Domain Adaptive Semantic Segmentation with Regional Contrastive Consistency Regularization

Qianyu Zhou[1*] Chuyun Zhuang[1*] Xuequan Lu[2†] Lizhuang Ma[1†]
[1]Shanghai Jiao Tong University, [2]Deakin University
{zhouqianyu, fallen}@sjtu.edu.cn, xuequan.lu@deakin.edu.au, ma-lz@cs.sjtu.edu.cn

Abstract

Unsupervised domain adaptation (UDA) aims to bridge the domain shift between the labeled source domain and the unlabeled target domain. However, most existing works perform the global-level feature alignment for semantic segmentation, while the local consistency between the regions has been largely neglected, and these methods are less robust to changing of outdoor environments. Motivated by the above facts, we propose a novel and fully end-to-end trainable approach, called regional contrastive consistency regularization (RCCR) for domain adaptive semantic segmentation. Our core idea is to pull the similar regional features extracted from the same location of different images to be closer, and meanwhile push the features from the different locations of the two images to be separated. We innovatively propose momentum projector heads, where the teacher projector is the exponential moving average of the student. Besides, we present a region-wise contrastive loss with two sampling strategies to realize effective regional consistency. Finally, a memory bank mechanism is designed to learn more robust and stable region-wise features under varying environments. Extensive experiments on two common UDA benchmarks, i.e., *GTAV to Cityscapes and SYNTHIA to Cityscapes, demonstrate that our approach outperforms the state-of-the-art methods.*

1. Introduction

Semantic segmentation aims to assign a semantic class label to each pixel for a given image, and it is a fundamental task in computer vision. It plays an essential role in many downstream applications such as autonomous driving, medical analysis, and remote sensing. Deep learning models and techniques of semantic segmentation [6, 7, 39, 41, 85] have achieved great progresses in popular semantic segmentation benchmarks [14,18,38]. However, these methods typically require a large amount of labeled training data, and such sufficient labeled data may not be always available in real-world scenarios. Labeling these pixel-wise images is extremely expensive, and time-consuming [14]. For instance, pixel-wise labeling for one Cityscapes image takes 90 minutes on average [14]. Recent progress in computer graphics such as rendering makes it possible to automatically generate synthetic images with free pixel-wise annotations from virtual 3D environments, *e.g.*,, GTAV [56], SYNTHIA [57], Virtual KITTI [22], etc. Thus, a natural idea is using synthetic data to supervise the segmentation model instead of real data. However, such data cannot fully match the real-world distributions to guarantee reliable performance due to the existing domain shifts. Thus, it is necessary to reduce the labeling cost and improve the generalization ability of the segmentation models under different distributions.

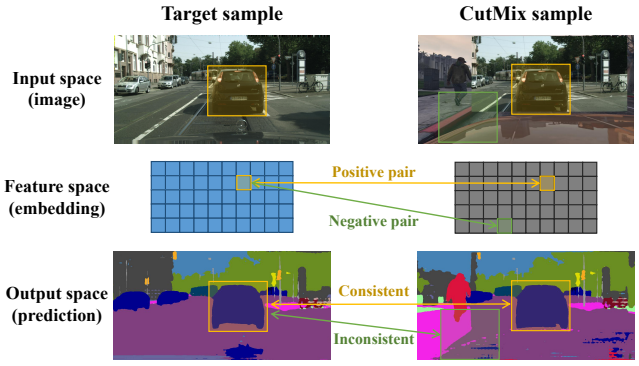

Figure 1. Previous domain adaptation methods overlook the regional consistency across different domains. To address this problem, our key idea builds on region-level contrastive learning by maximizing the inter-region differences and minimizing intra-region disagreement. 1). On the output space, the predicted label should be invariant to cross-domain environmental augmentations, *e.g.,* CutMix. 2). On the feature space, we pull the similar regional embeddings extracted on the same location from the target image and mixed image to be closer, and push the dissimilar embeddings from the different locations of the two images to be separated.

[0]*Equal Contribution. † Joint Corresponding author.

To cope with this problem, unsupervised domain adaptation (UDA) for semantic segmentation has been recently explored and has rapidly grown with a bunch of approaches. This task aims to bridge the existing domain gap between the labeled source domain and the unlabeled target domain. Many mainstream approaches perform the adaptation in input-level [11, 24, 25, 27, 29, 37, 77, 80, 81], feature-level [4, 12, 12, 42, 43, 64, 65, 67, 91, 92], and output-level [9, 44, 45, 52, 63, 69, 78]. However, most of them heavily depend on the computationally involved adversarial objectives [9, 44, 45, 52, 63, 64, 69], offline self-training [37, 46, 48, 58, 91, 92] and image translation [25, 27, 29, 37, 77, 80, 81], which makes the training process too complicated and hard to converge.

Recently, consistency regularization emerges [2, 13, 20, 47, 54, 61, 76, 87, 88], and tackles this problem by employing the consistency constraint on the target prediction between the student model and the teacher model, respectively. This kind of consistency-based method usually performs the feature-level domain alignment between the student model and the teacher model with an online ensemble. The teacher model is an exponential moving average (EMA) of the student model, and then the teacher model could transfer the learned knowledge to the student.

Unfortunately, such methods usually employ an inconsistent penalty on the global level for the prediction map, while largely neglecting the region-wise consistency on the local level, *i.e.,* some contextual object occurrence should be consistent regardless of the outdoor changes of environments. We observe that only capturing the pattern information from the global level is not powerful enough to enhance the feature-level representation. If lacking this property, the segmentation result of objects will inevitably suffer from a non-marginal performance drop in the target domain. To prevent the model from abusing the contexts, we aim to make the learned representations more robust to the changing environments by exploring the regional consistency in a fine-grained manner.

Motivated by the above facts, we propose a regional contrastive consistency regularization (RCCR) framework for domain adaptive semantic segmentation, which is fully end-to-end trainable. Our key idea builds on region-level contrastive learning by maximizing the inter-region differences and minimizing intra-region disagreement. To produce cross-domain environmental changes, we extend the CutMix strategy to the cross-domain setting, *i.e.,* a random region of the unlabeled target image is cut and pasted onto the source image. Two key components are presented to tackle the aforementioned problem. Firstly, we design momentum projector heads after the encoder architecture to produce the low-dimensional features, namely, student projector and teacher projector, where the teacher projector is the exponential moving average (EMA) of the student projector. Instead of directly using the output features of the encoder, the projector heads can prevent the classifier head from overlooking too much local information for adaptation. Secondly, we present the region-wise contrastive (RWC) loss between the latent embeddings of the student and teacher projector, respectively. The main intuition is to pull the similar regional features extracted from the same location from the target image and mixed image to be closer, and meanwhile push the features from the different locations of the two images to be separated.

To further improve the power of contrastive learning in domain adaptive semantic segmentation, two techniques are proposed in the whole architecture. Firstly, we present two sampling strategies for positive and negative samples, respectively. For positive sampling, we consider the output confidence of the segmentation head in a certain location while taking the label or the pseudo label into account for negative sampling. Secondly, we introduce a memory bank mechanism to store the negative features created in the last few batches to learn more robust and stable region-wise features under varying environments.

