An Adversarial Perturbation Oriented Domain Adaptation Approach for Semantic Segmentation

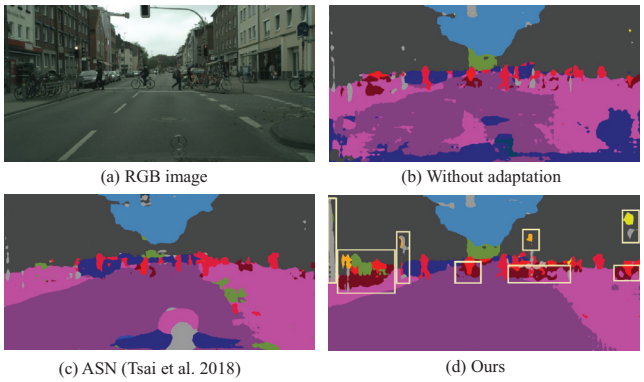
Jihan Yang,[1,2] **Ruijia Xu,**[1] **Ruiyu Li,**[2] **Xiaojuan Qi,**[3] **Xiaoyong Shen,**[2] **Guanbin Li,**[1*] **Liang Lin**[1,4]
[1]School of Data and Computer Science, Sun Yat-sen University, China
[2]Tencent YouTu Lab, [3]University of Oxford
[4]DarkMatter AI Research
{jihanyang13, ruijiaxu.cs}@gmail.com, {royryli, dylanshen}@tencent.com, xiaojuan.qi@eng.ox.ac.uk,
liguanbin@mail.sysu.edu.cn, linliang@ieee.org

Abstract

We focus on Unsupervised Domain Adaptation (*UDA*) for the task of semantic segmentation. Recently, adversarial alignment has been widely adopted to match the marginal distribution of feature representations across two domains *globally*. However, this strategy fails in adapting the representations of the tail classes or small objects for semantic segmentation since the alignment objective is dominated by head categories or large objects. In contrast to adversarial alignment, we propose to explicitly train a domain-invariant classifier by generating and defensing against *pointwise* feature space adversarial perturbations. Specifically, we firstly perturb the intermediate feature maps with several attack objectives (i.e., discriminator and classifier) on each individual position for both domains, and then the classifier is trained to be invariant to the perturbations. By perturbing each position individually, our model treats each location evenly regardless of the category or object size and thus circumvents the aforementioned issue. Moreover, the domain gap in feature space is reduced by extrapolating source and target perturbed features towards each other with attack on the domain discriminator. Our approach achieves the state-of-the-art performance on two challenging domain adaptation tasks for semantic segmentation: GTA5 $\rightarrow$ Cityscapes and SYNTHIA $\rightarrow$ Cityscapes.

Introduction

Semantic segmentation is a fundamental problem in computer vision with many applications in robotics, autonomous driving, medical diagnosis, image editing, etc. The goal is to assign each pixel with a semantic category. Recently, this field has gained remarkable progress via training deep convolutional neural networks (CNNs) (Long, Shelhamer, and Darrell 2015) on large scale human annotated datasets (Cordts et al. 2016). However, models trained on specific datasets may not generalize well to novel scenes (see Figure 1(b)) due to the inevitable visual domain gap between training and testing datasets. This seriously limits the applicability of the model in diversified real-world scenarios. For instance, an autonomous vehicle might not be able to sense its surroundings in a new city or a changing weather condition. To this end, learning domain-invariant representations for semantic segmentation has drawn increasing attentions.



(a) RGB image (b) Without adaptation (c) ASN (Tsai et al. 2018) (d) Ours

Figure 1: Comparison of semantic segmentation output. This example shows our method can evenly capture information of different categories, while classical adversarial alignment method such as ASN (Tsai et al. 2018) might collapse into head (i.e., background) classes or large objects.

Towards the above goal, Unsupervised Domain Adaptation (**UDA**) has shown promising results (Vu et al. 2019; Luo et al. 2019). UDA aims to close the gap between the annotated source domain and unlabeled target domain by learning domain-invariant while task-discriminative representations. Recently, adversarial alignment has been recognized as an effective way to obtain such representations (Hoffman et al. 2016; 2018). Typically, in adversarial alignment, a discriminator is trained to distinguish features or images from different domains, while the deep learner tries to generate features to confuse the discriminator. Recent representative approach ASN (Tsai et al. 2018) is proposed to match the source and target domains in the output space and has achieved promising results.

*Corresponding author is Guanbin Li. This work was supported in part by the State Key Development Program under Grant No.2016YFB1001004, in part by the National Natural Science Foundation of China under Grant No.61976250 and No.U1811463, in part by the Fundamental Research Funds for the Central Universities under Grant No.18lgpy63. This work was also supported by SenseTime Research Fund.

However, adversarial alignment based approaches can be easily overwhelmed by dominant categories (i.e., background classes or large objects). Since the discriminator is only trained to distinguish the two domains globally, it can not produce category-level or object-level supervisory signal for adaptation. Thus, the generator is not enforced to evenly capture category-specific or object-specific information and fails to adapt representations for the tail categories. We term this phenomenon as *category-conditional shift* and highlight it in the Figure 1. ASN performs well in adapting head categories (e.g., road) and gains improvement when viewed globally, but fails to segment the tail categories such as "sign", "bike". Missing the small instances (e.g., traffic light) is generally intolerable in real-world applications. While we can moderate this issue by equipping the segmentation objective with some heuristic re-weighting schemes (Berman, Rannen Triki, and Blaschko 2018), those solutions usually rely on implicit assumptions about the model or the data (e.g., L-Lipschitz condition, overlapping support (Wu et al. 2019)), which are not necessarily met in real-world scenarios. In our case, we empirically show that the adaptability achieved by those approximate strategies are sub-optimal.

