Learning Pseudo-Relations for Cross-domain Semantic Segmentation

Dong Zhao, Shuang Wang✉ , Qi Zang, Dou Quan, Xiutiao Ye, Rui Yang, Licheng Jiao
School of Artificial Intelligence, Xidian University, Shaanxi, China
zhaodong01@stu.xidian.edu.cn; shwang@mail.xidian.edu.cn

Abstract

*Domain adaptive semantic segmentation aims to adapt a model trained on labeled source domain to unlabeled target domain. Self-training shows competitive potential in this field. Existing methods along this stream mainly focus on selecting reliable predictions on target data as pseudo-labels for category learning, while ignoring the useful relations between pixels for relation learning. In this paper, we propose a pseudo-relation learning framework, **R**elation **Tea**cher (RTea), which can exploitable pixel relations to efficiently use unreliable pixels and learn generalized representations. In this framework, we build reasonable pseudo-relations on local grids and fuse them with low-level relations in the image space, which are motivated by the **reliable local relations prior** and **available low-level relations prior**. Then, we design a pseudo-relation learning strategy and optimize the class probability to meet the relation consistency by finding the optimal sub-graph division. In this way, the model's certainty and consistency of prediction are enhanced on the target domain, and the cross-domain inadaptation is further eliminated. Extensive experiments on three datasets demonstrate the effectiveness of the proposed method. The code will be available at https://github.com/DZhaoXd/RTea.*

1. Introduction

Semantic segmentation is a challenging problem of assigning each pixel a class label in an image. Driven by deep neural networks, significant progress has been made in this field. Despite these efforts, a segmentation model trained with a specific domain does not generalize well to other domains. It is known to be caused by the domain gap between the training (source) and testing (target) domains [13]. To solve this problem, unsupervised domain adaptation (UDA) is proposed to improve the segmentation model's adaptability to the target domain.

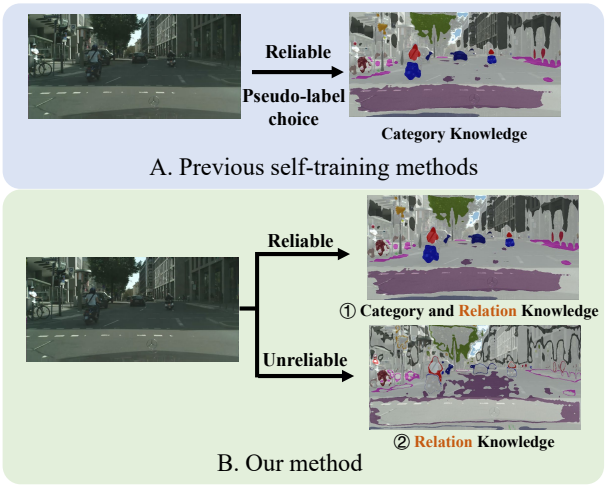

Figure 1: Overview of our motivation. In A, previous self-training methods select reliable pseudo-labels for category learning. In B, our method reasonably utilizes reliable pseudo-labels and unreliable ones for relation learning.

Domain alignment is one of the mainstream UDA semantic segmentation methods, aiming to align the distribution of source and target domains in input [30, 60, 25, 4], feature [53, 32, 19, 22, 35, 65, 50, 56], or output spaces [48, 33, 49]. Works along this line achieve positive adaptation benefits but the lack of specific target domain knowledge leads to slight improvement [39, 68, 66].

To this end, self-training methods [27, 10, 1, 69, 64, 67] are proposed to mine target-specific knowledge. These methods use the pseudo-labels generated by the pre-adapted model to further train the model on the target domain. Consequently, the quality of pseudo-labels for training directly determines the performance of self-training. Following this key point, reliability measure-based and uncertainty estimation-based self-training methods are proposed [71, 27, 69, 10, 1, 64, 52]. These methods reduce the noise interference of pseudo-labels for category learning, bring-

This work is supported by the National Key R&D Program of China under Grant No. 2021ZD0110400, the National Natural Science Foundation of China(No.62271377, No.62201407), the Key Research and Development Program of Shannxi (No.2021ZDLGY01-06, No.2022ZDLGY01-12), the China Postdoctoral Science Foundation (No. 2022M722496), the Foreign Scholars in University Research and Teaching Program's 111 Project (B07048).

ing considerable performance improvement.

In this paper, we explore the potential of self-training from another perspective, as shown in Fig. 1. In Fig. 1 A, previous self-training methods perform category learning on reliable pseudo-labels and discard unreliable ones. However, we find that not only category learning but relation learning can be performed in pseudo-labels to further improve the adaptability of the model. In Fig. 1 B, we argue that relation learning in pseudo-labels can be performed in two ways: ① The relations between reliable pseudo-labels can be additionally used for representation learning to build a more generalized representation space. ② By establishing the relations between reliable and unreliable pseudo-labels, discarded pixels can also be effectively used for self-training to increase the certainty of the model. In this way, the available knowledge contained in pseudo-labels can be fully exploited, both reliable and unreliable pixels.

To achieve the above goals, building reasonable relations between pixels is the core. Dense pixel relations can be represented by a relation matrix (or affinity matrix) [8, 20, 29], modeling the similarity between pixels on an image. Due to massive noise contained in the pseudo-labels, the relation matrix constructed by them also contain noisy relations. We explore two observational priors to guide the building of relations, as shown in Fig. 2. We use the predictions of the unadapted model for the target images to observe noise distribution of relations. Comparing Fig. 2 (e) and (f), we observe that high-level relations built by pseudo-labels are noisy in long-distance association but are reliable within local areas, which are termed as ***reliable local relations prior***. We analyze this because the insufficiently adapted model cannot transfer global semantics and can only give reasonable relations in local regions. Besides, in Fig. 2(d), we further explore the low-level relations built on each local grid in image space using Gaussian kernel. We find that the low-level relations in local grids can capture the boundaries of objects and contain exploitable relations, which are termed as ***available low-level relations prior***. We argue that such relations, although lacking in semantics, provides class boundary clues can be reasonably exploited.

With these aspects in mind, we propose a pseudo-relation learning-based self-training framework, **R**elation **Tea**cher (RTea), forcing the student model to learn the pseudo-relations between pixels from the teacher model and achieve co-evolution for both models. In this framework, with the guidance of the above two priors, we first use pseudo-labels to build high-level relations in each divided grid, which avoids being misled by long-distance relationships. Then, we fuse low-level relations in image space into high-level relations to attenuate noisy relations and assist semantic relations in identifying category boundaries. Next, we explore the way of learning pseudo-relations and devise a novel pseudo-relations loss, which optimizes the class

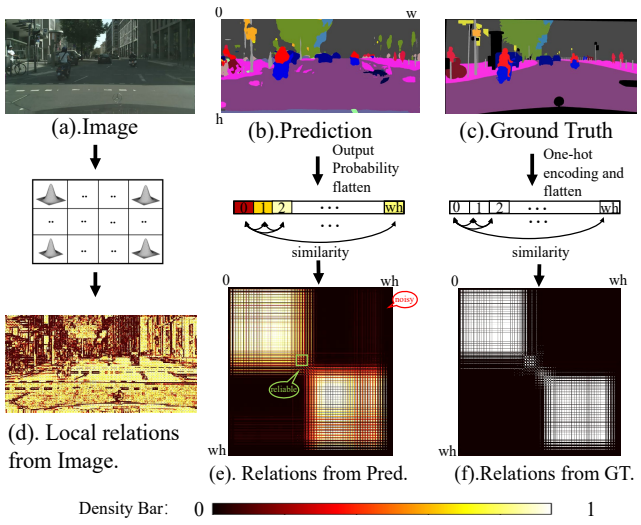

Figure 2: Overview of two observational priors. (b) is the source-only model's prediction, (d) is the local relation map in image space built by Gaussian kernel in each grid, (e) is the pseudo-relation matrix. (f) is the relation matrix built by ground truth.

probability to meet the relation consistency by finding the optimal sub-graph division from the global pseudo-relation of each class. It has two advantages over the naive relation-learning loss: one is global relational modeling, which can be easily implemented by matrix multiplication; the other is threshold-free learning by dynamically weighting class probabilities and pseudo-relations.

