SimT: Handling Open-set Noise for Domain Adaptive Semantic Segmentation

Xiaoqing Guo\textsuperscript{1} \quad Jie Liu\textsuperscript{1} \quad Tongliang Liu\textsuperscript{2} \quad Yixuan Yuan\textsuperscript{1}\textsuperscript{*}
\textsuperscript{1}City University of Hong Kong \quad \textsuperscript{2}University of Sydney
\{xqguo.ee, jliu.ee\}@my.cityu.edu.hk \quad tongliang.liu@sydney.edu.au \quad yxyuan.ee@cityu.edu.hk

\textbf{Abstract}

This paper studies a practical domain adaptive (DA) semantic segmentation problem where only pseudo-labeled target data is accessible through a black-box model. Due to the domain gap and label shift between two domains, pseudo-labeled target data contains mixed closed-set and open-set label noises. In this paper, we propose a simplex noise transition matrix (SimT) to model the mixed noise distributions in DA semantic segmentation and formulate the problem as estimation of SimT. By exploiting computational geometry analysis and properties of segmentation, we design three complementary regularizers, i.e., volume regularization, anchor guidance, convex guarantee, to approximate the true SimT. Specifically, volume regularization minimizes the volume of simplex formed by rows of the nonsquare SimT, which ensures outputs of segmentation model to fit into the ground truth label distribution. To compensate for the lack of open-set knowledge, anchor guidance and convex guarantee are devised to facilitate the modeling of open-set noise distribution and enhance the discriminative feature learning among closed-set and open-set classes. The estimated SimT is further utilized to correct noise issues in pseudo labels and promote the generalizability of segmentation model on target domain data. Extensive experimental results demonstrate that the proposed SimT can be flexibly plugged into existing DA methods to boost the performance. The source code is available at \url{https://github.com/CityU-AIM-Group/SimT}.

1. Introduction

DA semantic segmentation aims to adapt a segmentation model for the target domain through transferring knowledge from a source domain [22]. Considering the inefficient transmission and privacy issues in source domain knowledge, effective adaptation with only pseudo-labeled target data obtained through a black-box model is practical in real-world deployments [19]. This paper mainly focuses on this realistic scenario, and the challenge of this DA semantic segmentation problem lies in noisy pseudo labels generated for target domain data. To mitigate noisy adaptation, some methods leverage confidence score [13, 25, 50] or uncertainty [17, 46] to select reliable pseudo labels. For example, prior work [50] defines a threshold to eliminate low-confidence pseudo-labeled pixels. However, simply dropping confusing pixels with highly potential noise labels might lead to learn from a biased data distribution [48]. In fact, confusing pixels are more prone to be mislabeled and critical for accurate predictions, hence need to be carefully addressed. Instead of discarding unreliable pseudo-labeled target data, other methods denoise pseudo labels based on prototypes [42] or correct supervision signal through noise transition matrix (NTM) [11]. Nevertheless, these approaches heavily rely on the assumption of a shared label set between two domains and only consider the closed-set noisy pseudo labels. In real-world scenarios, the label space of target domain may not be consistent with that of source domain, and open-set class instances are commonly encountered in target domain. Hence, enhancing the segmentation performance on target domain needs to carefully address confusing pixels and open-set label noises.

With the aforementioned aims, we model the closed-set and open-set noise distributions in target pseudo labels through a SimT, which is a class-dependent and instance-independent transition matrix\textsuperscript{4}. Then the modeled noise distribution is leveraged to rectify supervision signals (i.e., loss) derived from noisy labels, thus all target samples are fully utilized. Therefore, the problem of alleviating closed-set and open-set noise issues can be treated as the problem of estimating SimT. To give an intuitive explanation of SimT, we interpret SimT through a geometric analysis on a toy example in Figure 1. In this toy example, SimT is defined as a matrix $T \in [0, 1]^{5 \times 3}$, as illustrated in Figure 1 (a). Each row, \textit{i.e.} $T_{c,.}$, indicates the probability of label $c$ flipping to three classes, which can be regarded as a three-dimensional point within the blue triangle of Figure 1 (b) since coordinates of this point add up to 1. The first three rows $[T_{1,.}, T_{2,.}, T_{3,.}]$ of SimT model closed-set noise distribution.