In this paper, we propose to perform domain adaptation via feature space adversarial perturbation inspired by (Goodfellow, Shlens, and Szegedy 2014). Our approach mitigates the category-conditional shift by iteratively generating pointwise adversarial perturbations and then defensing against them for both the source and target domains. Specifically, we firstly perturb the feature representations for both source and target samples by appending gradient perturbations to their original features. The perturbations are derived with adversarial attacks on the discriminator to assist in filling in the representation gap between source and target, as well as the classifier to capture the vulnerability of the model. This procedure is facilitated with the proposed Iterative Fast Gradient Sign Preposed Method (I-FGSPM) to mitigate the huge gradient gap among multiple attack objectives. Taking the original and perturbed features as inputs, the classifier is further trained to be domain-invariant by defensing against the adversarial perturbations, which is guided by the source domain segmentation supervision and the target domain consistency constraint.

Instead of aligning representations across domains globally, our perturbation based strategy is conducted on each individual position of the feature maps, and thus can capture the information of different categories evenly and alleviate the aforementioned category-conditional shift issue. In addition, the adversarial features also capture the vulnerability of the classifier, thus the adaptability and capability of the model in handling hard examples (typically tail classes or small objects) is further improved by defensing against the perturbations. Furthermore, since we extrapolate the source adversarial features towards the target representations to fill in the domain gap, our classifier can be aware of the target features as well as receiving source segmentation supervision, which further promotes our classifier to be domain-invariant. Extensive experiments on GTA5 $\rightarrow$ Cityscapes and SYNTHIA $\rightarrow$ Cityscapes have verified the state-of-the-art performance of our method.

Related Work

Semantic Segmentation is a highly active and important research area in visual tasks. Recent fully convolutional network based methods (Chen et al. 2017a; Zhao et al. 2017) have achieved remarkable progress in this field by training deep convolutional neural networks on numerous pixel-wise annotated images. However, building such large-scale datasets with dense annotations takes expensive human labor. An alternative approach is to train model on synthetic data (e.g., GTA5 (Richter et al. 2016), SYNTHIA (Ros et al. 2016)) and transfer to real-world data. Unfortunately, even a subtle departure from the training regime can cause catastrophic model degradation when generalized into new environments. The reason lies in the different data distributions between source and target domains, known as domain shift.

Unsupervised Domain Adaptation approaches have achieved remarkable success in addressing aforementioned problem. Existing methods mainly focus on minimizing the statistic distance such as Maximum Mean Discrepancy (MMD) of two domains (Long et al. 2015; 2017). Recently, inspired by GAN (Goodfellow et al. 2014), adversarial learning is successfully explored to entangle feature distributions from different domains (Ganin and Lempitsky 2015; Ganin et al. 2016). Hoffman et al. (2016) applied feature-level adversarial alignment method in UDA for semantic segmentation. Several following works improved this framework for pixel-wise domain adaption (Chen et al. 2017b; Chen, Li, and Van Gool 2018). Besides alignment in the bottom feature layers, Tsai et al. (2018) found that output space adaptation via adversarial alignment might be more effective. Vu et al. (2019) further proposed to align output space entropy maps. On par with feature-level and output space alignment methods, the remarkable progress of unpaired image to image translation (Zhu et al. 2017) inspired several methods to address pixel-level adaptation problems (Hoffman et al. 2018; Zhang et al. 2018). Among some other approaches, Zou et al. (2018) used self-training strategy to generate pseudo labels for unlabeled target domain. Saito, Ushiku, and Harada (2017) utilized tri-training to assign pseudo labels and obtain target-discriminative representations, while Luo et al. (2019) proposed to compose tri-training and adversarial alignment strategies to enforce category-level feature alignment. Recently, Xu et al. (2019) reveals that progressively adapting the task-specific feature norms of the source and target domains to a large range of values can result in significant transfer gains.

Adversarial Training injects perturbed examples into training data to increase robustness. These perturbed examples are designed for fooling machine learning models. To the best of our knowledge, adversarial training strategy is originated in (Szegedy et al. 2013) and further studied by Goodfellow, Shlens, and Szegedy (2014). Several attack methods are further designed for efficiently generating adversarial examples (Kurakin, Goodfellow, and Bengio 2016; Dong et al. 2018). As for UDA, Volpi et al. (2018) generated adversarial examples to adaptively augment the dataset. Liu et al. (2019) produced transferable examples to fill in the domain gap and adapt classification decision boundary. However, the above approach is only validated on the clas-