In a nutshell, our contributions are three-fold:

- We propose a regional contrastive consistency regularization framework for domain adaptive semantic segmentation, which keep the local regional consistency on the feature space and output label space, respectively, under the cross-domain environmental augmentations.
- We present a region-wise contrastive loss, and momentum projector heads to realize effective regional consistency in domain adaptation. We also introduce a memory bank mechanism and two sampling strategies to further improve the power of the regional contrastive consistency regularization.
- We provide extensive experiments with analysis and demonstrate the state-of-the-art performance on two challenging domain adaptation benchmark datasets for semantic segmentation, *i.e.,* GTAV [56] $\rightarrow$ Cityscapes [14] and SYNTHIA [57] $\rightarrow$ Cityscapes [14].

2. Related work

2.1. Unsupervised domain adaptation for semantic segmentation

Unsupervised domain adaptation (UDA) is attracting wide attention in the past few years, and it aims to learn a generalized model on the labeled source domain and the unlabeled target domain. This problem has been well-studied in image recognition [15, 23, 30, 35, 36, 70, 84]. However, these methods only work on simple and small

classification datasets, e.g., MNIST [32] and SVHN [49], and may have quite limited performance in more challenging and higher-structured tasks, *e.g.,* semantic segmentation. Thus, researching unsupervised domain adaptation in semantic segmentation is quite necessary and important. Many recent approaches are proposed to tackle the domain gap between the source data and the target data on different levels. These domain adaptation methods mainly can be divided into three categories: namely, the input-level [11, 24, 25, 27, 29, 37, 77, 80, 81], feature-level [4, 12, 12, 42, 43, 64, 65, 67, 89, 91, 92], and output-level adaptation [9, 44, 45, 52, 63, 69, 78]. However, most recent methods [29, 37, 69, 81] involving many sophisticated sub-components, *e.g.,* computationally involved adversarial objectives [9, 44, 45, 52, 63, 64, 69], offline self-training [37,46,48,58,74,91,92] and image translation models [25, 27, 29, 37, 77], which are quite complex and hard to converge, and cannot be trained in an end-to-end manner. In contrast, our proposed method is simple yet effective, and fully end-to-end trainable.

2.2. Consistency Regularization

Consistency regularization is initially proposed in semi-supervised image learning tasks [60]. Recently, this architecture and its advanced variants have achieved state-of-the-art performance in the semi-supervised learning (SSL) [5, 75, 79, 83, 86] and unsupervised domain adaptation (UDA) benchmarks [3, 13, 17, 34, 55, 76, 87, 88]. Their frameworks include a student and teacher model, where the teacher model uses the EMA weight of the student models. Besides the supervised loss, the inconsistency between the output of the two models are treated as an additional penalty to push the student model to learn more domain-invariant features from the teacher. Mixup has been recently adopted as a high-dimensional augmentation to produce perturbation on the input for the Mean Teacher architecture [10, 19, 21, 50, 61, 88]. However, most of them employ an inconsistent penalty on the global level, while largely neglecting the region-wise consistency on the local level. In contrast to these approaches, we aim to make the learned representations more robust to the changing environments on a fine-grained manner.

2.3. Contrastive Learning

Great progress in contrastive learning [1,8,31,40,51,59, 68, 68, 71–73] has been achieved by encouraging the positive pairs to get closer and pulling the negative pairs apart. For semantic segmentation tasks, [1, 68, 73] are proposed to fit the dense pixel prediction requirements. The definition of positive pairs and negative pairs can be various, and [59, 68] treated the same category samples as the positive pairs and others as the negative pairs. [40] divided the positive pairs and negative pairs according to the label distribution similarity between different patches. There are also some works that investigated the contrastive learning methods [31, 72] in Semi-Supervised Semantic Segmentation (SSS). *Our method differs from these methods in several aspects*. Firstly, we tackle a more complicated task UDA rather than SSS, where the domain shifts exist between the source and the target domain. Secondly, another main difference is that most of them only consider category-wise contrastive learning patterns while largely neglecting the region-wise consistency. In contrast, we keep the regional consistency in the feature space and output space, respectively. Finally, in addition to the introduced region-wise contrastive loss, we also take the category of samples into account in the sampling strategy.

3. Method

In this section, we describe our approach to UDA in the autonomous driving setting. Sec. 3.1 explains the notation and problem formulation; Sec. 3.2 presents the RCCR framework design, including the momentum projector head (Sec. 3.2.1), cross-domain environmental perturbations (Sec. 3.2.2), region-wise contrastive loss (Sec. 3.2.3), two sampling strategies (Sec. 3.2.4), and memory bank mechanism (Sec. 3.2.5); Sec. 3.3 describes the overall optimization and total training procedure.

3.1. Problem Formulation

In the UDA task, we have access to the source domain with labels, denoted as $D_s = \{(x_s, y_s) \mid x_s \subset \mathbb{R}^{H\times W\times 3}, y_s \subset \mathbb{R}^{H\times W}, y_s \in [1, C]\}$, and the target domain without labels denoting as $D_t = \{(x_t) \mid x_t \subset \mathbb{R}^{H\times W\times 3}\}$. Our primary goal is to bridge the domain gap between the D_s and D_t.

Feature extractor F_{enc} receives images x as input, and produces a high-dimensional feature map $M \in \mathbb{R}^{h\times w\times D}$ Then, a segmentation head F_{seg} maps M into a C-dimensional prediction map P after upsampling and the softmax layer: $P = f_{\text{seg}}(M) \in \mathbb{R}^{H\times W\times C}$. For the source domain with access to the source domain label, we optimize the network parameters θ by constraining it with the cross-entropy loss:

$$L_{CE} = -\sum_{n=1}^{H\times W}\sum_{c=1}^{C} y_s^{n,c} \log P_s^{n,c} \tag{1}$$

3.2. Regional Contrastive Consistency Regularization

3.2.1 Momentum Projector Head.

As shown in Fig. 2, we design momentum projector heads, namely student projector and teacher projector, where the teacher projector is an exponential moving average (EMA)