\textsuperscript{4}The complex instance-dependent label noise in real-world can be well approximated by class-dependent label noise when the noise rate is low [4].

\textsuperscript{*}Yixuan Yuan is the corresponding author.

This work was supported by Hong Kong Research Grants Council (RGC) General Research Fund 11211221(CityU 9043152).
which can detect open-set noises from pseudo-labeled target data and approximate the true SimT.

To this end, we propose three regularizers, i.e., volume regularization, anchor guidance and convex guarantee, for the estimation of SimT. Specifically, volume regularization for non-square SimT estimation is designed to approximate the true SimT, which is the one with minimum volume \( \text{Sim}(T) \) among all simplices enclosing \( p(y \mid x) \) (e.g., \( \text{Sim}(T'') \)) in Figure 1 (c). The minimum volume of SimT ensures outputs of segmentation model to fit into clean class posteriors. To prevent the trivial solution of open-set part in SimT, we devise anchor guidance regularization. In particular, we first define anchor point as the most representative pixel of each class with the largest clean class posterior probability as in Figure 1 (c). \( \text{Sim}(T'') \) formed by all anchor points is an approximation to the true SimT, hence we leverage anchor points to guide the estimation of SimT. To further facilitate the detection of anchor points in open-set classes, an auxiliary loss enhancing the discriminative feature learning among closed-set and open-set data is designed. Moreover, the trivial solution of open-set part may lead to non-convex \( \text{Sim}(T'') \). Hence, we advance convex guarantee to preserve the convex property of SimT, and the convex SimT can implicitly push the learned feature distribution of open-set classes away from closed-set classes.

We summarize our contributions in three aspects.

- We present a general and practical DA semantic segmentation framework to robustly learn from noisy pseudo-labeled target data, including closed-set and open-set label noises. To our best knowledge, it represents the first attempt to address the problem of open-set noisy label in DA.
- We, for the first time, model the closed-set and open-set noise distributions of target pseudo labels through SimT, and utilize computational geometry analysis to estimate it. The optimized SimT is used to benefit loss correction for noisy pseudo-labeled target data.
- Extensive experiments are conducted to verify the effect of the proposed SimT and demonstrate that it can be plugged into existing DA methods, such as UDA or SFDA, to boost the performance.

2. Related Work

2.1. Domain Adaptive Semantic Segmentation

DA semantic segmentation aims to adapt a model for the target domain through transferring knowledge from a source domain. Unsupervised domain adaptation (UDA) methods leverage adversarial learning to align data distributions of two domains [5,14,34]. Source-free domain adaptation (SFDA) methods [15,23,28,45] mine useful knowledge embedded in source model to perform DA without source data. These DA works require the coexistence of source data (or the detailed specification of source models) and tar-
get data since concurrent access can help to characterize the domain gap [15]. However, it is not always practical in real-world deployment scenarios, where the source domain data transmission may be inefficient when it is extremely large [45], data sharing may be restrained due to privacy [28] or the specific knowledge of source models may be not accessible [19]. Different from UDA and SFDA, we study a more general and practical DA problem where only pseudo-labeled target data is accessible through a black-box model.


However, almost all self-training based methods ignore the open-set pseudo label noises in target domain, which are commonly encountered in practice [20]. Although confidence or uncertainty based self-training methods omit potentially noisy data and may reject some open-set pseudo label noises, simply dropping pixels will lead to a biased data distribution learning especially when pixels lie around the decision boundary. In this paper, we model the closed-set and open-set noise distributions of pseudo labels in target domain with a SimT, and optimize the segmentation model with corrected supervision signals on all target data.

2.2. Deep Learning with Noisy Labels

Types of noisy labels can be divided into two categories: Closed-set noisy labels. ‘Closed-set’ means that true class labels of noisy labeled samples are within known classes of noisy data. Many efforts have been devoted to tackling the closed-set noisy label issue, such as sample reweighting [33], label correction [44], loss correction [11,38]. Shu et al. [33] leverage multiple layer perceptions to automatically suppress noisy samples. Zhang et al. [44] conduct a dynamic linear combination of noisy label and prediction to refurbish labels. Wang et al. [38] utilize a set of trusted clean-labeled samples to estimate NTM for loss correction.