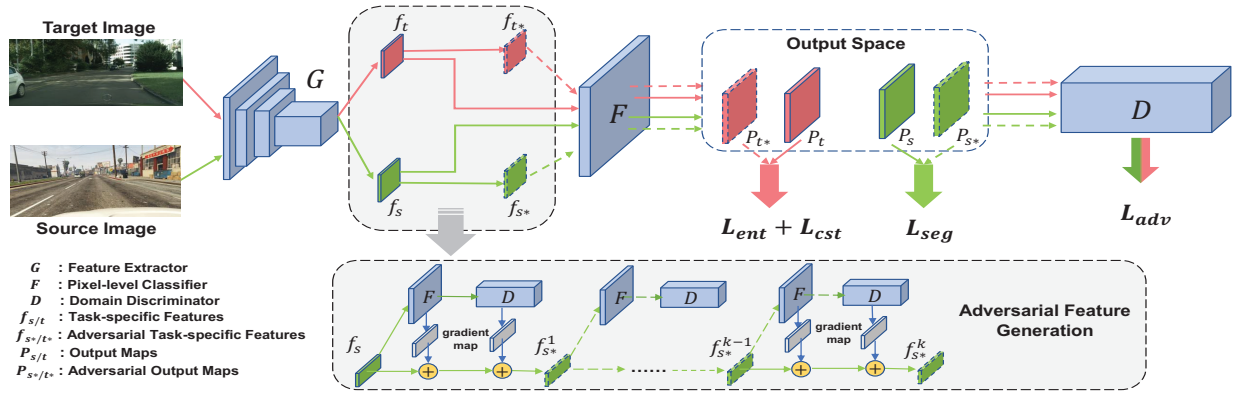


Figure 2: Framework Overview. We illustrate step 2 in the shaded area where source features are taken as an example. In light of $f_{s/t}$ extracted from the feature extractor G, we employ the multi-objective adversarial attack with our proposed I-FGSPM on the classifier F as well as discriminator D and then accumulate the gradient maps. Therefore, we obtain the mutated features $f_{s*/t*}$ after appending the perturbations to the original copies. Furthermore, these perturbed and original features are trained by an adversarial training procedure (i.e., step 3), which is presented in the upper right. We have highlighted the different training objectives for the output maps of their corresponding domains, which are predicted by the classifier F and then followed by the discriminator D to produce domain prediction maps. The green and red colors stand for the source and target flows respectively.

sification task for unsupervised domain adaptation. Our approach shares similar spirit with Liu et al., while we investigate adversarial training in the field of semantic segmentation to generate pointwise perturbations that improve the robustness and domain invariance of the learners.

Method

Considering the problem of unsupervised domain adaptation in semantic segmentation. Formally, we are given a source domain $\mathcal{S}$ and a target domain $\mathcal{T}$. We have access to the source data $x_s \in \mathcal{S}$ with pixel-level labels y_s and the target data $x_t \in \mathcal{T}$ without labels. Our overall framework is shown in Figure 2. Feature extractor G takes images x_s and x_t as inputs and produces intermediate feature maps f_s and f_t; Classifier F takes features f_s and f_t from G as inputs and predicts C-dimensional segmentation softmax outputs P_s and P_t; Discriminator D is a CNN-based binary classifier with a fully-convolutional output to distinguish whether the input (P_s or P_t) is from the source or target domain.

To address aforementioned category-conditional shift, we propose a framework that alternatively generates pointwise perturbations with multiple attack objectives and defenses against these perturbed copies via an adversarial training procedure. Since our framework conducts perturbations for each point independently, it circumvents the interference of different categories. Our learning procedure can also be seen as a form of active learning or hard example mining, where the model is enforced to minimize the worst case error when features are perturbed by adversaries. Our framework consists of three steps as follow:

Step 1: Initialize G and F. We train both the feature extractor G and classifier F with source samples. Since we need G and F to learn task-specific feature representations, this step is crucial. Specifically, we train the feature extractor and classifier by minimizing cross entropy loss as follow:

$$L_{ce}(x_s, y_s) = -\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{C} y_s^{(h,w,c)} \log P_s^{(h,w,c)}, \quad (1)$$

where input image size is $H \times W$ with C categories, and $P_s = (F \circ G)(x_s)$ is the softmax segmentation map produced by the classifier.

Step 2: Generation of adversarial features. The adversarial features $f_{s*/t*}$ are initialized with $f_{s/t}$ extracted by G from $x_{s/t}$, and iteratively updated with our proposed I-FGSPM combining several attack objectives. These perturbed features are designed to confuse the discriminator and the classifier with our tailored attack objectives.

Step 3: Training with adversarial features. With adversarial features from step 2, it is crucial to set proper training objectives to defense against the perturbations and enable the classifier to produce consistent predictions. Besides, robust classifier and discriminator can contiguously generate confusing adversarial features for further training.

During training, we freeze G after step 1, and alternate step 2 and step 3 to obtain a robust classifier against domain shift as well as category-conditional shift. We detail the step 2 and step 3 in the following sections.

Generation of Adversarial Features

In this part, we first introduce the attack objectives and then propose our Iterative Fast Gradient Sign Preposed Method (I-FGSPM) for combining multiple attack objectives.

Attack objectives. On the one hand, the generated perturbations are supposed to extrapolate the features towards domain-invariant regions. Therefore, they are expected to confuse the discriminator which aims to distinguish source domain from the target one by minimizing the loss function in Eq. (2), so that the gradient of $L_{adv}(P)$ is capable of pro-

ducing perturbations that help fill in the domain gap.

$$L_{adv}(P) = -\mathbb{E}[\log(D(P_s))] - \mathbb{E}[\log(1 - D(P_t))]. \quad (2)$$

On the other hand, to further improve the robustness of the classifier, the adversarial features should capture the vulnerability of the model (e.g., the tendency of classifier to collapse into head classes). In this regard, we conduct an adversarial attack on segmentation classifier and employ the Lovász-Softmax (Berman, Rannen Triki, and Blaschko 2018) as our attack objective in Eq (3). Since the perturbations are actually hard examples for the classifier, they carry rich information of the failure mode of the segmentation classifier. Lovász-Softmax is a smooth version of the jaccard index and we empirically show that our attack objective can produce proper segmentation perturbations as well as boosting the adaptability of the model.

$$L_{seg}(P_s, y_s) = \text{Lovász-Softmax}(P_s, y_s). \quad (3)$$

In addition, excessive perturbations might degenerate the semantic information of feature maps, so that we control the L_2-distance between the original features and their perturbed copies to self-adaptively constraint their divergence. Eventually, we accumulate gradient maps from all attack objectives and generate adversarial features with our proposed Iterative Fast Gradient Sign preposed Method (I-FGSPM).

