution to measure how domain gap is reduced in the feature level. We compared with several classification losses and homogeneous losses to show its superiority and flexibility.
Visualizations. To verify the effect of adaptation on the distribution, we use t-

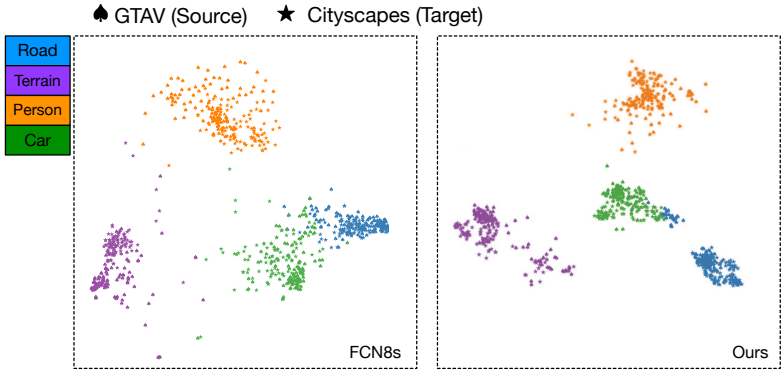
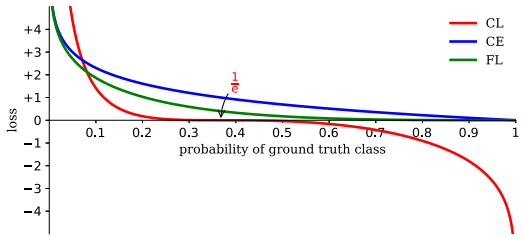


Fig. 4: We show the effect of adaptation on the distribution of the extracted features. ♠ denotes the point from source domain and ★ is from target domain

SNE [43] to visualize feature distributions in Figure 4. 100 images are randomly selected from each domain and for each image the features from last convolutional layer (the channel size equals to class categories.) are extracted. We compare the distributions of our model with FCN8s (No adaptation). Four categories are sampled to display for a clearly visual effect. We observe that with the

adaptation applying, the distance between two domains with same class becomes closer and the discrepancy between different classes also gets clear.



Loss	mIoU
Cross Entropy	34.4
Focal Loss	35.8
Conservative Loss	38.1

Fig. 5: The left figure shows three classification losses, including Cross Entropy loss (CE) in blue, Focal Loss (FL) in green and Conservative Loss (CL) in red. The right table shows the results of all three losses on GTAV → Cityscapes adaptation experiment

Comparison with other classification losses. In this experiment, we compare the Conservative Loss to Cross Entropy Loss and Focal Loss [44]. The Cross Entropy Loss is given by $\mathrm{CE}(p_t) = -\log(p_t)$, which is plotted in Figure 5 with green line. To ensure fairness, we utilize the α-balanced Focal Loss $\mathrm{FL}(p_t) = -\alpha_t(1-p_t)^2\log(p_t)$ and warm start in the experiment of Focal Loss, and apply $\alpha_t = 5$ by using a cross-validation.

From the right table in Figure 5, it can be observed that the Focal Loss obtains a better performance compared with the cross entropy loss because it focuses learning on hard negative examples. However, in the domain adaptation, the domain-invariant representations are crucial to achieve good adaptation performance. The Conservative Loss does enable the network to be insensitive to domain changes by punishing the extreme cases. It can be seen that the Conservative Loss yields higher result (**38.1**), and obtains more performance gain (**3.7**) than the Focal Loss (1.4) based on the cross entropy loss.

Effect of homogeneous losses. As shown in Section 3.3, the Conservative Loss has two properties: (1) when the p_t is low, the Conservative Loss enforces the network to learn discriminative features. (2) when the p_t is high, the loss enables the network to learn domain invariant features by gradient ascend method, which aims to penalize the extremely good cases. There are several losses that also maintain these two properties, for example the cubic equation. In this experiment, we propose several homogeneous losses to verify the effect of these two

properties, which are given by:

$$\text{Loss}_1 = -\lambda_1(p_t - 0.5)^3, \tag{4}$$

$$\text{Loss}_2 = -\lambda_2(p_t - \frac{1}{\text{e}})^3, \tag{5}$$

$$\text{Loss}_3 = \begin{cases} -\alpha * (p_t - \frac{1}{\text{e}})^3, & p_t < \frac{1}{\text{e}}, \\ -\beta * (p_t - \frac{1}{\text{e}})^3, & p_t \geq \frac{1}{\text{e}}. \end{cases} \tag{6}$$

Equation 4 and 5 demonstrate the λ-balanced cubic equations with different intersection points, i.e., 0.5 and $\frac{1}{\text{e}}$, respectively. Equation 6 is a piecewise function, which is more similar to the Conservative Loss due to these two balanced factors.

Table 6: Results of homogeneous losses

Loss Function	CE	FL	Loss_1	Loss_2	Loss_3	CL
mIoU	34.4	35.8	36.5	36.7	37.8	38.1

We apply the adaptation experiment on GTAV $\rightarrow$ Cityscapes to verify their capabilities. The results are reported in Table 6. In order to ensure fairness, all experiments are performed based on the warm start and those hyper-parameters $(\lambda_1, \lambda_2, \alpha, \beta)$ are chosen by using the cross-validation. We can observe that all homogeneous losses perform better than the cross entropy loss (34.4) and Focal Loss (35.8). Therefore, we can find that the exact form of the Conservative Loss is not crucial, and several homogeneous losses also yield comparable results and perform better than cross entropy loss and Focal Loss. Generally, we expect any loss function with similar properties as Conservative Loss to be equally effective.