Open-set noisy labels. ‘Open-set’ indicates that true class labels of noisy labeled samples are outside known classes of noisy data. Contrast to the widely explored closed-set noisy label problem, it has been seldom invested, which is our focus in this study. Prior works concentrate on resampling approach [20,31,36]. This solution is intuitive but suboptimal as the discarded data may contain useful information for shaping the decision boundary. Label correction and loss correction are vulnerable with open-set label noises since these methods are based on the assumption that training samples belong to known classes in the dataset. Hence, Yu et al. [41] transfer the ‘open-set’ problem to ‘closed-set’ one and then correct noisy labels. Xia et al. [39] extend the traditional NTM with an open-set class, so as to model the open-set label noises. Our work represents the first effort to combat open-set noisy labels in DA.

3. Methodology

3.1. Self-training with SimT

We focus on the problem of domain adaptive semantic segmentation, wherein we only have access to target domain images \( \mathcal{X}_T = \{X_t \in \mathbb{R}^{H \times W \times 3}\}_{t \in T} \) with spatial dimension of \( H \times W \) and the corresponding pseudo labels \( \mathcal{Y}_T = \{Y_t \in \{0,1\}^{H \times W \times C}\}_{t \in T} \). Note that \( Y_t \) is the pixel-wise label map with \( C \) classes, and pseudo labels are derived from a given black-box model \( f_b(\cdot) \). We aim to learn a segmentation model \( f(\cdot) \), parameterized by \( w \) that can correctly categorize pixels for target data \( \mathcal{X}_T \). Self-training based methods [11,42,46,50] utilize pseudo labels as approximated ground truth labels for model training, and the cross-entropy loss over the target dataset for self-training is \( \mathcal{L}_{ST} = - \sum_{t \in T} Y_t \log f(X_t) \). By minimizing \( \mathcal{L}_{ST} \) on target data, the optimized model can be discriminatively adapted to the target domain. However, noises in pseudo-labeled target domain data, including closed-set label noises and open-set label noises, inevitably deteriorate the performance of self-training based DA methods.

To enhance the noise tolerance property of \( \mathcal{L}_{ST} \), we propose SimT to benefit loss correction, as shown in Figure 2. SimT is formulated as \( T \in [0,1]^{(C+n)xC} \), where \( T_{1:C} \): models closed-set noise distribution and \( T_{C+1:C+n} \): models open-set noise distribution in target domain. \( n \) is a hyper-parameter that indicates the potential open-set class number, and we set \( n > 1 \) to implicitly model the diverse semantics within open-set classes. The defined SimT specifies the probability of ground truth label (\( Y_t = j \)) flipping to noisy label (\( Y_t = k \)) by \( T_{jk} = p(Y_t = k | Y_t = j) \), and we have \( \sum_{k=1}^{C} T_{jk} = 1 \). Assuming the class posterior probability for noisy pseudo label is \( p(\bar{Y}_t = k | X_t; w) \in [0,1]^C \) and for ground truth label is \( p(Y_t = j | X_t; w) \in [0,1]^{C+n} \), SimT bridges the generated pseudo label \( \bar{Y}_t \) and the ground truth label \( Y_t \) through:

\[
p(\bar{Y}_t = k | X_t; w) = \sum_{j=1}^{C+n} T_{jk} \cdot p(Y_t = j | X_t; w),
\]

\[
\Rightarrow p(\bar{Y}_t | X_t; w) = p(Y_t | X_t; w) \cdot T.
\]
target image $x_t$; Segmentation Net $f(\cdot;w)$; Clean class posterior $p(Y_t|X_t;w)$; Noisy class posterior $p(\tilde{Y}_t|X_t;w)$; Target pseudo label $\tilde{Y}_t$;

Figure 2. The proposed domain adaption framework contains a segmentation net and a SimT. Target images are passed through the segmentation net $f(\cdot;w)$ to perform semantic segmentation, and supervision signals are corrected by the learnable SimT. SimT is estimated through the proposed (a) volume regularization, (b) anchor guidance and (c) convex guarantee.