Original I-FGSM. While we can follow the practice in (Liu et al. 2019) to directly regard the gradients as perturbations, we have empirically found that this strategy may suffer from gradient vanishing in our case. Instead, we draw a link from adversarial attack to generate more stable and reasonable perturbations. Specifically, to generate the perturbations, we adopt the Iterative Fast Gradient Sign Method (I-FGSM) (Kurakin, Goodfellow, and Bengio 2016) as Eq. (4):

$$\begin{aligned} f_{s*}^{k+1} =& f_{s*}^{k} + \epsilon \cdot \text{sign}(\beta_1 \nabla_{f_{s*}^k} L_{seg}(P_{s*}^k, y_s) \\ & - \beta_2 \nabla_{f_{s*}^k} L_2(f_{s*}^k, f_s) + \beta_3 \nabla_{f_{s*}^k} L_{adv}(P_{s*}^k)), \end{aligned} \quad (4)$$

where β_1, β_2 and β_3 indicate the hyper-parameters to balance the gradients values from different attack objectives and ϵ represents the magnitude of the overall perturbation. We repeat this generating process for K iterations with $k \in \{0, 1, \cdots, K-1\}$. It is noteworthy that $f_{s*}^0 = f_s$.

However, this practice also raises some concerns when we execute I-FGSM under the circumstance of multiple adversarial attack objectives. Such concerns are attributed to the significant gradient gaps among different attack objectives. It is worth mentioning that, at each iteration, the final signs of the accumulated gradients are indeed dominated by one of the attack objectives. As illustrated in Figure 3, we plot the gradient log-intensity of each attack objective by using Eq. (4) to obtain adversarial features. In Figure 3, the gradients of L_{seg} and L_2 alternatively surpasses the others overwhelmingly with at least several orders of magnitude and therefore determine the final signs. Furthermore, the gradient value of a specific attack objective fluctuates by varying iterations and does not appear proportional tendency with its counterparts, so that it is not trivial to balance the gradient perturbations by simply adjusting the trade-off constants.

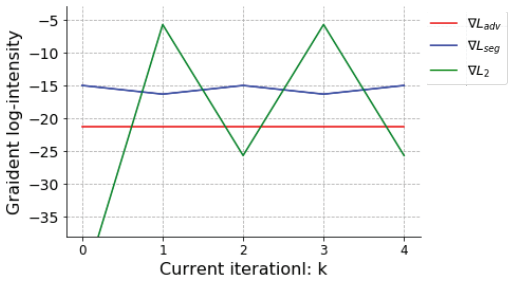


Figure 3: Gradient log-intensity tendencies with I-FGSM method in generation procedure.

Our I-FGSPM. To this end, we propose the Iterative Fast Gradient Sign Preposed Method (**I-FGSPM**) to fully exploit the contributions of each individual attack objective. Rather than placing the sign operator at the end of the overall gradient fusion which suffers from the gradient domination issue, we instead put ahead the sign calculations of each adversarial gradient and then balance these signed perturbations with intensity ϵ. The procedure is formulated as Eq. (5) and (6) for target and source perturbations respectively.

$$\begin{aligned} f_{t*}^{k+1} = & f_{t*}^{k} + \epsilon_1 \text{sign}(\nabla_{f_{t*}^k} L_{adv}(P_{t*}^k)) \\ & - \epsilon_2 \text{sign}(\nabla_{f_{t*}^k} L_2(f_{t*}^k, f_t)), \end{aligned} \quad (5)$$

$$\begin{aligned} f_{s*}^{k+1} = & f_{s*}^{k} + \epsilon_1 \text{sign}(\nabla_{f_{s*}^k} L_{adv}(P_{s*}^k)) \\ & - \epsilon_2 \text{sign}(\nabla_{f_{s*}^k} L_2(f_{s*}^k, f_s)) \\ & + \epsilon_3 \text{sign}(\nabla_{f_{s*}^k} L_{seg}(P_{s*}^k, y_s)). \end{aligned} \quad (6)$$

Training with Adversarial Features

Now, we are equipped with adversarial features which can reduce the domain gap and capture the vulnerability of the classifier. To obtain a domain-invariant classifier F and a robust domain discriminator D, we should design proper constraints that can guide the learning process to utilize these adversarial features to train F and D.