5 Conclusion

In this paper, we have proposed a novel loss, the Conservative Loss, for the semantic segmentation adaptation. To enforce the network to learn the discriminative and domain-invariant representations, our loss combines the gradient descend and gradient ascend method together, in which it penalizes the extreme cases and encourages moderate cases. We further introduce the adversarial networks to our full model for supplementing the domain alignment. Extensive experiments demonstrate our model achieves state-of-the-art. Exploratory experiments show that the Conservative Loss has high flexibility without limiting to exact form.

Acknowledgments This work is partially supported by the Big Data Collaboration Research grant from SenseTime Group (CUHK Agreement No. TS1610626).

References

1. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR. (2017) 2881–2890 1, 3
2. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014) 1, 9
3. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. (2015) 3431–3440 1, 3, 10
4. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR. (2016) 3213–3223 1, 2, 3, 8
5. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: ECCV, Springer (2016) 102–118 1, 2, 3, 8
6. Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N.D.: Dataset shift in machine learning. The MIT Press (2009) 1
7. Geng, B., Tao, D., Xu, C.: Daml: Domain adaptation metric learning. IEEE Transactions on Image Processing **20**(10) (2011) 2980–2989 2, 4
8. Gong, B., Shi, Y., Sha, F., Grauman, K.: Geodesic flow kernel for unsupervised domain adaptation. In: CVPR, IEEE (2012) 2066–2073 2, 4
9. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: ICML. (2015) 97–105 2, 4
10. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: CVPR. Volume 1. (2017) 4 2, 3, 4, 5
11. Liu, M.Y., Tuzel, O.: Coupled generative adversarial networks. In: NIPS. (2016) 469–477 2, 4, 5
12. Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M.: Domain-adversarial neural networks. arXiv preprint arXiv:1412.4446 (2014) 2, 4
13. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: ICML. (2015) 1180–1189 2, 4
14. Zhang, Y., David, P., Gong, B.: Curriculum domain adaptation for semantic segmentation of urban scenes. In: ICCV. Volume 2. (2017) 6 2, 4, 8, 9, 10, 12
15. Hoffman, J., Wang, D., Yu, F., Darrell, T.: Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. arXiv preprint arXiv:1612.02649 (2016) 2, 3, 4, 8, 9, 10
16. Chen, Y.H., Chen, W.Y., Chen, Y.T., Tsai, B.C., Wang, Y.C.F., Sun, M.: No more discrimination: Cross city adaptation of road scene segmenters. In: ICCV, IEEE (2017) 2011–2020 2, 4
17. Mansour, Y., Mohri, M., Rostamizadeh, A.: Domain adaptation: Learning bounds and algorithms. arXiv preprint arXiv:0902.3430 (2009) 3, 4, 6
18. Ben-David, S., Blitzer, J., Crammer, K., Pereira, F.: Analysis of representations for domain adaptation. In: NIPS. (2007) 137–144 3, 6
19. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. Machine learning **79**(1-2) (2010) 151–175 3, 4, 6
20. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS. (2014) 2672–2680 3, 4, 5, 8
21. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: CVPR. (2016) 3234–3243 3, 8, 9, 10

22. Wu, Z., Shen, C., Hengel, A.v.d.: Wider or deeper: Revisiting the resnet model for visual recognition. arXiv preprint arXiv:1611.10080 (2016) 3
23. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: ICCV. (2015) 1520–1528 3
24. Gong, B., Shi, Y., Sha, F., Grauman, K.: Geodesic flow kernel for unsupervised domain adaptation. In: CVPR, IEEE (2012) 2066–2073 4
25. Fernando, B., Habrard, A., Sebban, M., Tuytelaars, T.: Unsupervised visual domain adaptation using subspace alignment. In: ICCV. (2013) 2960–2967 4
26. Kulis, B., Saenko, K., Darrell, T.: What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In: CVPR. (2011) 1785–1792 4
27. Carlucci, F.M., Porzi, L., Caputo, B., Ricci, E., Bulò, S.R.: Autodial: Automatic domain alignment layers. In: ICCV. (2017) 4
28. Lu, H., Zhang, L., Cao, Z., Wei, W., Xian, K., Shen, C., van den Hengel, A.: When unsupervised domain adaptation meets tensor representations. In: ICCV. Volume 2. (2017) 4
29. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Deeper, broader and artier domain generalization. In: ICCV, IEEE (2017) 5543–5551 4
30. Ghifary, M., Kleijn, W.B., Zhang, M., Balduzzi, D., Li, W.: Deep reconstruction-classification networks for unsupervised domain adaptation. In: ECCV, Springer (2016) 597–613 4
31. Busto, P.P., Gall, J.: Open set domain adaptation. In: ICCV. Volume 1. (2017) 4
32. Motiian, S., Piccirilli, M., Adjeroh, D.A., Doretto, G.: Unified deep supervised domain adaptation and generalization. In: ICCV. Volume 2. (2017) 4
33. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: CVPR. (2018) 4, 9, 10
34. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. (2015) 3431–3440 9
35. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR, IEEE (2009) 248–255 9
36. Li, C., Wand, M.: Precomputed real-time texture synthesis with markovian generative adversarial networks. In: ECCV, Springer (2016) 702–716 9
37. Zhao, J., Mathieu, M., LeCun, Y.: Energy-based generative adversarial network. arXiv preprint arXiv:1609.03126 (2016) 9
38. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 9
39. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016) 9
40. Tirumala, S.S., Ali, S., Ramesh, C.P.: Evolving deep neural networks: A new prospect. In: Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), 2016 12th International Conference on, IEEE (2016) 69–74 9
41. Zinkevich, M.: Theoretical analysis of a warm start technique. In: NIPS 2011 Workshop on BigLearn, Citeseer (2011) 9
42. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015) 9, 10
43. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(Nov) (2008) 2579–2605 12
44. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. ICCV (2017) 13