With the formulated noise distribution ($T$), we correct the self-training loss over noisy pseudo-labeled target data as

$$\mathcal{L}_{LC} = - \sum_{t \in T} \tilde{Y}_t \log [f (X_t)_w T].$$

This corrected loss encourages the similarity between noisy class posterior probabilities $p(\tilde{Y}_t \mid X_t; w) = f (X_t)_w T$ and noisy pseudo labels $\tilde{Y}_t$. It is obvious that once the true SimT is obtained, we can recover the desired estimation of clean class posterior probability $p(Y_t \mid X_t; w)$ by the softmax of first $C$ classes output $f (X_t)_w$ even training the segmentation model with noisy data. Considering no clean-labeled target domain data is accessible, previous work estimates NTM through a meta-learning strategy [11]. Different from prior art, we form NTM to make up a simplex and solve it through computational geometry analysis without the requirement of a clean labeled target data set.

### 3.2. Volume Regularization of SimT

Semantic segmentation can be regarded as pixel-wise classification, thus we make geometry analysis at pixel level. Given a pixel $x$ within target domain image $X_t$, the noisy class posterior probability $p(\tilde{y} \mid x) = [p(\tilde{y} = 1 \mid x), p(\tilde{y} = 2 \mid x), \ldots, p(\tilde{y} = C \mid x)]$ can be regarded as a point in $C$-dimensional space, and the clean class posterior probability $p(y \mid x) = [p(y = 1 \mid x), p(y = 2 \mid x), \ldots, p(y = C + n \mid x)]$ is a $(C + n)$ dimensional vector. These two class posterior can be bridged by the proposed SimT via $p(\tilde{y} \mid x) = p(y \mid x)T$ as in Eq. (1). In the meanwhile, conditions of $\sum_{j=1}^{C+n} p(y = j \mid x) = 1$ and $p(y \mid x) > 0$ are satisfied, thus, the noisy class posterior probability $p(\tilde{y} \mid x)$ can be interpreted as a convex combination of $T$ rows with factor of $p(y \mid x)$. In other words, the simplex formed by rows of $T$ encloses $p(\tilde{y} \mid x)$ for any input pixel $x$ [16]. However, SimT is not identifiable if no assumption is made since there exists an infinite number of simplices enclosing $p(\tilde{y} \mid x)$.

To approximate the true SimT, we follow the sufficiently scattered assumption [16] and devise volume regularization. Among all simplices enclosing $p(\tilde{y} \mid x)$, the true SimT should be the one with minimum volume under the sufficiently scattered assumption. In this circumstance, the volume consisted by the target clean class posterior will converge the maximum one, and outputs of optimized segmentation model thus approximate to the ground truth label distribution. This motivates us to propose the following volume regularization for the optimization of SimT:

$$\mathcal{L}_{Volume}^{SimT} = \log [\text{Vol}(T)] = \log \sqrt{\det(T^TT)/(N!)}$$

where $\sqrt{\det(T^TT)/(N!)}$ denotes the volume of a r-parallelogram defined by the non-square matrix $T$ in geometry [8, 10]. Since $N!$ is a constant number, we discard this denominator in experiments. $\log(\cdot)$ function is to stabilize the optimization for computed determinant, following [8].

To make $T$ differentiable and satisfy the condition of $T \in [0, 1]^{(C+n)\times C}$, $\sum_{j=1}^{C} T_{jk} = 1$, we utilize the reparameterization method. Specifically, we first randomly initialize a matrix $U \in \mathbb{R}^{(C+n)\times C}$. To preserve the diagonally dominant property of closed-set part of SimT ($T_{1:C}$), the diagonal prior $I$ is introduced, which has been widely used in the literature of NTM estimation [16, 26]. Considering segmentation model tends to classify samples into majority categories, the class distribution of pseudo labels $C$ is involved. Incorporating these prior informations, we obtain $V = C \cdot \sigma(U) + I$, where $\sigma(\cdot)$ is the sigmoid function to avoid negative value in SimT. Then we do the normalization of $T_{jk} = \frac{V_{jk}}{\sum_{k=1}^{C} V_{jk}C}$ to derive SimT ($T$). Since both sigmoid function and normalization operation are differentiable, SimT can be updated through gradient descent on $U$.