For this purpose, the solution appears straightforward for the source domain since we still hold the strong supervision y_s for its adversarial features f_{s*}. On the contrary, when it comes to the unlabeled target domain, we are supposed to explore other supervision signals to satisfy the goal. Our considerations are two folds. First, we follow the practice in (Liu et al. 2019) that forces the classifier to make consistent predictions for f_t and f_{t*} as follow:

$$L_{cst}(P_t, P_{t*}) = \mathbb{E}[\|P_t - P_{t*}\|_2]. \quad (7)$$

Noted that this action does not guarantee the discriminative and reductive information for specific tasks. Instead, as the perturbations intend to confuse the classifier, the prediction maps of adversarial features are empirically subject to have more uncertainty with increasing entropy. To address this issue, we draw on the idea of the entropy minimization technique (Springenberg 2015; Long et al. 2018) as Eq (8) to provide extra supervision, which can be viewed as

Table 1: Results of adapting GTA5 to Cityscapes. The tail classes are highlighted in **blod**. The top and bottom parts correspond to VGG-16 and ResNet-101 based models separately.

Method	road	sidewalk	building	**wall**	**fence**	**pole**	**light**	**sign**	veg	**terrain**	sky	**person**	**rider**	car	**truck**	**bus**	**train**	**mbike**	**bike**	mIoU
ASN (Tsai et al. 2018)	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
CLAN (Luo et al. 2019)	88.0	30.6	79.2	23.4	20.5	26.1	23.0	14.8	81.6	34.5	72.0	45.8	7.9	80.5	26.6	29.9	0.0	10.7	0.0	36.6
Ours	88.4	34.2	77.6	23.7	18.3	24.8	24.9	12.4	80.7	30.4	68.6	48.9	17.9	80.8	27.0	27.2	6.2	19.1	10.2	38.0
Source Only	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
ASN (Tsai et al. 2018)	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
CLAN (Luo et al. 2019)	87.0	27.1	79.6	27.3	23.3	28.3	35.5	24.2	83.6	27.4	74.2	58.6	28.0	76.2	33.1	36.7	6.7	31.9	31.4	43.2
AdvEnt(Vu et al. 2019)	**89.9**	36.5	**81.6**	29.2	25.2	**28.5**	32.3	22.4	**83.9**	34.0	77.1	57.4	27.9	83.7	29.4	39.1	1.5	28.4	23.3	43.8
ASN + Weighted CE	82.8	**42.4**	77.1	22.6	21.8	28.3	**35.9**	**27.4**	80.2	25.0	77.2	58.1	26.3	59.4	25.7	32.7	3.6	29.0	31.4	41.4
ASN + Lovász	88.0	28.6	80.7	23.6	14.8	25.9	33.3	19.6	82.8	31.1	74.9	58.1	24.6	72.6	34.2	31.2	0.0	24.9	36.4	41.3
Ours	85.6	32.8	79.0	**29.5**	**25.5**	26.8	34.6	19.9	83.7	**40.6**	**77.9**	**59.2**	**28.3**	**84.6**	**34.6**	**49.2**	**8.0**	**32.6**	**39.6**	**45.9**

a soft-assignment variant of the pseudo-label cross entropy loss (Vu et al. 2019).

$$L_{ent}(P) = \mathbb{E}[\frac{-1}{\log(C)} \sum_{c=1}^{C} P^{(h,w,c)} \log P^{(h,w,c)}]. \quad (8)$$

Finally, by combining the objectives in (3), (7) and (8), we are capable of obtaining robust and discriminative classifier F as follow, where α_1, α_2 and α_3 are trade-off factors:

$$\min_F L_{cls} = L_{seg}(P_{s*}, y_s) + L_{seg}(P_s, y_s) + \alpha_1 L_{cst}(P_t, P_{t*}) + \alpha_2 L_{ent}(P_t) + \alpha_3 L_{ent}(P_{t*}). \quad (9)$$

In addition, we conduct a similar procedure to defense against domain-related perturbations, which forces the discriminator D to assign the same domain labels for the mutated features with respect to their original ones. Furthermore, it is beneficial for the discriminator to contiguously generate perturbations that extrapolate the features towards more domain-invariant regions and then bridge the domain discrepancy more effectively.

Experiments

Dataset

We evaluate our method along with several state-of-the-art algorithms on two challenging *synthesized-2-real* UDA benchmarks, i.e., **GTA5 → Cityscapes** and **SYNTHIA → Cityscapes**. Cityscapes is a real-world image dataset, consisting of 2,975 images for training and 500 images for validation. GTA5 contains 24,966 synthesized frames captured from the video game. We use the 19 classes of GTA5 in common with the Cityscapes for adaptation. SYNTHIA is a synthetic urban scenes dataset with 9,400 images. Similar to Vu et al. (2019), We train our model with 16 common classes in both SYNTHIA and Cityscapes, and evaluate the performance on 13-class subsets.

Implementations details

We use PyTorch for implementation. Similar to Tsai et al. (2018), we utilize the DeepLab-v2 (Chen et al. 2017a) as our backbone segmentation network. We employ Atrous Spatial Pyramid Pooling (ASPP) as classifier followed by an up-sampling layer with softmax output. For domain discriminator D, we use the one in DCGAN (Radford, Metz, and Chintala 2015) but exclude batch normalization layers. Our experiments are based on two different network architectures: VGG-16 (Simonyan and Zisserman 2014) and ResNet-101 (He et al. 2016). During training, we use SGD (Bottou 2010) for G and C with momentum 0.9, learning rate 2.5×10^{-4} and weight decay 10^{-4}. We use Adam (Kingma and Ba 2014) with learning rate 10^{-4} to optimize D. And we follow the polynomial annealing procedure (Chen et al. 2017a) to schedule the learning rate. When generating adversarial features, the iteration K of I-FGSPM is set to 3. Note that we set the ϵ_1, ϵ_2 and ϵ_3 in Eq. (5) and (6) as 0.01, 0.002 and 0.011 separately. α_1, α_2 and α_3 are 0.2, 0.002 and 0.0005 separately.