With RTea, the model's certainty and consistency of prediction can be enhanced on the target domain, and the cross-domain inadaptation is further eliminated. Sufficient experiments show our method can further mine the available knowledge in pseudo-labels, and it can be easily incorporated into existing self-training method to further boost their performance.

2. Related work

Unsupervised Domain Adaptation (UDA). UDA methods can be divided into domain alignment and self-training. Alignment-based methods narrow the domain gap by aligning distributions at different levels, *e.g.*, input [12, 30, 25, 60], feature [32, 33, 7, 50, 56, 22] and output [48, 49, 59]. Although they achieve positive adaptation benefits, the lack of specific target domain knowledge leads to slight improvement. To this end, self-training methods are proposed to train the network with pseudo-labels for the target domain, which can be divided into two categories, offline and online pseudo-label-based methods. Offline self-training (OFFST) methods saves the target domain pseudo-labels generated by the pre-trained model and use them to train the model iteratively [27, 10, 37]. To avoid noise interfer-

ence, these methods select high-quality pseudo-labels via threshold setting [70, 71], consistent prediction [68], and pseudo-label prototypes [64]. These methods require additional storage of pseudo-labels and an iterative training, which is not conducive to practical use. Online self-training (ONST) uses the model's output during training as supervision without saving intermediate results [58, 4, 51, 1]. ONST avoids the inconvenience of multiple training rounds and manual intervention between consecutive rounds in OFFST. Our work is designed in online self-training fashion. Besides, our method further mines the available knowledge in pseudo-labels from the novel perspective of pixel relations, which can further promote pseudo-label learning and boosting self-training performance.

Relation Learning (RL) in Semantic Segmentation. In supervised semantic segmentation, capturing relations between pixels is continuously studied at feature [8, 20, 29] and output layers[24, 17]. At the feature layer, several work [8, 20, 62, 29] use the label-guided pixel relation to build a compact feature space, in which features of the same class are close and of different classes are far. At the output layer, AAF [24] and CDGC [17] exploit pixel relations in the ground truth to correct the output probabilities, forcing the model to adjust both the classifier and the feature extractor to output reasonable relations and semantic structure. These supervised RL methods achieve huge performance improvements, emphasizing the importance of maintaining the relation between pixels. However, in the UDA task, the lack of annotations makes it impossible to build relations as these works. Our method proposes a novel pseudo-relation building and learning strategy so that the above advantages can be also achieved to the UDA segmentation task.

Pseudo Label Correction. As a classic technique of semi-supervised learning, pseudo-labeling has shown favorable competitive advantages in many visual tasks with limited labels. Due to the limitation of confidence bias and error accumulation [41], pseudo-labeling are easy to overfit the noise and lead to model divergence[61]. Some pioneering works alleviate the problems from pseudo-label selection[63], negative label learning[41], contrastive learning [55], model calibration [23, 54]. However, the task-specific and domain-dependent design limits the application of these methods in complex cross-domain tasks. In contrast, our work proposes a novel perspective to rectify pseudo-labels and a tailored solution for cross-domain segmentation tasks.

3. Methodology

3.1. Background and Overview

This paper focuses on the unsupervised domain adaptation (UDA) semantic segmentation, where source domain data $\mathcal{X}_s = \{x_s\}$ with pixel-level labels $\mathcal{Y}_s = \{y_s\}$ and unlabeled target domain data $\mathcal{X}_t = \{x_t\}$ are given. Our goal is to train a segmentation model G that can work well on target domain. G consists of a feature extractor F and a classifier C. Given sampled image $x \in \mathcal{X}_s \cup \mathcal{X}_t$, F mapping x to the feature space $f = F(x)$, and C categorizes each feature in f to obtain a class probability map $p = C(F(x)) \in \mathbb{R}^{h \times w \times K}$. For the source domain, the cross entropy loss L_s is calculated to optimize the model,

$$L_s = -\frac{1}{|\mathcal{X}_s|} \sum_{x_s \in \mathcal{X}_s} \log C(F(x_s))(y_s). \tag{1}$$

For the target domain, to narrow the domain gap, self-training methods adopt pseudo-labels $\hat{y}_t$ generated by the pre-adapted model to retrain the unadapted model,

$$L_{st} = -\frac{1}{|\mathcal{X}_t|} \sum_{x_t \in \mathcal{X}_t} \log C(F(x_t))(\hat{y}_t). \tag{2}$$

In this way, the model's adaptability can be enhanced by relearning the knowledge of the pseudo-labels in the target domain. However, using Eq. 2 tends to interfere with training because the pseudo-labels contain massive noise. Thus, some works select high-confidence pseudo-labels by setting threshold ζ^k for each class,

$$\hat{y}_t = \begin{cases} \arg\max_k p_t^k\,, & if\ \max(p_t^k) > \zeta^k \\ ignore\,, & otherwise, \end{cases} \tag{3}$$

where p_t^k is the k-th class probability score. In Eq. 3, the ζ^k determines the quality of the selected pixels, which needs to be dynamically adjusted according to the adaptation degree. Thus, in the subsequent works, scholars [58, 4, 51, 1] mainly focus on setting reasonable thresholds or uncertainty estimation strategies to select reliable pixels.

In this paper, we explore the potential of self-training from another perspective. Different from mining category knowledge in pseudo-labels, we propose to mine relation knowledge in pseudo-relations between pixels. The overview of our method is shown in Fig 3. For the source domain, images are input to the student model, which is optimized by the L_s. For the target domain, the original image and its data-transformed version are input into the student model and teacher model, respectively. After that, the output of the teacher model is used to build pseudo-relations S_{pr} to further supervise the student model of category learning by L_{st} and relation learning by L_{pr}. We detail the pseudo-relation S_{pr} building in Sec.3.2, and detail the pseudo-relation learning by optimizing L_{pr} in Sec.3.3.

3.2. Pseudo-relation Building

In the supervised segmentation task, relations can be driven by ground truth, which is not achievable for the UDA segmentation. Thus, how to build the relations between