### 3.3. Anchor Guidance of SimT

Considering the diagonally dominant assumption is not available for the open-set part $T_{C:C+n}$ of SimT, the open-set part will eventually collapse to a trivial solution by minimizing the volume regularization $\mathcal{L}_{Volume}^{SimT}$. To prevent the trivial solution of open-set part in SimT, we devise anchor guidance, where anchor points are detected to guide
the estimation of SimT. Formally, anchor point of class $c$ is defined as the pixel whose clean class posterior are $p(y = c \mid x^c) = 1$ and $p(y \neq c \mid x^c) = 0$. Thus we have $p(y = k \mid x^c) = \sum_{j=1}^{C+n} T_{jk} \cdot p(y = j \mid x^c) = T_{ck}$, indicating that the noisy class posterior probability of anchor point $p(y = k \mid x^c)$ lies at one vertex of SimT ($T_{ck}$). Hence, the simplex formed by noisy class posteriors of anchor points can recover the true SimT in theory. However, under mixed open-set and closed-set noisy labels, anchor points are hard to detect. Therefore, we define anchor point as the most representative pixel within each closed-set and open-set class with the largest clean class posterior probability. The anchor point in class $c$ is thus derived via

$$x^c = \arg\max_x p(y = c \mid x; w).$$

(4)

In geometrical perspective, the estimated $x^c$ is the point, whose noisy class posterior $p(y = k \mid x^c)$ is closest to the vertex $T_{ck}$ in $C$-dimensional space. Then we leverage computed anchor points to guide the approximation of the true SimT, and the anchor guidance is formulated as

$$L_{\text{Anchor}} = \sum_{c=1}^{C+n} \sum_{k=1}^{C} \| T_{ck} - p(y = k \mid x^c; w^{\text{fixed}}) \|^2,$$

where the noisy class posterior of anchor point $p(y = k \mid x^c; w^{\text{fixed}})$ is computed from the segmentation model warmed up with pseudo labels, denoted as $f(\cdot)_{w^{\text{fixed}}}$ with parameter of $w^{\text{fixed}}$. For each target image, only occurred classes would derive the corresponding anchor points. Then those detected anchor points are leveraged to calculate the regularization of anchor guidance. The accurate detection of anchor points in open-set classes can prevent the trivial solution and guide to approximate the true SimT.

**Auxiliary loss for anchor point detection.** To facilitate the detection of anchor points in open-set classes, an auxiliary loss enhancing the discriminative feature learning among closed-set and open-set data is designed. Firstly, we introduce a threshold to select confident pixels as non-noisy samples, and class posterior probabilities of those selected pixels are directly supervised to optimize the segmentation model. This direct supervision explicitly enables the model to encode the diverse semantics for both closed-set and open-set classes. Specifically, the confidence score is derived from the segmentation model $f(\cdot)_{w^{\text{fixed}}}$. We select confident closed-set pixels $X_K = \{x_k\}_{k \in K}$ through the condition of $\max f(x_k)_{w^{\text{fixed}}} > \tau_{\text{high}}$, and the corresponding labels $\tilde{y}_k$ are one-hot vectors of $\tilde{y}_k = \arg\max f(x_k)_{w^{\text{fixed}}}$. Confident open-set pixels $X_U = \{x_u\}_{u \in U}$ are those with $\max f(x_u)_{w^{\text{fixed}}} < \tau_{\text{low}}$ & $\arg\max f(x_u)_{w} > C$, and their one-hot ground truth labels $\tilde{y}_u$ are derived from $\tilde{y}_u = \arg\max f(x_u)_{w}$. Note that $\tau_{\text{high}}, \tau_{\text{low}}$ are hyper-parameters and empirically set to be 0.8, 0.2. To further detect open-set class samples, we make the open-set classifier to output the second-largest probability for closed-set pixels $X_K$, which is constrained by the second term of following auxiliary loss:

$$L_{\text{Aux}} = - \sum_{i \in K \cup U} \tilde{y}_i \log f(x_i)_{w} - \lambda \sum_{k \in K} \tilde{y}_o \log[f(x_k)_{w} \setminus \tilde{y}_k],$$

(6)

where $\tilde{y}_o$ is the one-hot form of $\tilde{y}_o = \arg\max_{c \in [C,C+n]} f(x_k)_{w}$, and $\lambda$ is a hyper-parameter to control the contribution of the second term in auxiliary loss. The first item maintains performance in the closed-set classes and detects open-set class pixels. The second term matches the masked-probability with the most possible open-set class, which prevents the open-set classifier far away from closed-set classifier and facilitates the discovery of open-set class pixels. The discriminative feature learning induced by auxiliary loss simplifies the detection of anchor points, hence enhancing the performance of anchor guidance.