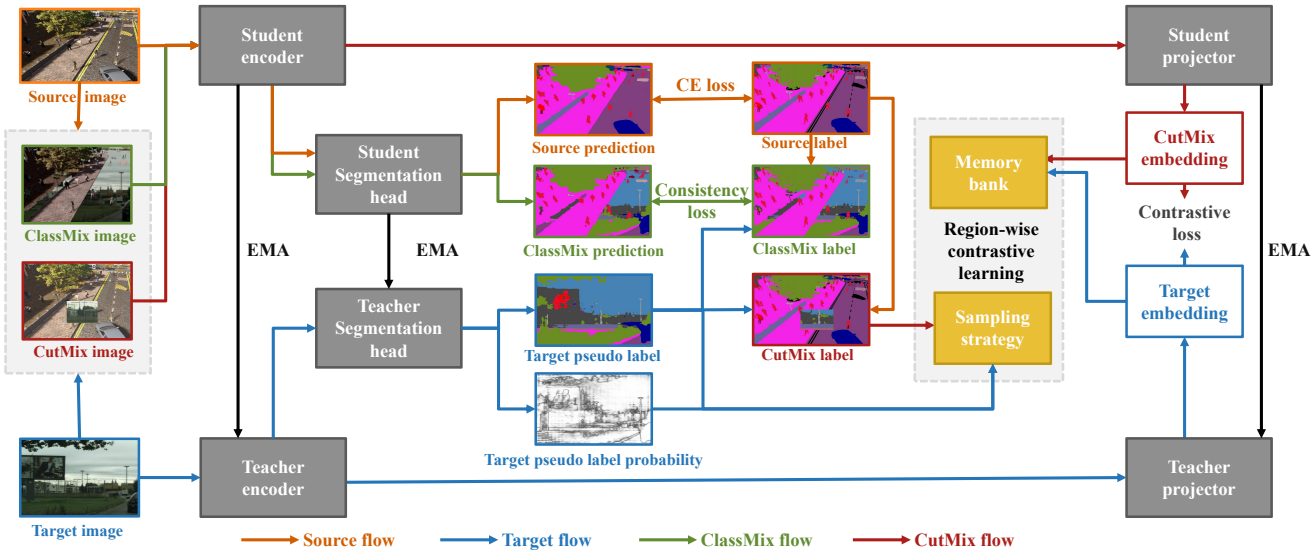

Figure 2. Overview of the regional contrastive consistency regularization (RCCR) architecture. Firstly, to produce cross-domain environmental changes, we cut a region from the target and paste it onto the source image to generate CutMix images. Then, we design the student and teacher projector to extract CutMix embeddings and target embeddings, respectively. Given these embeddings, we compute the proposed region-wise contrastive (RWC) loss by maximizing the inter-region differences and maximizing intra-region agreement. Moreover, we introduce sampling strategies and memory banks to further enhance our contrastive learning paradigm.

of the student projector, as described in Equ. 2. The projector heads F_{proj} behind the feature extractor F_{enc} aims to map the latent high-dimensional representations $M \in \mathbb{R}^{h\times w\times D}$ of F_{enc} to the low-dimensional representations $E \in \mathbb{R}^{h\times w\times K}$, where channel number $K<D$.

$$\hat{\theta}^{(t)}_{proj} = \alpha\hat{\theta}^{(t-1)}_{proj} + (1-\alpha)\theta^{(t)}_{proj} \tag{2}$$

The main intuition of using the embeddings E of the projector rather than output feature M of the feature extractor is to prevent losing too many semantic cues in the adaptation and overlooking the regional consistency on the local level for contrastive learning. As for the network architecture of F_{proj}, we implement it with two consecutive 1×1 convolutional layers with ReLU. This projector head F_{proj} maps M into a K-dimensional embedding map: $E = F_{\text{proj}}(M) \in \mathbb{R}^{h\times w\times K}$. The size h and w are eight times down-sampling of the input image in the ResNet101 backbone, *i.e.*,, $h = \frac{H}{8}$, $w = \frac{W}{8}$.

3.2.2 Cross-Domain Environmental Perturbations.

To produce the perturbations of the outdoor environments, we extend the semi-supervised augmentation strategy *e.g.*, CutMix [19], to the cross-domain setting, thus creating augmented inputs, denoted by x_{cut}. To be specific, we form x_{cut} by randomly cutting a square region from the target image x_t and paste it to the same location in a corresponding source image x_s, given as Equ. 3

$$x_{cut} = m \odot x_t + (1-m) \odot x_s, \tag{3}$$

The side length S and the top-left coordinate of the square region should be divided by 8. After that, we can feed the x_{cut} to the student model (encoder F_{enc} following by F_{proj}) to get the CutMix embedding E_{cut}, and we then obtain the target embedding $\hat{E}_t$ after we feed x_t to the teacher model. The reason why we use CutMix [19] for contrastive learning is that a location in the projector embedding maps to a 8×8 pixels region in the input image, and CutMix can achieve region-level cross-domain mixup for further local feature alignment.

3.2.3 Region-wise Contrastive Loss.

The core idea of region-wise contrastive (RWC) loss is to keep the regional consistency on a fine-grained level by maximizing the inter-region differences and minimizing the intra-region disagreements. In other words, we aim to pull the embedding on the same location of the overlap region between the CutMix embedding E_{cut} and the target embedding $\hat{E}_t$ to be closer, and push the embeddings on other locations to be separated. The proposed contrastive loss function is defined as follows:

$$L_{cont} = \frac{1}{s^2}\sum_{i=1}^{h}\sum_{j=1}^{w}\Phi^{i,j}L_{cont(E_{cut(i,j)})} \tag{4}$$

$$\mathcal{L}_{cont(E_{cut(i,j)})} = -\log \frac{exp(sim(E_{cut(i,j)}, \hat{E}_{t(i,j)})/\tau)}{exp(sim(E_{cut(i,j)}, \hat{E}_{t(i,j)})/\tau) + \Delta} \tag{5}$$

$$\Delta = \sum_{k=1}^{h}\sum_{l=1}^{w} \Omega^{k,l} exp(sim(E_{cut(i,j)}, \hat{E}_{t(k,l)})/\tau) + \sum_{k=1}^{h}\sum_{l=1}^{w} \Omega^{k,l} exp(sim(E_{cut(i,j)}, E_{cut(k,l)})/\tau) \tag{6}$$

$$\Phi^{i,j} = \mathbf{1}\{(i,j) \subset O(E_{cut}, \hat{E}_t)\} \tag{7}$$

$$\Omega^{k,l} = \mathbf{1}\{(k,l) \neq (i,j)\} \tag{8}$$

where $O(E_{cut}, \hat{E}_t)$ means the overlap region between the CutMix embedding E_{cut} and the target embedding $\hat{E}_t$, and $s = \frac{S}{8}$ is the size of $O(E_{cut}, \hat{E}_t)$. $sim(u, v) = u^\top v/\|u\|\|v\|$ is the cosine similarity between two embedding vectors with temperature term τ. For any $E_{cut(i,j)}$, there is only one relevant positive pair $(E_{cut(i,j)}, \hat{E}_{t(i,j)})$, where $\hat{E}_{t(i,j)}$ denotes the projector embedding in the same location (i, j) of the target image. The different locations in E_{cut} and $\hat{E}_t$ are all negative samples for $E_{cut(i,j)}$, and therefore there are $(2hw - 2)$ relevant negative pairs.

Thus, the discrepancy between the elements of the positive pair should be minimized, thus maximizing the intra-region agreements and performing the intra-domain adaptation. On the other hand, the discrepancy between the elements of the negative pair should be constrained, thus maximizing the inter-region difference and performing the inter-domain adaptation. Note that our proposed RWC loss is a kind of unidirectional contrastive learning strategy since the teacher model tends to output higher confident prediction than the student model, and therefore we only make E_{cut} close to $\hat{E}_t$.

3.2.4 Sampling Strategies.

Semantic segmentation tasks always require pixel-wise classification, and the outputs of the classifier also have obscure semantic relationships. To avoid treating every sample equally, we propose two sampling strategies of positive and negative samples in region-level contrastive learning.