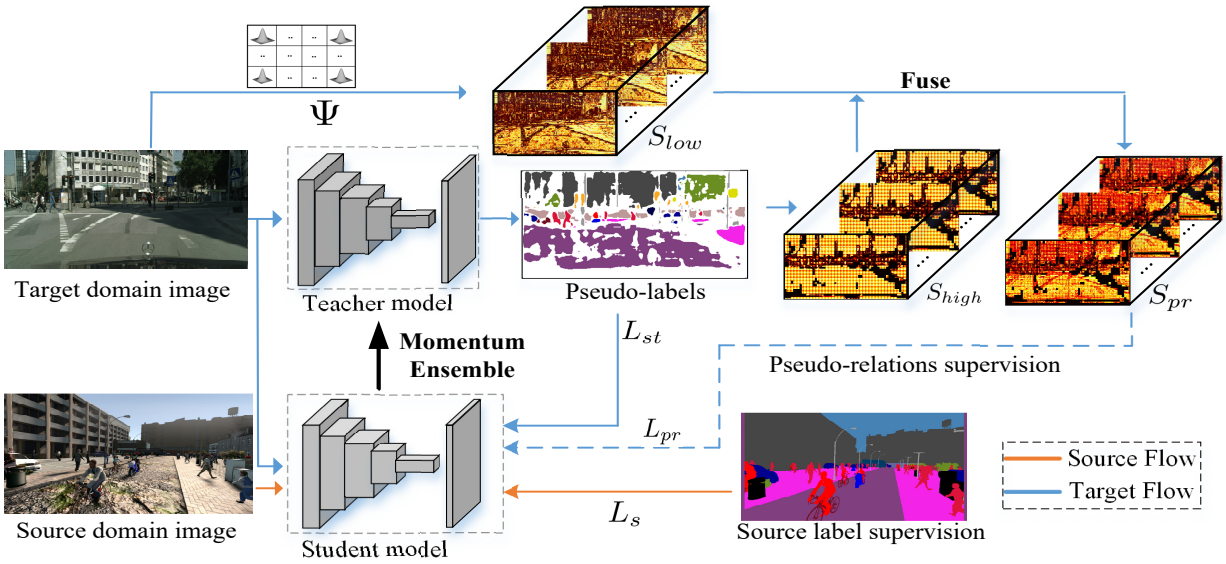

Figure 3: The overall pipeline of the proposed **R**elation **Tea**cher that includes a student model and a teacher model. S_{low}, S_{high} and S_{pr} represent low-level relation matrix, high-level relation matrix and pseudo-relation matrix. L_{st} is the traditional loss function for pseudo-label learning. L_{pr} is the proposed loss function for pseudo-relation learning.

pixels in the target domain is the key. A naive solution is to build by measuring the class similarity between pseudo-labels from a pre-adapted model. Given any two pixels $p_{(t,i)}$ and $p_{(t,j)}$ $(i, j = 0, 1, 2, ..., w \times h)$ on the class probability map $p_t \in \mathbb{R}^{h \times w \times K}$ of the target domain, the high-level relation between them can be defined as,

$$s_{(i,j)} = \mathrm{SIM}(p_{(t,i)}, p_{(t,j)}), \tag{4}$$

where $\mathrm{SIM}(\cdot, \cdot)$ is the cosine similarity measure. Extending this relation to all pixels in an image, we can derive a high-level relation matrix $S = \{s(i, j)\} \in \mathbb{R}^{hw \times hw}$. However, such a relation modeling mechanism will introduce massive noise due to the unadapted model. As mentioned in *reliable local relations prior*, these noises are mainly in the long-distance relations but relations in local grids are relatively accurate.

To this end, we devise a local relation modeling mechanism. We divide the target image into $N \times N$ grids $\{G_l\}_{l=1}^{hw/N^2}$ and model the relations inside the grid. Then the high-level relation matrix for each grid can be obtained,

$$S_{high} = \{s_{(m,n)}\}_{(m,n) \in G_l} \in \mathbb{R}^{N^2 \times N^2}. \tag{5}$$

Nonetheless, local pseudo-relations may still contain noise due to domain shifts.

To further mitigate unreliable relations in local grids, we introduce low-level relation constraints in each grid. The motivation stems from *available low-level relations prior* that local low-level relations (RGB space) can be well used to capture the edges and internal structures of objects. Drawing on the traditional potential energy function constructing the potential energy field [43, 44, 38], we use the Gaussian function to calculate the energy on the RGBXY[43] space as low-level relation constraints,

$$\begin{aligned} S_{low} &= \{\Psi_{(m,n)}\}_{(m,n) \in G_l} \in \mathbb{R}^{N^2 \times N^2}, \\ \Psi(m, n) &= exp\left(- \left| \frac{I_m - I_n}{\sigma} \right|^2 \right), \end{aligned} \tag{6}$$

where $I_{(\cdot)}$ is the feature vector on RGBXY space, σ is the bandwidth parameter of the Gaussian function. Due to the nature of the Gaussian function[16], the value range of the S_{low} are $[0, 1]$, which is the same as that of S_{high}. A higher value indicates a stronger association between pixels. Although it lacks high-level semantics, it can better regularize relations (See Tab. 4 for verification). We linearly combine the low- and high-level relations as final pseudo-relations,

$$S_{pr} = \alpha S_{low} + (1 - \alpha) S_{high}, \tag{7}$$

where the α is a hyper-parameter and is simply set to 0.5.

3.3. Pseudo-relation Learning

Following the paradigm of pseudo-label learning, we can exploit the reliable relations in the pseudo-relation matrix to perform pseudo-relation learning. Specifically, for any pixel pair $p_{(t,i)}$ and $p_{(t,j)}$ in l-th grid, we can perform a *p*airwise

*r*elation loss [24, 17] in a pseudo manner as follows,

$$\hat{L}_{pr} = \sum_{l} \sum_{i,j} \mathrm{KL}(p_{(t,i)} || p_{(t,j)}) \cdot \mathbb{1}(S_{pr;l}(i,j) > M_{up}) + (1 - \mathrm{KL}(p_{(t,i)} || p_{(t,j)})) \cdot \mathbb{1}(S_{pr;l}(i,j) < M_{low}). \tag{8}$$

$\mathrm{KL}(\cdot||\cdot)$ is the Kullback-Leibler divergence between two distributions. $\mathbb{1}(\cdot)$ is the indicator function for thresholding. M_{up} and M_{low} are upper and lower thresholds for filtering unreliable relations. Optimizing this loss pulls the class distribution of the pixel pairs with a pseudo-relation higher than M_{up} closer and pushes those with a pseudo-relation lower than M_{low} further away. In this way, pseudo-relation learning can assist pseudo-label learning (in Eq.2) in the following aspects. Let the pixels used by Eq.2 be denoted as reliable pixels, and the rest are unreliable ones. 1). Relations between reliable pixels. Those reliable inter- and intra-class relations are informative and can be incorporated into learning as structural inference. 2). Relations between reliable and unreliable pixels. Pushing or Pulling the class probabilities of the unreliable pixels towards reliable ones will greatly enhance the confidence of the model. 3). Relations between unreliable pixels. Although we cannot give these pixel pairs explicit class targets, implicit class cues may help the model to enhance confidence. See further instructions in Appendix A.

In practice, in $\hat{L}_{pr}$, the pair of class distribution input to $\mathrm{KL}(\cdot||\cdot)$ should assign the more sharp one as the target. This makes it require pair-by-pair index calculation, which greatly increases the training time. Besides, M_{up} and M_{low} requires hyper-parameter search under different adaptation tasks, hindering the flexibility of application. Here, we learn from the idea of graph cut [57] and devise a new pseudo-relation loss, which can be performed by matrix multiplication without explicit thresholding.