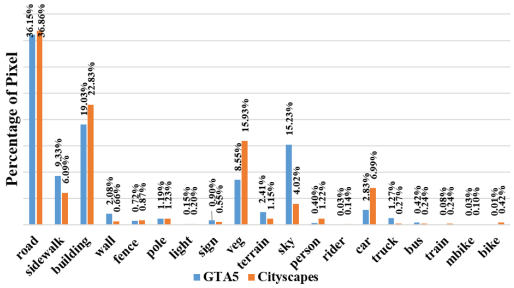


Figure 4: Category distribution on GTA5 → Cityscapes.

Result Analysis

We compare our model with several state-of-the-art domain adaptation methods on semantic segmentation performance in terms of mIoU. Table 1 shows that our ResNet-101 based model brings +9.3% gain compared to source only model on GTA5 → Cityscapes. Besides, our method also outperforms state-of-the-arts over +1.4% and +2.1% in mIoU on VGG-16 and ResNet-101 separately. To further illustrate the effectiveness of our method on tail classes, we show the marginal category distributions counted in 19 common classes on GTA5 and Cityscapes datasets in Figure 4, and highlight the tail classes with **blod** in Table 1. For example, the category "bike" accounts for only 0.01% ratio in the GTA5 category distribution, and the ResNet-101 based adversarial alignment methods suffer from a huge performance degradation compared to the source only model. Specifi-

Table 2: Results of adapting SYNTHIA to Cityscapes. The tail classes are highlighted in **blod**.

Method	road	sidewalk	building	**light**	**sign**	veg	sky	person	**rider**	car	**bus**	**mbike**	**bike**	$mIoU_{13}$
ASN (Tsai et al. 2018)	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
CLAN (Luo et al. 2019)	80.4	30.7	74.7	1.4	8.0	77.1	79.0	46.5	8.9	73.8	18.2	2.2	9.9	39.3
Ours	82.9	31.4	72.1	10.4	9.7	75.0	76.3	48.5	15.5	70.3	11.3	1.2	29.4	41.1
Source Only	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
ASN (Tsai et al. 2018)	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
CLAN (Luo et al. 2019)	81.3	37.0	**80.1**	16.1	13.7	78.2	81.5	53.4	21.2	73.0	32.9	22.6	30.7	47.8
AdvEnt (Vu et al. 2019)	**87.0**	**44.1**	79.7	4.8	7.2	80.1	**83.6**	56.4	**23.7**	72.7	32.6	12.8	33.7	47.6
ASN + Weighted CE	74.9	37.6	78.1	10.5	10.2	76.8	78.3	35.3	20.1	63.2	31.2	19.5	43.3	44.5
ASN + Lovász	77.3	40.0	78.3	14.4	13.7	74.7	83.5	55.7	20.9	70.2	23.6	19.3	40.5	47.1
Ours	86.4	41.3	79.3	**22.6**	**17.3**	**80.3**	81.6	**56.9**	21.0	**84.1**	**49.1**	**24.6**	**45.7**	**53.1**

cally, AdvEnt can deliver a +7.2% performance improvement on average, but the category "bike" itself suffers 12.7% performance degradation. On the contrary, our approach can still improve the performance of the "bike" category by benefiting from the pointwise perturbation strategy. In fact, our framework can achieve the best performance at the majority of tail categories, showing the effectiveness of our algorithm in mitigating the category-conditional shift.

Table 2 provides the comparative performance on SYNTHIA → Cityscapes. SYNTHIA has significantly different layouts as well as viewpoints compared to Cityscapes, and less training samples than GTA5. Hence, models trained in SYNTHIA might suffer from serious domain shift when generalized into Cityscapes. It is noteworthy that our adversarial perturbation framework generates hard examples that strongly resist adaptation, thus our model can efficiently improve performance in the difficult task by considering these augmented features. As a result, our method significantly outperforms the state-of-the-art methods by +1.8% and +5.5% in mIoU based on VGG-16 and ResNet-101 separately. Specifically, even when compared to CLAN method, which aims at aligning category-level joint distribution, our framework still achieves higher performance on tail classes. Some qualitative results are presented in Figure 6.

Furthermore, we re-implement ASN with some category balancing mechanisms (e.g., weighted cross entropy and Lovász-Softmax loss) based on ResNet-101 for fair comparison. As shown in Table 1 and 2, we show that only ASN + Lovász brings +1.2% gain in SYNTHIA → Cityscapes, while others even suffer from performance degradation. As shown in Figure 4 and 5, marginal category distributions are varying across domains, and thus re-weighting mechanisms can not guarantee adaptability on the target domain.

Ablation Study

Different Attack Methods. A basic problem of our framework is how to generate proper perturbations. We compare several attack methods widely used in adversarial attack community and their modified sign-preposed versions. Specifically, we compare our proposed I-FGSPM with I-FGSM and modified sign-preposed version of FGSM (Goodfellow, Shlens, and Szegedy 2014) as well as

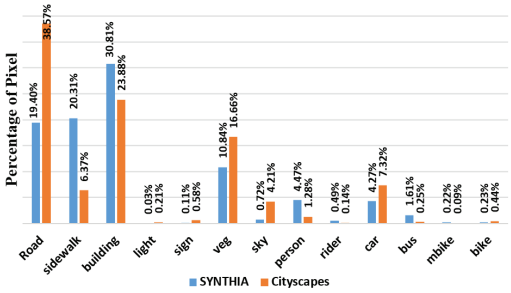


Figure 5: Category distribution on SYNTHIA → Cityscapes.