Negative Sampling. We introduce random sampling and category-aware sampling strategies for negative samples. Specifically, we first randomly select half samples from the original negatives. The intuition of this sampling strategy is similar to the projector head, aiming to prevent from overlooking too much regional consistency for cross-domain segmentation. Secondly, we filter the negative embeddings that have the same label or pseudo label with $E_{cut(i,j)}$ due to the fact that some works [59,68] have proved that gathering the embeddings of the same category can be beneficial to semantic segmentation. Note that we do not enlarge the embedding distance for the same category.

Positive Sampling. As mentioned above, for every positive pair $(E_{cut(i,j)}, \hat{E}_{t(i,j)})$, we push $E_{cut(i,j)}$ to be close to $\hat{E}_{t(i,j)}$ in a specified direction, since the output of the teacher model is more credible than the student model, but this fact does not mean that the teacher embedding can be completely trusted, especially the embedding output has not been fully supervised constrained as the segmentation head. Thus, we propose a positive sampling strategy based on the segmentation output. In particular, we set a threshold δ to select positive samples, for every $E_{cut(i,j)}$, only the prediction probability in the same location of corresponding target image larger than δ will be included in contrastive learning.

3.2.5 Memory Bank Mechanism.

To further enhance the contrastive learning scheme, we develop a memory bank mechanism to better explore the latent embedding space. To be specific, the embedding outputs produced in the past few batch images are also considered as negatives in the current iteration, and we construct a memory bank to save useful information: the projector embedding and their corresponding label or pseudo label. We keep the memory bank driven by the motivation that: the environments between different images may have similar distribution and can be semantically related since they come from the same dataset. Therefore this strategy can enable learning more robust and stable region-wise features under varying environments. Besides, this method can also reduce the memory occupation and computational resources of segmentation.

3.3. Overall Optimization and End-to-End Training

After each iteration, the student's weights θ are optimized by the following loss:

$$L_{(\theta)} = L_{CE} + \lambda L_{cons} + \mu L_{cont} \tag{9}$$

where λ and μ are hyper-parameters that balance the three loss parts. And the teacher's weight $\hat{\theta}$ are updated as an exponential moving average (EMA) of the student's weights:

$$\hat{\theta}^{(t)} = \alpha\hat{\theta}^{(t-1)} + (1-\alpha)\theta^{(t)} \tag{10}$$

where α is a smoothing coefficient hyperparameter.

Algorithm 1 described the whole pipeline of the regional contrastive consistency regularization in the end-to-end training procedure. Given the labeled source domain

Algorithm 1: Regional Contrastive Consistency Regularization

Input: source domain dataset D_s and target domain dataset D_t, the student model F_θ and teacher model $\hat{F}_\theta$
Output: teacher model $(\hat{F}_{enc}, \hat{F}_{seg})$

Initialize network parameters θ randomly;
for $t \leftarrow 1$ **to** N **do**
  $x_{class}, x_{cut} \leftarrow ClassMix(x_s, x_t), CutMix(x_s, x_t)$; // produce mix images
  $P_s \leftarrow F_{seg}(F_{enc}(x_s))$; // fed source image to the student model
  $\hat{M}_t \leftarrow \hat{F}_{enc}(x_t)$; // fed target image to the teacher model
  $\hat{P}_t, \hat{E}_t \leftarrow \hat{F}_{seg}(\hat{M}_t), \hat{F}_{proj}(\hat{M}_t)$; // Get target predictions, and target embeddings
  $\hat{y}_t \leftarrow argmax(\hat{P}_t)$; // Get target pseudo label
  $y_{class} \leftarrow ClassMix(\hat{y}_t, y_s)$; // form ClassMix label
  $P_{class} \leftarrow F_{seg}(F_{enc}(x_{class}))$; // fed ClassMix image to get the ClassMix prediction
  $E_{cut} \leftarrow F_{proj}(F_{enc}(x_{cut}))$; // fed CutMix image to get the CutMix embedding
  $L_{(\theta)} = L_{CE}(y_s, P_s) + \lambda L_{cons}(y_{class}, P_{class}) + \mu L_{cont}(E_{cut}, \hat{E}_t)$; // compute total loss
  Compute $\bigtriangledown_\theta L$ by backpropagation;
  Perform one step of stochastic gradient descent on θ;
  $\hat{\theta}^{(t)}_{enc} = \alpha\hat{\theta}^{(t-1)}_{enc} + (1-\alpha)\theta^{(t)}_{enc}$; // update the teacher feature extractor by EMA
  $\hat{\theta}^{(t)}_{seg} = \alpha\hat{\theta}^{(t-1)}_{seg} + (1-\alpha)\theta^{(t)}_{seg}$; // update the teacher Segmentation head by EMA
  $\hat{\theta}^{(t)}_{proj} = \alpha\hat{\theta}^{(t-1)}_{proj} + (1-\alpha)\theta^{(t)}_{proj}$; // update the teacher projector by EMA
end
return $(\hat{F}_{enc}, \hat{F}_{seg})$

D_s and the unlabeled target domain D_t, we firstly produce ClassMix image x_{class} and CutMix image x_{cut}, respectively. The former x_{class} is fed to the student model F_θ to get the ClassMix prediction P_{class}. The source label y_s and the target pseudo label $\hat{y}_t$ are mixed as y_{class} to compute the consistency loss L_{cons} with P_{class}. Secondly, as for the regional contrastive learning, we fed the target feature map $\hat{M}_t$ to the teacher projector head F_{proj} to get the target embedding $\hat{E}_t$. Meanwhile, the latter CutMix image x_{cut} is fed into the student model to get the CutMix embedding E_{cut}. Then, we calculate the region-wise contrastive loss L_{cont} given these two embeddings. Finally, the student model is fully supervised by a Cross-Entropy loss L_{CE}. In the inference phase, we only keep $\hat{F}_{enc}$ and $\hat{F}_{seg}$ remained.

4. Experiments

In this section, we first describe the experimental setup in Sec. 4.1, including the datasets and the implementation details. Then, in Sec. 4.2, we demonstrate the state-of-the-art performance of our proposed method on two popular UDA benchmark datasets of cross-domain semantic segmentation. We also provide more qualitative results of cross-domain segmentation to validate its efficacy. Finally, in Sec. 4.3, we conduct ablation studies to investigate the role of each component in our proposed method.

4.1. Experimental Setup

4.1.1 Datasets.

Following common UDA protocols [45, 63, 66, 69], our experiments are conducted on two widely-used UDA benchmarks, *i.e.,* GTAV $\rightarrow$ Cityscapes and SYNTHIA $\rightarrow$ Cityscapes. The target datasets Cityscapes [14] is a real urban scene dataset composed of 2,975 training and 500 validation samples, with 19 semantic classes. The source datasets GTAV [56] and SYNTHIA [57] contains 24,966 and 9,400 synthetic training samples, respectively. The former GTAV [56] is a synthetic dataset rendered by the gaming engine Grand Theft Auto V (GTAV). The latter SYNTHIA is another synthetic dataset that has semantically compatible annotations with Cityscapes. We use the same 19 classes as Cityscapes for GTAV [56], and 13 of the 19 classes for SYNTHIA [57].