### 3.4. Convex Guarantee of SimT

As analyzed in §1, the trivial solution of open-set part will lead to non-convex SimT. In other word, one of vertices in SimT may be the convex combination of the rest vertices. This means different ground truth classes may show the similar noise transition probability, and results in multiple clean class posteriors $p(y \mid x)$ correspond to the same noisy class posterior $p(y \mid x)$. Considering similar noise transition probabilities reveal close feature representations [39], the non-convex SimT may lead the segmentation model to produce undiscriminating features. Hence, we advance convex guarantee to prevent SimT from becoming non-convex. Inspired by the insight that each vertex of the convex hull cannot be represented by any convex combinations of rest vertices, we first introduce a weighting matrix $u \in [0, 1]^{(C+n) \times (C+n)}$ with constraints of $u_{j,k} = -1$ and $\sum_{j \neq k} u_{j,k} = 1$ to denote convex combination coefficients. The reparameterization method is introduced to make the weighting matrix $u$ differentiable and satisfy its constraints. Then the convex guarantee is formulated as a minimax problem, where the optimization objective is $\max \min u Tu^T$. By updating $u$ to minimize the optimization objective, the weighting matrix learns to represent each vertex with the convex combination of rest vertices. In turn, by updating $T$ to maximize the optimization objective, each vertex is pushed away from the convex combination of rest vertices, and SimT is encouraged to be convex. For simplification, convex guarantee regularization is formulated as

$$L_{\text{Convex}} = - \|u Tu^T\|_2,$$

(7)

The convex guaranteed SimT enforces the class posterior distribution to be scattered, thereby the segmentation model is encouraged to extract discriminative features among closed-set and open-set classes.
3.5. Optimization summary for SimT

With the proposed regularizers, SimT is optimized to estimate the mixed closed-set and open-set noise distributions, thereby benefiting the loss correction of pseudo-labeled target data. In summary, the overall objective function for our DA semantic segmentation framework is formulated as

\[
\mathcal{L}(w, T) = \mathcal{L}_{LC}(w, T) + \mathcal{L}_{Aux}(w) + \alpha \mathcal{L}_{SimT}^T(T) + \beta \mathcal{L}_{ConvexT}(T) + \gamma \mathcal{L}_{ConvexT}(T),
\]

where corrected loss \(\mathcal{L}_{LC}(w, T)\) (Eq. (2)) and \(\mathcal{L}_{Aux}(w)\) (Eq. (6)) optimize the segmentation model. \(\mathcal{L}_{LC}(w, T)\) and three regularizers \(\mathcal{L}_{SimT}^T(T)\) (Eq. (3)), \(\mathcal{L}_{SimT}^T(T)\) (Eq. (5)), \(\mathcal{L}_{ConvexT}(T)\) (Eq. (7)) are minimized to approximate the true SimT. \(\alpha, \beta, \gamma\) are regularization coefficients to balance contributions of three regularizers.

4. Experiments

4.1. Training Details

Datasets. We evaluate the performance of our method on two DA semantic segmentation tasks, including a synthetic-to-real scenario (GTAS [29]→Cityscapes [7]) and a surgical instrument type segmentation task (Endovis17 [2]→Endovis18 [1]). GTAS contains 24,966 images captured from a video game, and 19 classes in pixel-wise annotations are compatible with Cityscapes. Cityscapes is a real-world semantic segmentation dataset obtained in driving scenarios, containing 34 classes according to the manual annotations. Specifically, 2,975 unlabeled images are regarded as the target domain training data, and evaluations are performed on 500 validation images with manual annotations. Endovis17 contains 1800 source images captured from abdominal porcine procedure, and 3 instrument type classes are compatible with Endovis18. Endovis18 is an abdominal procedure dataset, consisting of 2384 target images with 6 instrument types. Following [21], we divide Endovis18 into 1639 unlabeled training images and 596 validation images.