Table 3: Evaluation on different attack methods.

Attack Method	$mIoU_{13}$ (SYNTHIA)
None	44.8
I-FGSM	51.8
FGSPM	52.9
MI-FGSPM	52.2
I-FGSPM (Ours)	**53.1**

Momentum I-FGSM (MI-FGSM) (Dong et al. 2018). Furthermore, we also provide a "None" version without any attacks. As illustrated in Table 3, ResNet-101 based adversarial attack methods bring obvious gain against "None" version. With sign-preposed operation, our I-FGSPM achieves +1.3% improvement compared to I-FGSM. FGSPM is the non-iterative version of our I-FGSPM and achieves comparable performance against I-FGSPM. Note that though MI-FGSM achieves remarkable results in adversarial attacks area, its sign-preposed version MI-FGSPM might excessively enlarge the divergence between original features with adversarial features, and causes performance degradation when employed by our framework.

Different perturbing layers. One natural question is whether it is better to perturb the input or the hidden layers of model. Szegedy et al. (2013) reported that adversarial perturbations yield the best regularization when applied to the hidden layers. Our experiments with ResNet-101 shown in Table 4 also verify that perturbing in feature-level achieves

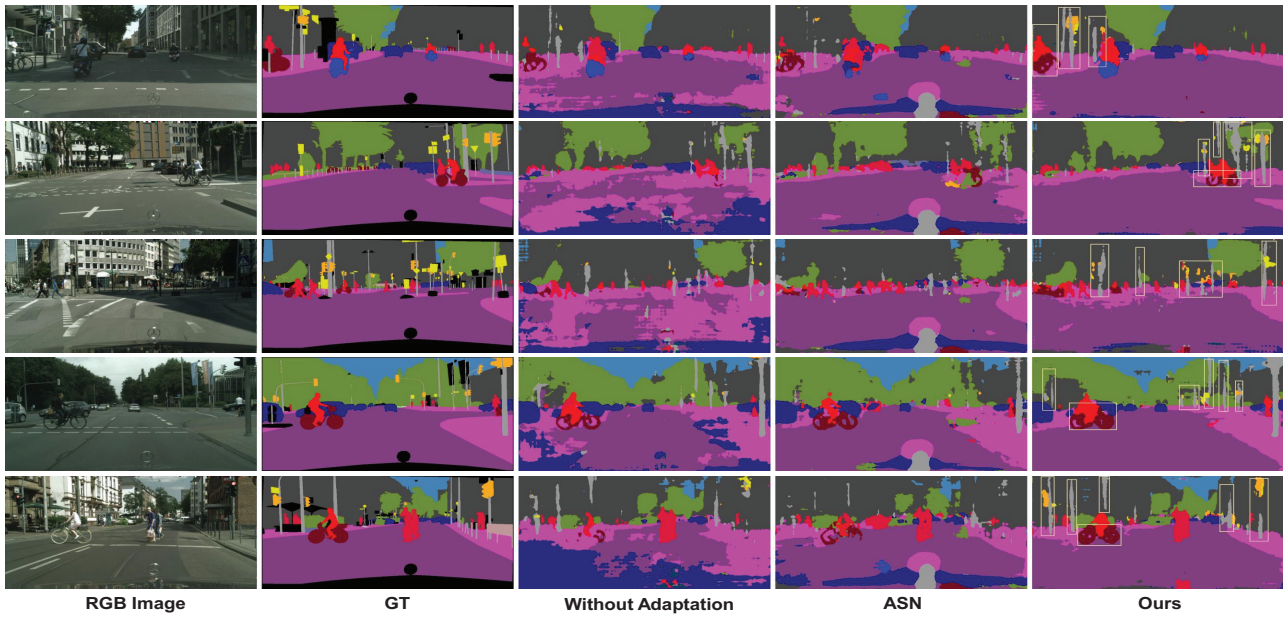


Figure 6: Qualitative results of UDA segmentation for SYNTHIA → Cityscapes. Along with each target image and its corresponding ground truth, we present the results of source only model (without adaptation), ASN and ours respectively.

Table 4: Evaluation on different perturbing layers.

Layer	mIoU_{13} (SYNTHIA)
Pixel-level	50.4
After layer1	45.0
After layer2	49.8
After layer3	50.6
After layer4 (Ours)	**53.1**

the best result. These might boil down to that the activation of hidden units can be unbounded and very large when perturbing the hidden layers (Goodfellow, Shlens, and Szegedy 2014). We also find that perturbing deeper hidden layers can further benefit our framework.

Component Analysis. We study how each component affects overall performance in terms of mIoU based on ResNet-101. As shown in the top part of Table 5, starting with source only model trained with Lovász-Softmax, we notice that the effect of Lovász-Softmax loss varies across different UDA tasks, which might depend on how different the marginal distributions across two domains are. Entropy minimization strategy can bring improvement on both benchmarks but lead to strong class biases, which has been verified in AdvEnt (Vu et al. 2019), while our overall model not only significantly lifts mIoU, but also remarkably alleviates category biases specially for tail classes.