Specifically, in l-th grid, we construct an undirected weighted graph $\mathcal{G}$, treating all pixels in this grid as nodes of the graph and the similarity between these pixels as the weights of edges. Thus, the corresponding $S_{pr;l}$ can be seen as a pseudo-adjacency matrix of the $\mathcal{G}$. With the $\mathcal{G}$, the K classification problem is regarded as a graph cutting problem of cutting a graph into K subgraphs. Then, the k-th class probability $p^k_{t;l}$ of this grid is regarded as the probability of the corresponding nodes cut into k-th subgraph ($k \in [0, K-1]$). According to the properties of the graph, we can calculate the soft cost of cutting the k-th subgraph as follows,

$$cut^k = \hat{p}^k_{t;l} S_{pr;l} \mathbf{1} - \hat{p}^k_{t;l} S_{pr;l} (\hat{p}^k_{t;l})^T. \tag{9}$$

$\hat{p}^k_{t;l} \in \mathbb{R}^{1 \times N^2}$ is the vector flatten by $p^k_{t;l} \in \mathbb{R}^{1 \times N \times N}$, $\mathbf{1} \in \mathbb{R}^{N^2 \times 1}$ is an all-ones vector. We show an example to illustrate Eq.9 in Fig.4. The smaller the cut value, the smaller

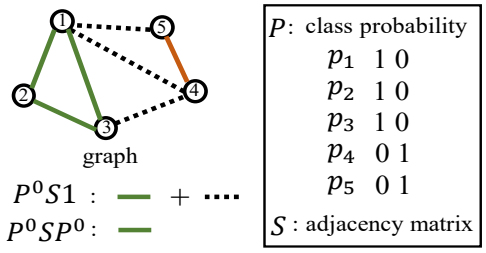

Figure 4: We show an example of binary classification to illustrate Eq.9. Given five samples and their class probability P, take the cut of class 0 as an example. $P^0S\mathbf{1}$ is the sum of all edges connected to the nodes of class 0, i.e., the green and dashed lines. P^0SP^0 is the sum of all edges connecting two nodes of class 0, i.e., the green lines, and $P^0S\mathbf{1} - P^0SP^0$ is the cost consumed of the 0 class. To minimize the cost, nodes with high affinity should be classified into the same class, and vice versa.

the correlation between the subgraphs and the larger the correlation within the subgraphs. Thus, cut^k can be regarded as a proxy, when the cut^k value is small, the class probability $p^k_{t;l}$ better conforms to the pseudo-relations $S_{pr;l}$, and vice versa. Moreover, the Eq.9 is derivable with respect to $p^k_{t;l}$, and its gradient w.r.t. p^k is,

$$\frac{\partial cut^k}{\partial p^k} = S_{pr}(\mathbf{1} - 2p^k). \tag{10}$$

The proxy's gradient is proportional to the S_{pr} and $\mathbf{1} - 2p^k$, which implicitly weight high-confidence relations and reliable pixels (class probability far from 0.5). This explains that unreliable edges (relations) on the graph play a minor role in pulling or pushing away the class distribution, enabling adaptive weighting without threshold. Consequently, we modify Eq.9 as a threshold-free pseudo-relation loss function to optimize p_t as follows,

$$L_{pr} = \sum_{l}^{hw/N^2} \sum_{k}^{K} \hat{p}^k_{t;l} S_{pr} (\mathbf{1} - \hat{p}^k_{t;l})^T. \tag{11}$$

L_{pr} can be calculated using matrix multiplication in a simple and quick way.

3.4. Relation Teacher

We embed the learning of pseudo-relation into the optimization of mean-teacher [45] framework to realize online self-training as shown in Fig 3. The final optimization objective is,

$$L_f = L_s + \lambda_{st} L_{st} + \lambda_{pr} L_{pr}. \tag{12}$$

where λ_{pr}, λ_{st} are the trade-off coefficients. λ_{st} is set as 0.001 following SAC [1]. λ_{pr} is empirically set to 0.01 to balance the loss value.

Method	road	sidewalk	Building	Wall	fence	pole	light	sign	vege.	terrain	sky	person	rider	car	truck	bus	train	mbike	bike	mIoU
FADA (ECCV 2020) [50]	91.0	50.6	86.0	43.4	29.8	36.8	43.4	25.0	86.8	38.3	87.4	64.0	38.0	85.2	31.6	46.1	6.5	25.4	37.1	50.1
FDA (CVPR 2020) [60]	92.5	53.3	82.3	26.5	27.6	36.4	40.5	38.8	82.2	39.8	78.0	62.6	34.4	84.9	34.1	53.1	16.8	27.7	46.4	50.4
IAST (ECCV 2020) [37]	93.8	57.8	85.1	39.5	26.7	26.2	43.1	34.7	84.9	32.9	88.0	62.6	29.0	87.3	39.2	49.6	23.2	34.7	39.6	51.5
MetaCorr (CVPR 2021) [10]	92.8	58.1	86.2	39.7	33.1	36.3	42.0	38.6	85.5	37.8	87.6	62.8	31.7	84.8	35.7	50.3	2.0	36.8	48.0	52.1
RPT (CVPR 2021) [66]	89.2	43.3	86.1	39.5	29.9	40.2	49.6	33.1	87.4	38.5	86.0	64.4	25.1	88.5	36.6	45.8	23.9	36.5	56.8	52.6
SAC (CVPR 2021) [1]	90.4	53.9	86.6	42.4	27.3	45.1	48.5	42.7	87.4	40.1	86.1	67.5	29.7	88.5	49.1	54.6	9.8	26.6	45.3	53.8
CFDAN (CVPR 2021) [36]	92.5	58.3	86.5	27.4	28.8	38.1	46.7	42.5	85.4	38.4	**91.8**	66.4	37.0	87.8	40.7	52.4	44.6	41.7	59.0	56.1
SDFA (ICCV 2021) [26]	94.8	59.4	86.2	40.5	29.5	25.5	43.8	34.7	85.9	34.9	89.5	63.4	30.8	88.3	42.6	50.7	25.3	35.7	40.9	52.8
ProDA (CVPR 2021) [64]	87.8	56	79.7	**46.3**	44.8	45.6	53.5	53.5	88.6	45.2	82.1	70.7	39.2	88.8	45.5	59.4	1.0	48.9	56.4	57.5
SimT (CVPR2022) [9]	94.2	60	88.5	30.3	39.7	41.2	47.8	**60.8**	88.6	47.3	89.3	71.5	**45.0**	90.7	54.2	60.2	0.0	51.8	58.4	58.9
CPST (CVPR 2022) [28]	92.3	59.9	84.9	45.7	29.7	**52.8**	**61.5**	59.5	87.9	41.5	85.0	73.0	35.5	90.4	48.7	**73.9**	26.3	**53.8**	53.9	60.8
Undoing (CVPR 2022) [31]	92.9	52.7	87.2	39.4	41.3	43.9	55	52.9	89.3	48.2	91.2	71.4	36	90.2	**67.9**	59.8	0	48.5	59.3	59.3
DAP (CVPR 2022) [21]	94.5	63.1	**89.1**	29.8	**47.5**	50.4	56.7	58.7	89.5	**50.2**	87	**73.6**	38.6	**91.3**	50.2	52.9	0	50.2	**63.5**	59.8
CaCo (CVPR 2022) [18]	93.8	64.1	85.7	43.7	42.2	46.1	50.1	54.0	88.7	47	86.5	68.1	2.9	88.0	43.4	60.1	31.5	46.1	60.9	58.0
ADPL (TPAMI 2023) [3]	93.4	60.6	87.5	45.3	32.6	37.3	43.3	55.5	87.2	44.8	88	64.5	34.2	88.3	52.6	61.8	**49.8**	41.8	59.4	59.4
RTea (Ours)	**95.4**	**67.1**	87.9	46.1	44.0	46.0	53.8	59.5	**89.7**	49.8	89.8	71.5	40.5	90.8	55.0	57.9	22.1	47.7	62.5	**61.9**
Daformer (CVPR 2022) [14]	95.7	70.2	89.4	53.5	48.1	49.6	55.8	59.4	89.9	47.9	92.5	72.2	44.7	92.3	74.5	78.2	65.1	55.9	61.8	68.3
Daformer + RTea (Ours)	96.1	71.7	89.1	57.8	50.4	55.9	59.3	66.7	90.4	48.2	94.5	74.8	46.5	93.8	78.7	81.6	65.8	57.1	62.8	70.6
HRDA (ECCV 2022) [15]	96.4	74.4	91.0	61.6	51.5	57.1	63.9	69.3	**91.3**	48.4	94.2	79.0	52.9	93.9	84.1	85.7	**75.9**	**63.9**	**67.5**	73.8
HRDA + RTea (Ours)	**97.1**	**75.2**	**92.6**	**63.5**	**51.8**	**58.2**	**66.5**	**71.2**	91.1	**49.0**	**96.8**	**81.5**	**54.2**	**94.2**	**84.8**	**86.6**	75.7	62.2	66.7	**74.7**

Table 1: Experimental results for GTA5 → Cityscapes adaptation task. The best results in every column are **highlighted**.