Implementation Details. We adopt DeepLab-v2 [3] backbone with encoder of ResNet-101 [12] as our segmentation model. Given the pseudo-labeled target domain data derived from a black-box model \(f_b(\cdot)\), we warm up the segmentation model to obtain \(f(\cdot)_{w, \text{init}}\) with \(C\)-way output probabilities. The classifier of segmentation model is then extended to output \((C + n)\)-way clean class posterior probabilities. The extended model is denoted as \(f(\cdot)_{w}\) we aim to optimize, and we obtain final predictions from first \(C\)-way of \(f(\cdot)_{w}\) during inference phase. Our method is implemented with the PyTorch library on Nvidia 2080Ti. We adopt polynomial learning rate scheduling to optimize the feature extractor with the initial learning rate of 6e-4, while it is set as 6e-3 for the optimization of classifier and SimT. The batch size is set as 1, and the maximum iteration number is 40000. Hyper-parameters of \(n, \lambda, \alpha, \beta, \gamma\) are set as 15, 0.1, 1.0, 1.0, 0.1 in our implementation. Performances are evaluated by the widely utilized metrics, intersection-over-union (IoU) of each class and the mean IoU (mIoU).

4.2. Results on GTAS→Cityscapes

UDA. We first verify the effectiveness of our approach in the GTAS→Cityscapes scenario, and the corresponding comparison results are listed in Table 1 with the first and second best results highlighted in bold and underline. Under the setting of UDA, our black-box model \(f_b(\cdot)\) can be any existing UDA models, and we leverage model of [22] to generate pseudo labels for unlabeled target domain data.
Overall, the proposed SimT method achieves at the state-of-the-art mIoU score of 58.9%, surpassing all other UDA methods by a large margin. In terms of the per class IoU score, the proposed SimT method shows the superior performance. To be specific, we arrive at the best on 10 out of 19 categories and the second best on 3 out of 19 categories. Moreover, SimT outperforms other self-training based methods [9,11,13,22,46,49,50] by a large margin of 18.8% in mIoU.

### 4.4. Ablation Study

**Ablation Experiments.** To investigate effects of individual components within SimT, we design ablation experiments under UDA setting (GTA5→Cityscapes), as shown in Table 3. In particular, we validate three proposed regularizers, i.e., volume regularization, anchor guidance, and convex guarantee of SimT. We adopt AdaptSegNet [34] as the black-box model to generate pseudo labels for target domain data. It is observed that incorporating SimT (row 11) to model mixed closed-set and open-set noise distributions can correct supervision signals on target data and obtain the performance gain with 5.6% increment in mIoU. Ablating each proposed component (rows 7-10) induces the performance degradation. It is because that the designed regularizers are complementary and combining all regularizers with properly chosen coefficients (i.e., $\alpha = 1.0$, $\beta = 1.0$, $\gamma = 0.1$) can reach a trade-off to better approximate the true SimT, as shown in Figure 3.

### Effectiveness of SimT

We then compare SimT with four typical self-training based UDA models [11,46,49,50]. For a fair comparison, all competed methods use AdaptSegNet to derive pseudo labels. As listed in Table 3, the proposed SimT (row 11) is superior to other self-training methods, including entropy minimization [49] (row 3), threshold [50] (row 4) and uncertainty [46] (row 5) based sample selection, MetaCorrection based rectification [11] (row 6), yielding increments of 2.9%, 3.6%, 1.9%, 0.7% mIoU.

### 4.3. Results on Endovis17→Endovis18

**UDA.** We further verify the effectiveness of SimT in a medical scenario, i.e., Endovis17→Endovis18, and the quantitative comparison results are listed in Table 2. We utilize the UDA model of [21] to generate pseudo labels for unlabeled target domain data. Obviously, the proposed SimT method obtains the state-of-the-art performance with 58.12% mIoU, outperforming all other UDA methods by a large margin. This observation proves the impact of the proposed method in medical image analysis.

**SFDA.** In the SFDA setting of Endovis17→Endovis18, performance of the proposed SimT is evaluated based on the black-box model $f_0(\cdot)$ is borrowed from [15]. The corresponding comparison results of our method and baseline models in Table 2 show that the proposed SimT is superior to all other SFDA methods with mIoU of 55.23%.

### 4.4. Ablation Study

**Ablation Experiments.** To investigate effects of individual components within SimT, we design ablation experiments under UDA setting (GTA5→Cityscapes), as shown in Table 3. In particular, we validate three proposed regularizers, i.e., volume regularization, anchor guidance, and convex guarantee, and the auxiliary loss devised for anchor guidance is also ablated to validate its effectiveness. We adopt AdaptSegNet [34] as the black-box model to generate pseudo labels for target domain data. It is observed that incorporating SimT (row 11) to model mixed closed-set and open-set noise distributions can correct supervision signals on target data and obtain the performance gain with 5.6% increment in mIoU. Ablating each proposed component (rows 7-10) induces the performance degradation. It is because that the designed regularizers are complementary and combining all regularizers with properly chosen coefficients (i.e., $\alpha = 1.0$, $\beta = 1.0$, $\gamma = 0.1$) can reach a trade-off to better approximate the true SimT, as shown in Figure 3.