4.1.2 Implementation Details.

Following common UDA protocols [37, 63, 66, 78, 92], we adopt the widely used Deeplabv2 [6] framework with a ResNet101 [26] backbone as our model. The backbone is pre-trained on ImageNet [16] . Following the common practice [6], we use the "poly" learning weight decay and the power is set to 0.9. SGD optimizer is implemented with weight decay 5×10^{-4} and momentum 0.9. The base learning rate values are set to 2.5×10^{-4} and 2.5×10^{-3} for

Table 1. Comparison results (mIoU) from GTAV to Cityscapes.

Method	Venue	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motocycle	bike	**mIoU**
Source Only	-	63.3	15.7	59.4	8.6	15.2	18.3	26.9	15.0	80.5	15.3	73.0	51.0	17.7	59.7	28.2	33.1	3.5	23.2	16.7	32.9
BDL [37]	CVPR'19	91.0	44.7	84.2	34.6	27.6	30.2	36.0	36.0	85.0	43.6	83.0	58.6	31.6	83.3	35.3	49.7	3.3	28.8	35.6	48.5
IntraDA [52]	CVPR'20	90.6	37.1	82.6	30.1	19.1	29.5	32.4	20.6	85.7	40.5	79.7	58.7	31.1	86.3	31.5	48.3	0.0	30.2	35.8	46.3
SIM [69]	CVPR'20	90.6	44.7	84.8	34.3	28.7	31.6	35.0	37.6	84.7	43.3	85.3	57.0	31.5	83.8	42.6	48.5	1.9	30.4	39.0	49.2
LTIR [29]	CVPR'20	92.9	55.0	85.3	34.2	31.1	34.9	40.7	34.0	85.2	40.1	87.1	61.0	31.1	82.5	32.3	42.9	0.3	36.4	46.1	50.2
FDA [81]	CVPR'20	92.5	53.3	82.4	26.5	27.6	36.4	40.6	38.9	82.3	39.8	78.0	62.6	34.4	84.9	34.1	53.1	16.9	27.7	46.4	50.5
WLabel [53]	ECCV'20	91.6	47.4	84.0	30.4	28.3	31.4	37.4	35.4	83.9	38.3	83.9	61.2	28.2	83.7	28.8	41.3	8.8	24.7	46.4	48.2
CrCDA [28]	ECCV'20	92.4	55.3	82.3	31.2	29.1	32.5	33.2	35.6	83.5	34.8	84.2	58.9	32.2	84.7	40.6	46.1	2.1	31.1	32.7	48.6
FADA [67]	ECCV'20	92.5	47.5	85.1	37.6	32.8	33.4	33.8	18.4	85.3	37.7	83.5	63.2	39.7	87.5	32.9	47.8	1.6	34.9	39.5	49.2
LDR [77]	ECCV'20	90.8	41.4	84.7	35.1	27.5	31.2	38.0	32.8	85.6	42.1	84.9	59.6	34.4	85.0	42.8	52.7	3.4	30.9	38.1	49.5
CCM [33]	ECCV'20	93.5	57.6	84.6	39.3	24.1	25.2	35.0	17.3	85.0	40.6	86.5	58.7	28.7	85.8	49.0	56.4	5.4	31.9	43.2	49.9
ASA [90]	TIP'21	89.2	27.8	81.3	25.3	22.7	28.7	36.5	19.6	83.8	31.4	77.1	59.2	29.8	84.3	33.2	45.6	16.9	34.5	30.8	45.1
CLAN [44]	TPAMI'21	88.7	35.5	80.3	27.5	25.0	29.3	36.4	28.1	84.5	37.0	76.6	58.4	29.7	81.2	38.8	40.9	5.6	32.9	28.8	45.5
DAST [82]	AAAI'21	92.2	49.0	84.3	36.5	28.9	33.9	38.8	28.4	84.9	41.6	83.2	60.0	28.7	87.2	45.0	45.3	7.4	33.8	32.8	49.6
BiMaL [62]	ICCV'21	91.2	39.6	82.7	29.4	25.2	29.6	34.3	25.5	85.4	44.0	80.8	59.7	30.4	86.6	38.5	47.6	1.2	34.0	36.8	47.3
Ours	-	**93.7**	**60.4**	**86.5**	**41.1**	32.0	**37.3**	38.7	38.6	**87.2**	43.0	85.5	**65.4**	**35.1**	**88.3**	41.8	51.6	0.0	**38.0**	**52.1**	**53.5**

Table 2. Comparison results (mIoU) from SYNTHIA to Cityscapes.

Method	Venue	road	sidewalk	building	light	sign	vegetation	sky	person	rider	car	bus	motocycle	bike	**mIoU**
Source Only	-	36.3	14.6	68.8	5.6	9.1	69.0	79.4	52.5	11.3	49.8	9.5	11.0	20.7	29.5
BDL [37]	CVPR'19	86.0	46.7	80.3	14.1	11.6	79.2	81.3	54.1	27.9	73.7	42.2	25.7	45.3	51.4
IntraDA [52]	CVPR'20	84.3	37.7	79.5	9.2	8.4	80.0	84.1	57.2	23.0	78.0	38.1	20.3	36.5	48.9
LTIR [29]	CVPR'20	92.6	53.2	79.2	1.6	7.5	78.6	84.4	52.6	20.0	82.1	34.8	14.6	39.4	49.3
SIM [69]	CVPR'20	83.0	44.0	80.3	17.1	15.8	80.5	81.8	59.9	33.1	70.2	37.3	28.5	45.8	52.1
FDA [81]	CVPR'20	79.3	35.0	73.2	19.9	24.0	61.7	82.6	61.4	31.1	83.9	40.8	38.4	51.1	52.5
CrCDA [28]	ECCV'20	86.2	44.9	79.5	9.4	11.8	78.6	86.5	57.2	26.1	76.8	39.9	21.5	32.1	50.0
WLabel [53]	ECCV'20	92.0	53.5	80.9	3.8	6.0	81.6	84.4	60.8	24.4	80.5	39.0	26.0	41.7	51.9
CCM [33]	ECCV'20	79.6	36.4	80.6	22.4	14.9	81.8	77.4	56.8	25.9	80.7	45.3	29.9	52.0	52.9
LDR [77]	ECCV'20	85.1	44.5	81.0	16.4	15.2	80.1	84.8	59.4	31.9	73.2	41.0	32.6	44.7	53.1
CLAN [44]	TPAMI'21	82.7	37.2	81.5	17.1	13.1	81.2	83.3	55.5	22.1	76.6	30.1	23.5	30.7	48.8
ASA [90]	TIP'21	91.2	48.5	80.4	5.5	5.2	79.5	83.6	56.4	21.9	80.3	36.2	20.0	32.9	49.3
DAST [82]	AAAI'21	87.1	44.5	82.3	13.9	13.1	81.6	86.0	60.3	25.1	83.1	40.1	24.4	40.5	52.5
BiMaL [62]	ICCV'21	92.8	51.5	81.5	17.6	15.9	82.4	84.6	55.9	22.3	85.7	44.5	24.6	38.8	53.7
Ours	-	79.4	45.3	**83.3**	**24.7**	**29.6**	68.9	**87.5**	**63.1**	**33.8**	**87.0**	**51.0**	32.1	**52.1**	**56.8**

backbone network parameters and others, respectively. The temperature τ is set to 0.1, and the positive threshold δ is set to 0.75. Following [37,63,66,78,92], the source images are scaled to 1280×720 for GTAV and 1280×760 for SYNTHIA, and the target images are resized to 512×1024 in the training. Then, both the source images and target images are randomly cropped to 512×512. In addition, we also apply Color jittering and Gaussian blurring on the mixed images following [61]. The projector F_{proj} is constructed by two 1×1 consecutive convolutions (2048 hidden layer channels and 128 output channels) with one intermediate ReLU layer. $\lambda = 1$ and $\mu(t) = 0.01 * e^{\left(-5(1-t/t_{\max})^{0.5}\right)}$ for the loss function. We use $batchsize = 2$ for $250k$ iterations in all experiments.