Method	road	sidewalk	Building	Wall*	fence*	pole*	light	sign	vege.	sky	person	rider	car	bus	mbike	bike	mIoU	mIoU*
FADA (ECCV 2020)[50]	84.5	40.1	83.1	4.8	0	34.3	20.1	27.2	84.8	84	53.5	22.6	85.4	43.7	26.8	27.8	45.2	52.5
FDA (CVPR 2020) [60]	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
IAST (ECCV 2020) [37]	81.9	41.5	83.3	17.7	4.6	32.3	30.9	28.8	83.4	85.0	65.5	30.8	86.5	38.2	33.1	52.7	49.8	57.0
MetaCorr (CVPR 2021) [10]	92.6	52.7	81.3	8.9	2.4	28.1	13.0	7.3	83.5	85.0	60.1	19.7	84.8	37.2	21.5	43.9	45.1	52.5
RPT (CVPR 2021) [66]	88.9	46.5	84.5	15.1	0.5	38.5	39.5	30.1	85.9	85.8	59.8	26.1	88.1	46.8	27.7	56.1	51.2	58.9
SAC (CVPR 2021) [1]	89.3	47.2	85.5	26.5	1.3	43.0	45.5	32.0	87.1	89.3	63.6	25.4	86.9	35.6	30.4	53.0	52.6	59.3
CFDAN (CVPR 2021) [36]	75.7	30.0	81.9	11.5	2.5	35.3	18.0	32.7	86.2	90.1	65.1	33.2	83.3	36.5	35.3	54.3	48.2	55.5
SFDA (ICCV 2021) [26]	90.5	50.0	81.6	13.3	2.8	34.7	25.7	33.1	83.8	89.2	66.0	34.9	85.3	53.4	46.1	46.6	52.0	60.1
ProDA (CVPR 2021) [64]	87.8	45.7	84.6	**37.1**	0.6	44.0	54.6	37.0	88.1	84.4	74.2	24.3	88.2	51.1	40.5	45.6	55.5	62.0
CPST (CVPR 2022) [28]	87.2	43.9	85.5	33.6	0.3	**47.7**	**57.4**	37.2	87.8	88.5	**79.0**	32.0	**90.6**	49.4	**50.8**	59.8	57.9	65.3
Undoing (CVPR 2022) [31]	82.5	37.2	81.1	23.8	0	45.7	57.2	**47.6**	87.7	85.8	74.1	28.6	88.4	**66.0**	47.0	55.3	56.7	64.5
CaCo (CVPR 2022) [18]	87.4	48.9	79.6	8.8	0.2	30.1	17.4	28.3	79.9	81.2	56.3	24.2	78.6	39.2	28.1	48.3	46.0	53.6
ADPL (TPAMI 2023) [3]	86.1	38.6	85.9	29.7	1.3	36.6	41.3	47.2	85	90.4	67.5	**44.3**	87.4	57.1	43.9	51.4	55.9	63.6
RTea (ours)	**93.2**	**59.6**	**86.3**	31.3	**4.8**	43.1	41.8	44.0	**88.6**	**90.5**	70.4	42.6	89.5	56.7	40.2	**59.9**	**58.9**	**66.4**
Daformer (CVPR 2022) [14]	84.5	40.7	88.4	41.5	**6.5**	50.0	55	54.6	86	89.8	73.2	48.2	87.2	53.2	53.9	61.7	67.4	60.9
Daformer + RTea (Ours)	85.9	43.2	90.1	45.1	6.3	52.4	60.5	57.1	87.8	92.2	75.3	51.8	87.4	55.9	54.1	62.6	69.5	63.0
HRDA (ECCV 2022) [15]	85.2	47.7	88.8	49.5	4.8	57.2	65.7	60.9	85.3	92.9	79.4	52.8	89	64.7	**63.9**	**64.9**	72.4	65.8
HRDA + RTea (Ours)	**87.8**	**49.0**	**90.3**	**50.3**	5.5	**58.6**	**66.0**	**61.4**	**86.8**	**93.1**	**79.5**	**53.1**	**89.5**	**65.1**	63.7	64.6	**73.0**	**66.5**

Table 2: Experimental results for SYNTHIA → Cityscapes adaptation task. The best results in every column are **highlighted**. The mIoU and mIoU* are averaged over 16 and 13 categories, respectively.

4. Experiment

4.1. Datasets and Experimental Setup

Datasets. We use one real dataset (Cityscapes [5]) and two synthetic datasets (GTA5 [40] and SYNTHIA [42]). The Cityscapes dataset contains 2,975 training images and 500 validation images of resolution 2048×1024. The GTA5 dataset contains 24,966 images with resolution 1914×1052 and has 19 common categories with Cityscapes. The SYNTHIA dataset contains 9,400 images with resolution 1280×760 and has 16 common categories with Cityscapes.

Implementation Details. We adopt Deeplab-v2 [2] as the base network, ResNet-101 [11] as the feature extractor and the aspp [2] module as the classifier. The network is pre-trained on ImageNet. The optimizer is SGD with the momentum of 0.9 and weight decay of 10^{-4}. The initial learning rate is set to 2.5×10^{-4}, and then is reduced following a poly policy with a power of 0.9. The batch size is set as 4. The final λ_f and λ_o values are set to 0.025 and 0.005, respectively. We apply the L^{pr} to the model's output class probability, which is 8 times smaller than the original resolution. And the grid size N is set to 8. The weighting factor α is set to 0.5. The σ for XY and RGB space in Gaussian function is set to 6 and 0.1, following [43]. Data augmentation strategies performed on the data input to the student model include random flipping, Gaussian noise, color transformation, cutout[6], and contrast enhancement, a similar and common operation in online self-

training methods [58, 4, 51, 1]. Besides, we also preform the class-mixing resampling strategy [47] on the source domain to focus on the minority class. When performing L_{st} for pseudo-label learning, we adopt the thresholding selection method in SAC [1] to pick high-quality samples. After training with Rtea, we retrained the model using the distillation strategy [64] for better adaptation, which is commonly used in recently published UDA work [31, 21, 28, 3]. The detailed scores for each stage are in the Tab. 3. Our network is trained with four RTX3090 GPUs on PyTorch.

4.2. Comparisons with State-of-the-Arts

GTA5 $\rightarrow$ Cityscapes. We report the comparison results with existing methods on GTA5 $\rightarrow$ Cityscapes task in Tab. 1. We compared two structures based on resnet-101 and transformer structures. Overall, our RTea achieves new state-of-the-art performance than related works, and the category performance scores are also highly competitive, demonstrating the effectiveness of RTea. Compared with the domain alignment method, the performance of RTea show an advantage over the state-of-the-art method FDA [60] by 11.5%. Compared with the offline self-training method, the mIoU score of RTea is 10.4% and 9.8% higher than IAST [37] and RPT[66]. Compared with the online self-training method, RTea exceeds SAC [1] and ProDA [64] by 8.1% and 4.4%. Compared with the newly released method Undoing[31], DAP[21] and ADPL[3], our RTea still outperforms these methods by more than 2.0% mIoU scores, showing the potential of our method even more. Compared with the method of transformer structure, our method improves the mIoU score by 0.7% and 0.9% on Daformer [14] and HRDA [15] respectively, which shows that RTea has good scalability and transferability. Qualitative results on both tasks can be found in Appendix B.