As illustrated in the bottom part of Table 5, we consider our basic training strategy in step 1 as a component, and replace it with ASN. By cooperating with our perturbations strategy, ours + ASN brings +3.8% and +7.0% gain, while ASN + Lovàsz + Entropy only gets +0.9% and +1.5% improvement against ASN on GTA5 to Cityscapes and SYNTHIA to Cityscapes separately. A possible reason is that ASN can shape the feature extractor biased towards the head classes and miss representations from tail classes.

Table 5: Ablation studies of each component. "S" represents our strategy as discussed in step 1 while "ASN" indicates that our network weights are pre-trained by ASN in step 1.

Base	Perturbation	Lovász	Entropy	mIoU (GTA5)	mIoU_{13} (SYN)
S				36.6	38.6
S		√		35.0	41.3
S			√	41.8	42.5
S		√	√	38.5	44.8
S	√			41.7	45.7
S	√	√		44.6	49.9
S	√		√	43.6	47.0
S	√	√	√	**45.9**	**53.1**
ASN				41.4	45.9
ASN		√	√	42.3	47.4
ASN	√	√	√	45.2	52.9

Conclusion

In this paper, we reveal that adversarial alignment based segmentation DA might be dominated by head classes and fail to capture the adaptability of different categories evenly. To address this issue, we proposed a novel framework that iteratively exploits our improved I-FGSPM to extrapolate the perturbed features towards more domain-invariant regions and defenses against them via an adversarial training procedure. The virtues of our method lie in not only the adaptability of model but that it circumvents the intervention among different categories. Extensive experiments have verified that our approach significantly outperforms the state-of-the-arts, especially for the hard tail classes.

References

Berman, M.; Rannen Triki, A.; and Blaschko, M. B. 2018. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *CVPR*, 4413–4421.

Bottou, L. 2010. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*. 177–186.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017a. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE T-PAMI* 40(4):834–848.

Chen, Y.-H.; Chen, W.-Y.; Chen, Y.-T.; Tsai, B.-C.; Frank Wang, Y.-C.; and Sun, M. 2017b. No more discrimination: Cross city adaptation of road scene segmenters. In *ICCV*, 1992–2001.

Chen, Y.; Li, W.; and Van Gool, L. 2018. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *CVPR*, 7892–7901.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *CVPR*.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting adversarial attacks with momentum. In *CVPR*, 9185–9193.

Ganin, Y., and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *ICML*, 1180–1189.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *JMLR* 17(1):2096–2030.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NeuralIPS*, 2672–2680.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hoffman, J.; Wang, D.; Yu, F.; and Darrell, T. 2016. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*.

Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.-Y.; Isola, P.; Saenko, K.; Efros, A.; and Darrell, T. 2018. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kurakin, A.; Goodfellow, I.; and Bengio, S. 2016. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*.

Liu, H.; Long, M.; Wang, J.; and Jordan, M. 2019. Transferable adversarial training: A general approach to adapting deep classifiers. In *ICML*, 4013–4022.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. 2015. Learning transferable features with deep adaptation networks. In *ICML*, 97–105.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2017. Deep transfer learning with joint adaptation networks. In *ICML*, 2208–2217.

Long, M.; Cao, Y.; Cao, Z.; Wang, J.; and Jordan, M. I. 2018. Transferable representation learning with deep adaptation networks. *IEEE T-PAMI*.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*, 3431–3440.

Luo, Y.; Zheng, L.; Guan, T.; Yu, J.; and Yang, Y. 2019. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *CVPR*, 2507–2516.

Radford, A.; Metz, L.; and Chintala, S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Richter, S. R.; Vineet, V.; Roth, S.; and Koltun, V. 2016. Playing for data: Ground truth from computer games. In *ECCV*, 102–118.

Ros, G.; Sellart, L.; Materzynska, J.; Vazquez, D.; and Lopez, A. M. 2016. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 3234–3243.

Saito, K.; Ushiku, Y.; and Harada, T. 2017. Asymmetric tri-training for unsupervised domain adaptation. In *ICML*, 2988–2997.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Springenberg, J. T. 2015. Unsupervised and semi-supervised learning with categorical generative adversarial networks. *arXiv preprint arXiv:1511.06390*.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Tsai, Y.-H.; Hung, W.-C.; Schulter, S.; Sohn, K.; Yang, M.-H.; and Chandraker, M. 2018. Learning to adapt structured output space for semantic segmentation. In *CVPR*, 7472–7481.

Volpi, R.; Namkoong, H.; Sener, O.; Duchi, J. C.; Murino, V.; and Savarese, S. 2018. Generalizing to unseen domains via adversarial data augmentation. In *NeuralIPS*, 5334–5344.

Vu, T.-H.; Jain, H.; Bucher, M.; Cord, M.; and Pérez, P. 2019. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, 2517–2526.

Wu, Y.; Winston, E.; Kaushik, D.; and Lipton, Z. 2019. Domain adaptation with asymmetrically-relaxed distribution alignment. In *ICML*, 6872–6881.

Xu, R.; Li, G.; Yang, J.; and Lin, L. 2019. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *ICCV*.

Zhang, Y.; Qiu, Z.; Yao, T.; Liu, D.; and Mei, T. 2018. Fully convolutional adaptation networks for semantic segmentation. In *CVPR*, 6810–6818.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *CVPR*, 2881–2890.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2223–2232.

Zou, Y.; Yu, Z.; Vijaya Kumar, B.; and Wang, J. 2018. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*, 289–305.