4.2. Comparison to State-of-the-Art Methods

We compare our method with the state-of-the-art UDA methods on two common UDA benchmarks. We report the results of GTAV [56] $\rightarrow$ Cityscapes [14] in Table 1 and the results of SYNTHIA [57] $\rightarrow$ Cityscapes [14] in Table 2, respectively. We use the mean of class-wise Intersection-over-Union (mIoU) as the evaluation metric, and measure the performance on the validation set of Cityscapes.

As shown in these two tables, our proposed RCCR method outperforms the state-of-the-art approaches by $3\% \sim 6\%$ on two challenging tasks. It also surpasses the baseline ("Source Only") by around 20% and 27%, respectively. Specifically, we obtain a 53.5% mIoU on GTAV [56] and achieve the best per-class IoU performance for 11 classes among the total 19 classes. For SYNTHIA [57] dataset, we observe a 56.8% mIoU and get the best per-class IoU in 9 classes among the total 13 classes. These results reveal the effectiveness of our RCCR among different classes, *e.g.,* building, traffic light, traffic sign, person, rider, car, and etc. Most recent methods incorporate many sophisticated sub-components, *e.g.,* computationally involved adversarial

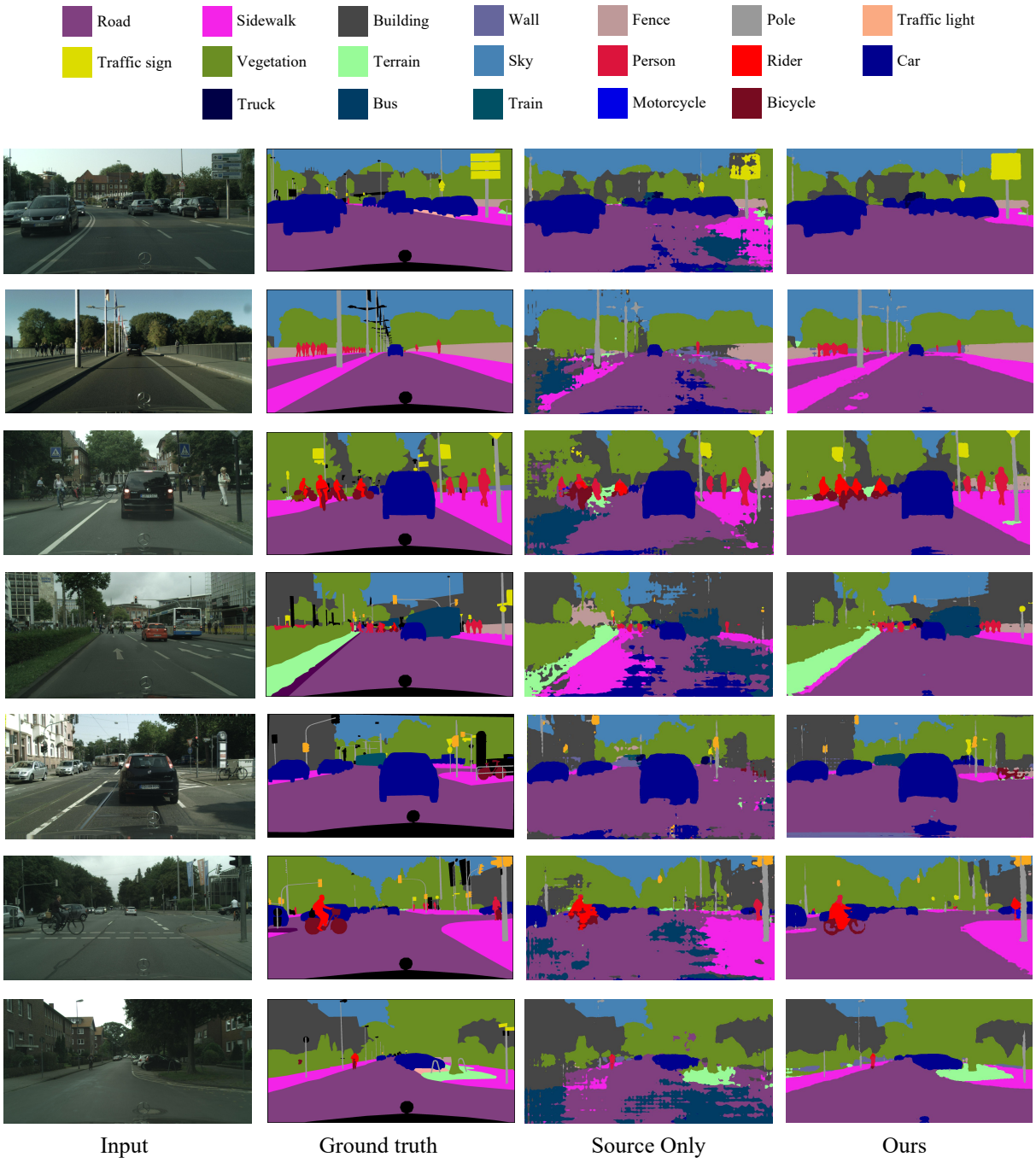

Figure 3. Qualitative segmentation results in the GTAV [56] to Cityscapes [14] benchmark. From left to right: target image, ground truth, source-only prediction, and predictions using our method.

training [44, 45, 52, 63, 69], image translation [27, 29, 37, 77, 81], offline self-training [37, 46, 48, 58, 91, 92], which are quite complex and hard to converge, and cannot be trained in an end-to-end manner. In contrast, our proposed method is simple yet effective, and fully end-to-end trainable.

Figure 3 visualizes some segmentation results in the GTAV $\rightarrow$ Cityscapes (19 classes) set-up. The four columns plot (a) RGB input images of the target domain, (b) ground truth of the target image, (c) the predictions of "Source Only" baseline, and (d) the predictions of our RCCR. As

Table 3. Ablations of each component in RCCR from SYNTHIA to Cityscapes. RWC: region-wise contrastive loss; NS(R): random sampling for negative samples; NS(C): category-wise sampling for negative samples; PS: positive sampling; MB: memory bank.

ID	RWC	NS(R)	NS(C)	PS	MB	mIoU
baseline						54.8
I	✓					55.3
II	✓	✓				55.6
III	✓	✓	✓			56.0
IV	✓	✓	✓	✓		56.3
V	✓	✓	✓	✓	✓	56.8

we can see from the figure, due to the lack of regional consistency, the baseline tends to incorrectly classify some categories *e.g.,* the road as sidewalk or terrain, and produces some false predictions on some sophisticated classes, *e.g.,* traffic sign. With the help of our proposed approach, we manage to produce correct predictions at a high level of confidence.