SYNTHIA $\rightarrow$ Cityscapes. The results of using SYNTHIA as the source domain are reported in Tab. 2, including mIoU/mIoU* covering 16/13 classes. On the whole, our method still achieves significant improvements, showing gains over advanced methods. In particular, SYNTHIA and Cityscapes suffer from significant visual domain differences in 'road', and 'sidewalk', which leads to the poor performance of the most UDA methods in these categories. Our method uses the pseudo-relations of these categories and achieves higher performance improvements, and we argue this is due to the local pseudo-relation learning to better capture the structure of these categories. Compared with the domain alignment and offline self-training methods, RTea maintains similar performance gains to the GTA5 transfer task. Compared with online self-training method, we achieve better performance than CPST [28] over 16 and 13 classes, exceeding its by 1.0% and 1.1% mIoU score. Compared with the adaptation method using transformer structure, our method improves the mIoU score by 2.1% and 0.7% on Daformer [14] and HRDA [15], which further verifies the effectiveness of the method.

L_s	L_{st}	*copy-paste*	$\hat{L}_{pr}$	L_{pr}	L^s_{pr}	$Dist$	GTA5	SYNTHIA
✓	✓						53.5	51.2
✓	✓	✓					55.8	53.9
✓	✓	✓	✓				58.9	56.3
✓	✓	✓		✓			59.6	56.8
✓	✓	✓		✓	✓		59.6	56.9
✓	✓	✓		✓	✓	✓	61.9	58.9

Table 3: Ablation experiments of each module in GTA5 and SYNTHIA $\rightarrow$ Cityscapes adaptation task. The basic self-training (L_s + L_{st}) is following SAC[1]. The *copy-paste* means that we adopt the resampling strategy in [47] for minority learning. Here, we report a result of $\hat{L}_{pr}$ with the fine-tuned thresholds by grid search. L^s_{pr} denotes that relation learning is performed on the source domain. $Dist$ denote the distillation strategy in [50, 64]. All numbers are mIoU($\%$) score.

4.3. Ablation Studies

Ablation for Each Module. Tab.3 reports the ablation results for different modules. We adopt SAC [1] with the copy-paste [47] argumentation as the baseline, achieving the 55.8% and 53.9% mIoU score on two adaptation tasks. Overall, Rtea achieves 3.8% and 3.0% performance improvements on the two tasks on this competitive baseline, respectively, showing the effectiveness of the pseudo-relation learning. Specifically, the naive $\hat{L}_{pr}$ loss, although, can improve performance, the screening of the threshold makes it difficult to be directly applied to practice, because the threshold may be very sensitive in different adaptation scenarios. The proposed L_{pr} loss does not require explicit threshold while improving the performance, showing its flexibility and effectiveness. In addition, we also tried to perform pseudo-relation learning on the source domain L^s_{pr}, and we did not find a significant performance improvement, which may be because it cannot directly benefit the target domain. After training with RTea, the self-distillation method can still be used to further improve the adaptability of the model in the target domain.

Ablation for Pseudo-Relation Building. Tab.4 shows the ablation results for building pseudo-relation. When the relations are built on RGBXY space (S_{low}), the performance achieves a good improvement, 2.0% and 2.1% on two tasks, respectively, which suggests that low-level relations on local grids can propagate effective category knowledge. When the relations are built on pseudo labels (S_{high}), the performance is only improved by 1.0% and 0.7% on two tasks, respectively. This is because it is not easy to find valuable relation pairs in this way, and their effects may overlap with label learning. Specifically, unreliable pixels discarded by pseudo-label learning are still difficult to use, and high

S_{low}	S_{high}	S_{pr}	GTA5	SYNTHIA
			55.8	53.9
✓			57.8	55.8
	✓		56.8	54.6
		✓	**59.6**	**56.8**

Table 4: Ablation experiments on the pseudo-relation building. S_{low}, S_{high} and S_{pr} are low-level, high-level and final pseudo relation matrix.

$R-R$	$R-U$	$U-U$	GTA5	SYNTHIA
			55.8	53.9
	✓	✓	58.2	55.2
✓		✓	57.2	54.9
✓	✓		59.3	56.6
✓	✓	✓	**59.6**	**56.8**

Table 5: Ablation experiments on the pseudo-relation learning. $R-R$ denotes pairs with two reliable pixels. $R-U$ denotes pairs with one reliable and one unreliable pixel. $U-U$ denotes pairs with two unreliable pixels.

confidence relationships still contain massive noise, When S_{low} and S_{high} are combined, the best effect is achieved. We argue that the advantage is that, the noisy relations in the high level are corrected and the lack of semantics of the low-level relations is also made up, and more unreliable pixels are better exploited.

Ablation for Pseudo-Relation Learning. To verify the effect of different pixel pairs on pseudo-relation learning, we devise the ablation study in Tab. 5. We follow 1), 2) and 3) explained in the Eq. 8, and divide them into R-R (two reliable pixels), R-U (one reliable and one unreliable pixel) and U-U (two unreliable pixels). During training, we clear all gradients for one of the pairs to verify its effect on adaptation. We find that R-U pairs play the most important role in adaptation, R-R pairs come second, and U-U pairs come last. We analyze learning R-U pairs can make full use of valuable pixels from unexploited target domains, which is the most direct to enhance the target domain adaptability.

4.4. Discussion

Sensitivity for reliable sample selection. In this section, we perform two experiments to explore how sensitive RTea is to reliable sample selection. First, we report the results when the parameters of the heuristic threshold (Eq. 3) are varied. The thresholding strategy consists of two hyperparameters, namely the upper threshold θ and the decay rate β. Table 6 shows that our method maintain a stable performance improvement when thresholding parameters vary within a certain range. Second, we report Rtea's performance on different reliable sample selection methods in Table 7. It verifies the effectiveness of RTea, which improves the performance of different self-training methods.

	β=0.0001	β=0.001	β=0.01
θ = 0.75	59.0	59.2	58.8
θ = 0.80	59.2	59.6	59.1

Table 6: The mIoU scores (%) on GTA5 → Cityscapes task with varying thresholding parameters in Eq. 3.

	PD	CD	AD	HT
without Rtea	55.0	54.4	53.2	55.8
with Rtea	58.8	56.8	54.8	59.6

Table 7: The mIoU scores (%) on GTA5 → Cityscapes task with different reliable sample selection strategies. PD denotes feature-prototype distance [64], CD denotes classifier discrepancy [69], AD denotes adversarial difficulty [34] and HT is our used heuristic threshold [1].

How local grids affect each relation pair. We perform experiments on GTA5 → Cityscapes to verify which relation pairs are most affected by local grids. Pixels are still classified as reliable (R) and unreliable (U) according to whether they are used for category learning in traditional self-training, see Tabel 8. In the first line, we find that (1) the global grid is good for $R-R$ because it extends the spatial scope of relational learning, while it does harm to $R-U$ and $U-U$ due to the long-range relation noise. As a result, performing pseudo-relation learning on global region lead model degradation. In the second line, we argue that (2) performing pseudo-relation on the local grids slightly reduces the gain of $R-R$, but greatly reduces the interference of noise on $R-U$ and $U-U$. Finally, the model can obtain better performance improvement.

	Baseline	$R-R$	$R-U$	$U-U$
Global	55.8	57.2 (+1.4)	51.4 (-4.4)	52.5 (-3.3)
Local grid	55.8	56.6 (+0.8)	57.9 (+2.1)	55.9(+0.1)

Table 8: The mIoU (%) of local grids effect on GTA5 → Cityscapes adaptation task.