**Effectiveness of SimT.** We then compare SimT with four typical self-training based UDA models [11,46,49,50]. For a fair comparison, all competed methods use AdaptSegNet to derive pseudo labels. As listed in Table 3, the proposed SimT (row 11) is superior to other self-training methods, including entropy minimization [49] (row 3), threshold [50] (row 4) and uncertainty [46] (row 5) based sample selection, MetaCorrection based rectification [11] (row 6), yielding increments of 2.9%, 3.6%, 1.9%, 0.7% mIoU.
With entropy minimization, considerable noisy labels inevitably lead to unsatisfactory adaptation. Although self-training methods in [46, 50] discard noisy samples and may filter out open-set label noises, but lose useful information in those discarded pixels, especially when they are around the decision boundary. Both MetaCorrection [11] and SimT distill effective information from all samples. Comparing with MetaCorrection that only models closed-set noise distribution, the proposed SimT additionally ameliorates open-set noise issues and shows the superior performance.

Robustness to Various Types of Noise. To explore the robustness of our SimT, we further evaluate it under different types of label noises. To be specific, we adopt four UDA models, including AdaptSegNet [34], LTIR [14], DSP [9] and BAPA-Net [22], as black-box models to generate noisy pseudo labels. Then SimT is applied to mitigate the closed-set and open-set noise problem within pseudo labels. As illustrated in Table 3, these four UDA models have significantly improved performance with the proposed SimT.

4.5. Visualization Results

Segmentation Visualization. As illustrated in Figure 4, we provide some typical qualitative segmentation results of target data in GTA5→Cityscapes scenario. Compare with the baseline BAPA-Net [22], the proposed SimT shows superior performance in identifying small-scale objectives (e.g., ‘traffic sgn’, rows 2, 3, 5) and differentiating confused class samples (e.g., ‘rider’, ‘person’ and ‘bike’, rows 1, 4). In practice, we observe that label noises in pseudo-labeled target data tend to appear in the minor categories and ambiguous categories, leading to the unsatisfactory performance of baselines. On the contrary, the proposed SimT corrects supervision signals for target data and alleviates the noise issues. We also illustrate qualitative results of target Endovis18 data under UDA setting, as shown in Figure 5.

SimT Visualization. Inspired by the fact that confusion matrix is a crude estimator of the true NTM [43], we visualize the learned SimT and the confusion matrix of pseudo labels for reference, as illustrated in Figure 6 (a, b). It is obvious that the optimized SimT is similar to confusion matrix, demonstrating the proposed SimT can model the closed-set and open-set noise distributions and generate corrected supervision signals for pseudo-labeled target data.

Open-set Class Visualization. We further visualize detected open-set class regions, which are denoted by ‘white’ areas in Figure 6 (c). From left to right, they are input image, ground truth and clean class posterior prediction, respectively. We observe the proposed method is able to detect open-set class pixels and inherently has the capability of alleviating open-set label noise issue in target data.

5. Conclusion

In this paper, we study a general and practical DA semantic segmentation setting where only pseudo-labeled target data is accessible through a black-box model. SimT is advanced to model mixed close-set and open-set noise distributions within pseudo-labeled target data. We formulate columns of SimT to make up a simplex and interpret it through the computational geometry analysis. Based on the geometry analysis, we devise three regularizers, i.e. volume regularization, anchor guidance, convex guarantee, to approximate the true SimT, which is further utilized to correct noise issues in pseudo labels and enhance the generalization ability of segmentation model on the target domain data. Extensive experimental results demonstrate that the proposed SimT can be flexibly plugged into existing UDA and SFDMA methods to boost the performance. In the future, we plan to introduce open-set class prior to enable the semantic separation inner detected open-set regions.
References


[34] 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 CVPR, pages 7472–7481, 2018. 2, 6, 7, 8


[42] 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 CVPR, 2021. 1, 3, 6, 7