4.3. Ablation Study

In this section, we perform the ablation study to investigate each role of our RCCR components, including the region-wise contrastive (RWC) loss, random sampling (NS(R)) and category-wise sampling (NS(C)) for negative samples, sampling strategies for positive samples (PS), and memory bank (MB). As shown in Table 3, with our proposed RWC loss, we reach the state-of-the-art performance of 55.3% mIoU, showing the effectiveness of region-level alignment under different environments, and the carefully designed projector can actually extract useful feature embeddings while preserving the necessary information for segmentation tasks. When taking the different sampling strategies into account, we find the gradual and non-marginal improvements by $0.3\% \sim 0.4\%$, which reveals that combining the category information derived from the label or target segmentation output can lead to a more powerful contrastive learning scheme. By default, we take the memory bank to store the negative samples created from the last three batches, and this mechanism makes further improvement by 0.5 points.

5. Conclusion

In this paper, we proposed regional contrastive consistency regularization (RCCR) for domain adaptive semantic segmentation. By maximizing the inter-region differences and minimizing intra-region disagreements, we could effectively keep the regional consistency in a fine-grained manner, *i.e.*, feature space and label space, regardless of the changing of outdoor environments. Firstly, a region-wise contrastive (RWC) loss with two sampling strategies is proposed to realize efficient regional consistency. Then, we introduce momentum projector heads, where the teacher projector is the exponential moving average of the student. Besides, we design a memory bank mechanism to learn more robust and stable region-wise features under varying environments. Experimental results on the two challenging benchmark datasets show that our RCCR achieves the state-of-the-art UDA segmentation performance.

References

[1] Inigo Alonso, Alberto Sabater, David Ferstl, Luis Montesano, and Ana C Murillo. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. *arXiv preprint arXiv:2104.13415*, 2021. 3

[2] Nikita Araslanov and Stefan Roth. Self-supervised augmentation consistency for adapting semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15384–15394, 2021. 2

[3] Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. Exploring object relation in mean teacher for cross-domain detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11457–11466, 2019. 3

[4] Wei-Lun Chang, Hui-Po Wang, Wen-Hsiao Peng, and Wei-Chen Chiu. All about structure: Adapting structural information across domains for boosting semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1909, 2019. 2, 3

[5] Cong Chen, Shouyang Dong, Ye Tian, Kunlin Cao, Li Liu, and Yuanhao Guo. Temporal self-ensembling teacher for semi-supervised object detection. *IEEE Transactions on Multimedia*, pages 1–1, 2021. 3

[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 Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2018. 1, 6

[7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[9] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1841–1850, 2019. 2, 3

[10] Ying Chen, Xu Ouyang, Kaiyue Zhu, and Gady Agam. Complexmix: Semi-supervised semantic segmentation via mask-based data augmentation. In *2021 IEEE International*

Conference on Image Processing (ICIP), pages 2264–2268. IEEE, 2021. 3

[11] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. Crdoco: Pixel-level domain transfer with cross-domain consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1791–1800, 2019. 2, 3

[12] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 1992–2001, 2017. 2, 3

[13] Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6830–6840, 2019. 2, 3

[14] 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 *Proc. CVPR*, pages 3213–3223, 2016. 1, 2, 6, 7, 8

[15] Gabriela Csurka. Domain adaptation for visual applications: A comprehensive survey. *arXiv preprint arXiv:1702.05374*, 2017. 2

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[17] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4091–4101, June 2021. 3

[18] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 1

[19] Geoff French, Samuli Laine, Timo Aila, Samuli Laine, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. In *British Machine Vision Conference*, 2020. 3, 4

[20] Geoffrey French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. In *Proceedings of the International Conference on Learning Representations*, 2018. 2

[21] Geoff French, Avital Oliver, and Tim Salimans. Milking cowmask for semi-supervised image classification. *arXiv preprint arXiv:2003.12022*, 2020. 3

[22] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4340–4349, 2016. 1

[23] Yaroslav Ganin and Victor S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proc. ICML*, volume 37, pages 1180–1189, 2015. 2

[24] Qiqi Gu, Qianyu Zhou, Minghao Xu, Zhengyang Feng, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. Pit: Position-invariant transform for cross-fov domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8761–8770, 2021. 2, 3

[25] Shaohua Guo, Qianyu Zhou, Ye Zhou, Qiqi Gu, Junshu Tang, Zhengyang Feng, and Lizhuang Ma. Label-free regional consistency for image-to-image translation. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 2, 3

[26] 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. 6

[27] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998, 2018. 2, 3, 8

[28] Jiaxing Huang, Shijian Lu, Dayan Guan, and Xiaobing Zhang. Contextual-relation consistent domain adaptation for semantic segmentation. In *European conference on computer vision*, volume 12360, pages 705–722. Springer, 2020. 7

[29] Myeongjin Kim and Hyeran Byun. Learning texture invariant representation for domain adaptation of semantic segmentation. In *Proc. CVPR*, pages 12975–12984, 2020. 2, 3, 7, 8

[30] Wouter M Kouw and Marco Loog. A review of single-source unsupervised domain adaptation. *arXiv preprint arXiv:1901.05335*, 2019. 2

[31] Xin Lai, Zhuotao Tian, Li Jiang, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. Semi-supervised semantic segmentation with directional context-aware consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1205–1214, 2021. 3

[32] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *IEEE Proc.*, 86(11):2278–2324, 1998. 3

[33] Guangrui Li, Guoliang Kang, Wu Liu, Yunchao Wei, and Yi Yang. Content-consistent matching for domain adaptive semantic segmentation. In *European conference on computer vision*, volume 12359, pages 440–456. Springer, 2020. 7

[34] Kang Li, Shujun Wang, Lequan Yu, and Pheng-Ann Heng. Dual-teacher: Integrating intra-domain and inter-domain teachers for annotation-efficient cardiac segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 418–427. Springer, 2020. 3

[35] Shuang Li, Chi Harold Liu, Limin Su, Binhui Xie, Zhengming Ding, CL Philip Chen, and Dapeng Wu. Discriminative transfer feature and label consistency for cross-domain image classification. *IEEE Trans. Neural Netw. Learn. Sys.*, pages 1–15, 2020. 2

[36] Shuang Li, Shiji Song, Gao Huang, Zhengming Ding, and Cheng Wu. Domain invariant and class discriminative fea-

ture learning for visual domain adaptation. *IEEE Trans. Image Process.*, 27(9):4260–4273, 2018. 2

[37] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6936–6945, 2019. 2, 3, 6, 7, 8

[38] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[39] Chenxi Liu, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, Wei Hua, Alan L. Yuille, and Fei-Fei Li. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 82–92, 2019. 1

[40] Weizhe Liu, David Ferstl, Samuel Schulter, Lukas Zebedin, Pascal Fua, and Christian Leistner. Domain adaptation for semantic segmentation via patch-wise contrastive learning. *arXiv preprint arXiv:2104.11056*, 2021. 3

[41] 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

[42] Zhihe Lu, Yongxin Yang, Xiatian Zhu, Cong Liu, Yi-Zhe Song, and Tao Xiang. Stochastic classifiers for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9111–9120, 2020. 2, 3