Computational overhead. Performing relation learning on dense prediction tasks such as semantic segmentation is time-consuming. Therefore, it is necessary to discuss the computational overhead introduced in Rtea. Overall, we think the pseudo-relation computational overhead is acceptable. First, for complexity, our computational cost is $\mathcal{O}(N^2WH)$ (H, W: width and height of feature map, N: grid size, $N \ll W, H$), which is greatly reduced compared to the global similarity $\mathcal{O}(W^2H^2)$. Second, the pseudo-relation can be computed in parallel using matrix operations in PyTorch. On a Nivida RTX-3090, it only takes almost 40 ms for each batch and only increases about 3% computational overhead in each iteration. Third, the cost is only for training and not for inference.

Align low- and high-level relation matrix with the same Gaussian function. In the method section, we use the cosine distance (Eq. 5) and Gaussian distance (Eq. 6) to model the similarity of high- and low-level, respectively. It is interesting to explore unifying the modeling of high- and low-level pseudo-relations into the same metric function, e.g. Gaussian distance. With the same metric function, for any local grid G_l, the local relation can be built as, $s(m,n) = \sum_{m,n \in G_l} \Psi_{\sigma_{rgbxy}}(m,n) \cdot \Psi_{\sigma_{prob}}(m,n)$, where σ_{rgbxy} and σ_{prob} are the bandwidth of the low- and high-level Gaussian function. This formula is similar to the kernels of bilateral filtering[46], inspiring us to understand the pseudo-relation learning from the filtering perspective, i.e. smoothing and correcting the noise on output probability. Table 9 presents the comparison results, showing unifying the same metric function can maintain performance with less computational overhead.

Metric	GTA5	SYN.	Time	Metric	GTA5	SYN.	Time
Current	59.6	56.9	40ms	Same Gaussian	59.6	57.1	29ms

Table 9: The mIoU (%) and computational cost (time/batch) using different similarity metric functions on GTA5 $\rightarrow$ Cityscapes adaptation task.

Hyper-parameters Impacts. We analyze the sensitivity of the hyper-parameters α (trade-off coefficient), grid size N, the results are presented in the Appendix C.

4.5. Visualization

Visualization of pixel relations. Fig. 5 visualizes the learned relations on GTA5 $\rightarrow$ Cityscapes task. Comparing Fig. 5 a , b and e, it shows that the feature associations of the baseline are chaotic while our method builds more accurate feature relations. Comparing Fig. 5 c, d and e, it indicts that our method captures category relations more accurately than baseline, resulting in more structured outputs.

Comparison of Pixel Utilization and Model Confidence. Fig. 6 compares the correct pixel utilization (PU) and average confidence (AC) of the model before and after adding RTea. With RTea, the PU is significantly improved and the AC also shows better results. It illustrates RTea can exploit more uncertain pixels around high-confidence pixels and propagate relation information to them, thereby fully improving the certainty of the model.

Visualization of Pseudo-Labels. Fig. 7 visualizes the pseudo-labels of the baseline and our RTea to demonstrate the benefits of pseudo-relation learning. It shows that the RTea makes better use of pseudo-labels than the baseline model. For the areas are hard to distinguish (framed by the red box), RTea can enlarge the reliable pseudo-label area by capturing relations between pixels. Moreover, RTea provides clear class boundaries for the classes such as roads, sidewalks and buildings, which is more helpful for pseudo-label learning.

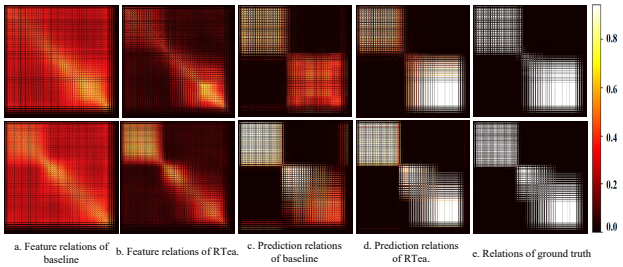

Figure 5: Visualization of the learned relations of RTea model on the GTA5 $\rightarrow$ Cityscapes task.

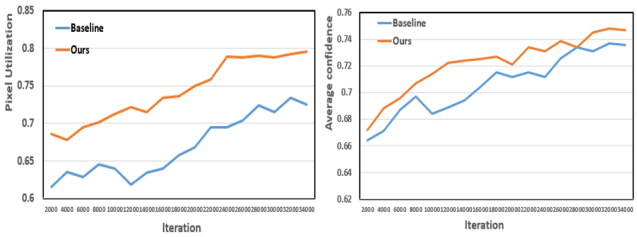

Figure 6: Comparison of the average correct pixel utilization and confidence of the target domain.

a. Baseline **b. Ours**

Figure 7: Comparison of pseudo-labels of baseline and RTea. The white masked area is unreliable pseudo-labels.

5. Conclusion

In this paper, we propose pseudo-relation learning framework for UDA semantic segmentation. In RTea, we provide two prior guidelines for pseudo-relation building, which may help more works exploiting pseudo-relations. Moreover, we explore how to use pseudo-relations from the constraints and do detailed analysis and experiments on the proposed solution. Sufficient experiments on two datasets demonstrate the effectiveness of the proposed method. In general, RTea provides a new idea for self-training methods and may inspire more works in this field.

References

[1] 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, June 2021. 1, 3, 5, 6, 7, 8

[2] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018. 6

[3] Yiting Cheng, Fangyun Wei, Jianmin Bao, Dong Chen, and Wenqiang Zhang. Adpl: Adaptive dual path learning for domain adaptation of semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 6, 7

[4] 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 Conference on International Conference on Computer Vision*, pages 6830–6840, 2019. 1, 3, 7

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, June 2016. 6

[6] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 6

[7] Liang Du, Jingang Tan, Hongye Yang, Jianfeng Feng, Xiangyang Xue, Qibao Zheng, Xiaoqing Ye, and Xiaolin Zhang. Ssf-dan: Separated semantic feature based domain adaptation network for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on International Conference on Computer Vision*, pages 982–991, October 2019. 2

[8] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019. 2, 3

[9] Xiaoqing Guo, Jie Liu, Tongliang Liu, and Yixuan Yuan. Simt: Handling open-set noise for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7032–7041, 2022. 6

[10] Xiaoqing Guo, Chen Yang, Baopu Li, and Yixuan Yuan. Metacorrection: Domain-aware meta loss correction for unsupervised domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3927–3936, June 2021. 1, 2, 6

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, June 2016. 6

[12] 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 Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1989–1998, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. 2

[13] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *CoRR*, abs/1612.02649, 2016. 1

[14] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9924–9935, June 2022. 6, 7

[15] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 372–391, Cham, 2022. Springer Nature Switzerland. 6, 7

[16] Cheng-Chun Hsu, Kuang-Jui Hsu, Chung-Chi Tsai, Yen-Yu Lin, and Yung-Yu Chuang. Weakly supervised instance segmentation using the bounding box tightness prior. In *Advances in Neural Information Processing Systems*, volume 32, 2019. 4

[17] Hanzhe Hu, Deyi Ji, Weihao Gan, Shuai Bai, Wei Wu, and Junjie Yan. Class-wise dynamic graph convolution for semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 1–17. Springer, 2020. 3, 5

[18] Jiaxing Huang, Dayan Guan, Aoran Xiao, Shijian Lu, and Ling Shao. Category contrast for unsupervised domain adaptation in visual tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1203–1214, June 2022. 6

[19] Jiaxing Huang, Shijian Lu, Dayan Guan, and Xiaobing Zhang. Contextual-relation consistent domain adaptation for semantic segmentation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Proceedings of the European Conference on Computer Vision*, pages 705–722, 2020. 1