[43] Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Significance-aware information bottleneck for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6778–6787, 2019. 2, 3

[44] Yawei Luo, Ping Liu, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Category-level adversarial adaptation for semantic segmentation using purified features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 2, 3, 7, 8

[45] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2507–2516, 2019. 2, 3, 6, 8

[46] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *European conference on computer vision*, pages 415–430. Springer, 2020. 2, 3, 8

[47] Luke Melas-Kyriazi and Arjun K Manrai. Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12435–12445, 2021. 2

[48] M. Naseer Subhani and Mohsen Ali. Learning from scale-invariant examples for domain adaptation in semantic segmentation. In *European conference on computer vision*, pages 290–306. Springer, 2020. 2, 3, 8

[49] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NeurIPS workshop*, 2011. 3

[50] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1369–1378, 2021. 3

[51] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[52] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *Unsupervised Intra-domain Adaptation for Semantic Segmentation through Self-Supervision*, pages 3764–3773, 2020. 2, 3, 7, 8

[53] Sujoy Paul, Yi-Hsuan Tsai, Samuel Schulter, Amit K. Roy-Chowdhury, and Manmohan Chandraker. Domain adaptive semantic segmentation using weak labels. In *European conference on computer vision*, volume 12354, pages 571–587. Springer, 2020. 7

[54] Christian S Perone, Pedro Ballester, Rodrigo C Barros, and Julien Cohen-Adad. Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. *NeuroImage*, 194:1–11, 2019. 2

[55] Rindra Ramamonjison, Amin Banitalebi-Dehkordi, Xinyu Kang, Xiaolong Bai, and Yong Zhang. Simrod: A simple adaptation method for robust object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision (ICCV)*, 2021. 3

[56] 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, 7, 8

[57] 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, 2, 6, 7

[58] Inkyu Shin, Sanghyun Woo, Fei Pan, and In So Kweon. Two-phase pseudo label densification for self-training based domain adaptation. In *European conference on computer vision*, pages 532–548. Springer, 2020. 2, 3, 8

[59] Shiyu Tang, Peijun Tang, Yanxiang Gong, Zheng Ma, and Mei Xie. Unsupervised domain adaptation via coarse-to-fine feature alignment method using contrastive learning. *arXiv preprint arXiv:2103.12371*, 2021. 3, 5

[60] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems 30*, pages 1195–1204, 2017. 3

[61] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1379–1389, 2021. 2, 3, 7

[62] Thanh-Dat Truong, Chi Nhan Duong, Ngan Le, Son Lam Phung, Chase Rainwater, and Khoa Luu. Bimal: Bijective maximum likelihood approach to domain adaptation in semantic scene segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 7

[63] 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. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7472–7481, 2018. 2, 3, 6, 7, 8

[64] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. Domain adaptation for structured output via discriminative patch representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1456–1465, 2019. 2, 3

[65] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. DADA: depth-aware domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7363–7372, 2019. 2, 3

[66] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019. 6, 7

[67] Haoran Wang, Tong Shen, Wei Zhang, Lingyu Duan, and Tao Mei. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. pages 642–659. Springer, 2020. 2, 3, 7

[68] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. *arXiv preprint arXiv:2101.11939*, 2021. 3, 5

[69] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-mei Hwu, Thomas S Huang, and Honghui Shi. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12635–12644, 2020. 2, 3, 6, 7, 8

[70] Garrett Wilson and Diane J Cook. A survey of unsupervised deep domain adaptation. *arXiv preprint arXiv:1812.02849*, 2018. 2

[71] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. 3

[72] Jinxi Xiang, Zhuowei Li, Wenji Wang, Qing Xia, and Shaoting Zhang. Self-ensembling contrastive learning for semi-supervised medical image segmentation. *arXiv preprint arXiv:2105.12924*, 2021. 3

[73] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16684–16693, 2021. 3

[74] Hongyi Xu, Fengqi Liu, Qianyu Zhou, Jinkun Hao, Zhijie Cao, Zhengyang Feng, and Lizhuang Ma. Semi-supervised 3d object detection via adaptive pseudo-labeling. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 3183–3187. IEEE, 2021. 3

[75] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. *arXiv preprint arXiv:2106.09018*, 2021. 3

[76] Yonghao Xu, Bo Du, Lefei Zhang, Qian Zhang, Guoli Wang, and Liangpei Zhang. Self-ensembling attention networks: Addressing domain shift for semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5581–5588, 2019. 2, 3

[77] Jinyu Yang, Weizhi An, Sheng Wang, Xinliang Zhu, Chaochao Yan, and Junzhou Huang. Label-driven reconstruction for domain adaptation in semantic segmentation. In *European conference on computer vision*, volume 12372, pages 480–498. Springer, 2020. 2, 3, 7, 8

[78] Jihan Yang, Ruijia Xu, Ruiyu Li, Xiaojuan Qi, Xiaoyong Shen, Guanbin Li, and Liang Lin. An adversarial perturbation oriented domain adaptation approach for semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12613–12620, 2020. 2, 3, 6, 7

[79] Qize Yang, Xihan Wei, Biao Wang, Xian-Sheng Hua, and Lei Zhang. Interactive self-training with mean teachers for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5941–5950, June 2021. 3

[80] Yanchao Yang, Dong Lao, Ganesh Sundaramoorthi, and Stefano Soatto. Phase consistent ecological domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9011–9020, 2020. 2, 3

[81] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020. 2, 3, 7, 8

[82] Fei Yu, Mo Zhang, Hexin Dong, Sheng Hu, Bin Dong, and Li Zhang. Dast: Unsupervised domain adaptation in semantic segmentation based on discriminator attention and self-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10754–10762, 2021. 7

[83] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 605–613. Springer, 2019. 3

[84] Jing Zhang, Wanqing Li, and Philip Ogunbona. Transfer learning for cross-dataset recognition: a survey. *arXiv preprint arXiv:1705.04396*, 2017. 2

[85] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 1

[86] Na Zhao, Tat-Seng Chua, and Gim Hee Lee. Sess: Self-ensembling semi-supervised 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[87] Qianyu Zhou, Zhengyang Feng, Qiqi Gu, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. Uncertainty-aware consistency regularization for cross-domain semantic segmentation. *arXiv preprint arXiv:2004.08878*, 2020. 2, 3

[88] Qianyu Zhou, Zhengyang Feng, Qiqi Gu, Jiangmiao Pang, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. Context-aware mixup for domain adaptive semantic segmentation. *arXiv preprint arXiv:2108.03557*, 2021. 2, 3

[89] Qianyu Zhou, Qiqi Gu, Jiangmiao Pang, Zhengyang Feng, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. Self-adversarial disentangling for specific domain adaptation. *arXiv preprint arXiv:2108.03553*, 2021. 3

[90] Wei Zhou, Yukang Wang, Jiajia Chu, Jiehua Yang, Xiang Bai, and Yongchao Xu. Affinity space adaptation for semantic segmentation across domains. *IEEE Transactions on Image Processing*, 30:2549–2561, 2020. 7

[91] Yang Zou, Zhiding Yu, BVK 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. 2, 3, 8

[92] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5982–5991, 2019. 2, 3, 6, 7, 8