[20] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on International Conference on Computer Vision*, pages 603–612, 2019. 2, 3

[21] Xinyue Huo, Lingxi Xie, Hengtong Hu, Wengang Zhou, Houqiang Li, and Qi Tian. Domain-agnostic prior for transfer semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7075–7085, 2022. 6, 7

[22] Guoliang Kang, Yunchao Wei, Yi Yang, Yueting Zhuang, and Alexander Hauptmann. Pixel-level cycle association: A new perspective for domain adaptive semantic segmentation. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and

H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3569–3580. Curran Associates, Inc., 2020. 1, 2

[23] Nazmul Karim, Mamshad Nayeem Rizve, Nazanin Rahnavard, Ajmal Mian, and Mubarak Shah. Unicon: Combating label noise through uniform selection and contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9676–9686, June 2022. 3

[24] Tsung-Wei Ke, Jyh-Jing Hwang, Ziwei Liu, and Stella X Yu. Adaptive affinity fields for semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 587–602, 2018. 3, 5

[25] Myeongjin Kim and Hyeran Byun. Learning texture invariant representation for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12975–12984, June 2020. 1, 2

[26] Jogendra Nath Kundu, Akshay Kulkarni, Amit Singh, Varun Jampani, and R Venkatesh Babu. Generalize then adapt: Source-free domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on International Conference on Computer Vision*, pages 7046–7056, 2021. 6

[27] Guangrui Li, Guoliang Kang, Wu Liu, Yunchao Wei, and Yi Yang. Content-consistent matching for domain adaptive semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 440–456. Springer, 2020. 1, 2

[28] Ruihuang Li, Shuai Li, Chenhang He, Yabin Zhang, Xu Jia, and Lei Zhang. Class-balanced pixel-level self-labeling for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11593–11603, 2022. 6, 7

[29] Xia Li, Yibo Yang, Qijie Zhao, Tiancheng Shen, Zhouchen Lin, and Hong Liu. Spatial pyramid based graph reasoning for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8950–8959, 2020. 2, 3

[30] 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, June 2019. 1, 2

[31] Yahao Liu, Jinhong Deng, Jiale Tao, Tong Chu, Lixin Duan, and Wen Li. Undoing the damage of label shift for cross-domain semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2022. 6, 7

[32] 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 Conference on International Conference on Computer Vision*, pages 6778–6787, October 2019. 1, 2

[33] Yawei Luo, L. Zheng, T. Guan, Junqing Yu, and Y. 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 2502–2511, 2019. 1, 2

[34] 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. 8

[35] Fengmao Lv, Tao Liang, Xiang Chen, and Guosheng Lin. Cross-domain semantic segmentation via domain-invariant interactive relation transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4334–4343, June 2020. 1

[36] Haoyu Ma, Xiangru Lin, Zifeng Wu, and Yizhou Yu. Coarse-to-fine domain adaptive semantic segmentation with photometric alignment and category-center regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4051–4060, June 2021. 6

[37] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision*, pages 415–430, 2020. 2, 6, 7

[38] Anton Obukhov, Stamatios Georgoulis, Dengxin Dai, and Luc Van Gool. Gated crf loss for weakly supervised semantic image segmentation. *arXiv preprint arXiv:1906.04651*, 2019. 4

[39] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3764–3773, June 2020. 1

[40] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Proceedings of the European Conference on Computer Vision*, pages 102–118. Springer International Publishing, 2016. 6

[41] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*, 2021. 3

[42] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 3234–3243, June 2016. 6

[43] Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers. Normalized cut loss for weakly-supervised cnn segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1818–1827, June 2018. 4, 6

[44] Meng Tang, Federico Perazzi, Abdelaziz Djelouah, Ismail Ben Ayed, Christopher Schroers, and Yuri Boykov. On regularized losses for weakly-supervised cnn segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 507–522, 2018. 4

[45] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets im-

prove semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*, 2017. 5

[46] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*, pages 839–846. IEEE, 1998. 9

[47] 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. 7

[48] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 7472–7481, June 2018. 1, 2

[49] 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 Conference on International Conference on Computer Vision*, pages 1456–1465, October 2019. 1, 2

[50] Haoran Wang, Tong Shen, Wei Zhang, Ling-Yu Duan, and Tao Mei. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Proceedings of the European Conference on Computer Vision*, pages 642–659, 2020. 1, 2, 6, 7

[51] Kaihong Wang, Chenhongyi Yang, and Margrit Betke. Consistency regularization with high-dimensional nonadversarial source-guided perturbation for unsupervised domain adaptation in segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10138–10146, 2021. 3, 7

[52] Shuang Wang, Qi Zang, Dong Zhao, Chaowei Fang, Dou Quan, Yutong Wan, Yanhe Guo, and Licheng Jiao. Select, purify, and exchange: A multisource unsupervised domain adaptation method for building extraction. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15, 2023. 1

[53] Shuang Wang, Dong Zhao, Chi Zhang, Yuwei Guo, Qi Zang, Yu Gu, Yi Li, and Licheng Jiao. Cluster alignment with target knowledge mining for unsupervised domain adaptation semantic segmentation. *IEEE Transactions on Image Processing*, 31:7403–7418, 2022. 1

[54] Xiao Wang, Hongrui Liu, Chuan Shi, and Cheng Yang. Be confident! towards trustworthy graph neural networks via confidence calibration. *Advances in Neural Information Processing Systems*, 34:23768–23779, 2021. 3

[55] Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4248–4257, 2022. 3

[56] 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, June 2020. 1, 2

[57] Z. Wu and R. Leahy. An optimal graph theoretic approach to data clustering: theory and its application to image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11):1101–1113, 1993. 5

[58] 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. 3, 7

[59] Jinyu Yang, Weizhi An, Sheng Wang, Xinliang Zhu, Chaochao Yan, and Junzhou Huang. Label-driven reconstruction for domain adaptation in semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 480–498. Springer, 2020. 2

[60] 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, June 2020. 1, 2, 6, 7

[61] Kun Yi and Jianxin Wu. Probabilistic end-to-end noise correction for learning with noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7017–7025, 2019. 3

[62] Changqian Yu, Jingbo Wang, Changxin Gao, Gang Yu, Chunhua Shen, and Nong Sang. Context prior for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12416–12425, 2020. 3

[63] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34:18408–18419, 2021. 3

[64] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12414–12424, 2021. 1, 3, 6, 7, 8

[65] Qiming Zhang, Jing Zhang, Wei Liu, and Dacheng Tao. Category anchor-guided unsupervised domain adaptation for semantic segmentation. In *Advances in Neural Information Processing Systems*, volume 32, 2019. 1

[66] Yiheng Zhang, Zhaofan Qiu, Ting Yao, Chong-Wah Ngo, Dong Liu, and Tao Mei. Transferring and regularizing prediction for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9621–9630, June 2020. 1, 6, 7

[67] Dong Zhao, Shuang Wang, Qi Zang, Dou Quan, Xiutiao Ye, and Licheng Jiao. Towards better stability and adaptability: Improve online self-training for model adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11733–11743, 2023. 1

[68] Zhedong Zheng and Yi Yang. Unsupervised scene adaptation with memory regularization in vivo. In *International Joint Conference on Artificial Intelligence*, 2020. 1, 3

[69] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision*, 129(4):1106–1120, 2021. 1, 8

[70] 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*, pages 289–305, 2018. 3

[71] Yang Zou, Zhiding Yu, Xiaofeng Liu, B.V.K. Vijaya Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE/CVF Conference on International Conference on Computer Vision*, pages 5982–5991, October 2019. 1, 3